University of Glasgow

Bone, Jordan (2025) *Exploring evolutionary dynamics of influenza viruses through the lens of equine influenza.* PhD thesis.

https://theses.gla.ac.uk/85122/

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

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

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

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

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

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

Exploring Evolutionary Dynamics of Influenza Viruses Through the Lens of Equine Influenza

Jordan Bone BSc (Hons), MSc

Submitted in fulfilment of the requirements for a Degree of Doctorate of Philosophy in Veterinary Medicine

School of Biodiversity, One Health and Veterinary Medicine

College of Medical, Veterinary & Life Sciences

University of Glasgow

June 2024

Abstract

Though many vaccines exist to confer protection from Influenza A viruses, they remain viruses of great concern in both mammalian and avian species. Influenza A viruses circulate in many wild, domestic and human species; often with the potential to transmit between these populations. Emergence in these other populations can lead to widespread transmission and/or severe pathology and indeed has occurred multiple times over just the last century.

As obligate pathogens, influenza viruses are only able to evolve during infection of hosts, as it is the only setting in which they can replicate. However, in addition to the local environment (i.e. the infected host) influenza viruses must also adapt to transmission between hosts. Differences between hosts can be minimal but may be as dramatic as alternative tissue tropism or even host species. This work aims to study how influenza viruses evolve both within infected hosts and across transmission events, and the interaction between these two, sometimes competing, ecological niches that viruses must adapt to. Integral to the continued success of influenza viruses is the ability to circumvent the host immune system. Host adaptive immunity places strong selective pressures upon viral populations. Transmission experiments were carried out in which mixed populations of horses (either vaccinated or unvaccinated) were sequentially exposed to one another to create a five-step chain of transmission. The first experiment mixed naive individuals with horses that had received a multivalent vaccine, the second mixed naive horses with hosts that had received a univalent vaccine. Horses were nasally-swabbed daily in order to collect shed virus particles which could then be quantified and deep-sequenced.

137 qPCR values and 53 sequences of viral populations were collected. Differences in viral load, consensus genomes and low-frequency mutations were observed across transmission chains and between vaccinated and unvaccinated hosts. Unvaccinated horses shed more virus than their vaccinated counterparts, though this difference was much greater when comparing naive hosts to those that received a multivalent vaccine. Conversely, genomic diversity at the consensus level appeared highest in hosts that received the monovalent vaccine - suggesting strong selective pressures that mutations are attempting to overcome. This genetic diversity however was not reflected in sub-consensus reads, where lessened selective pressures allowed for greater diversification of viruses replicating in unvaccinated hosts.

Table of Contents

List of Tables

List of Figures

Acknowledgement

After some of my first undergraduate lectures on viral evolution, I was so enthralled that I immediately spoke to the professor, inquiring how best to go about pursuing virology as a career path. The lecturer, Pablo Murcia, gave me his recommendations: secure a Master's degree, some work experience, then complete a PhD and move onto a life of research. Hence, I find myself here, submitting a thesis after four years of gruelling work and challenging circumstances. Truly, none of this would have been possible without the guidance of all four supervisors but I'd especially like to give thanks to Pablo for the years of support, endurance and stability. But I'm thankful to have also had Willie Weir's know-how & ability to handle any situation, Richard Orton's familiarity & expertise in the field and Roman Biek's tutelage starting with guest lectures in my undergraduate years to specialised training during my MSc programme and then his insight throughout my PhD.

Additionally, despite the difficulties and seclusion presented by the pandemic, many others have been of great assistance. From Chris Illingworth walking me through bioinformatic tools & reviewing sample writing to Matthew Arnold's guidance & enthusiasm for structural biology and the whole Murcia group for their reassurance, feedback & understanding of a less-than-reliable colleague; a PhD project is never a wholly individual experience and these are just a few of the people that eased the path.

Of course, outside of academic settings I was supported by myriad people: friends & family alike. Too numerous to list all here but know that you are always in my heart and I'd be a mere fraction of the man I am without you by my side.

Author's Declaration

I, Jordan Bone, certify that, except where explicit reference is made to the contribution of others, this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Dated 14th June 2024

Abbreviations

- AHT: Animal Health Trust
- AMR: Anti-Microbial Resistance
- BAM: Binary Alignment Map
- CCR5: Chemokine Receptor 5
- C_t: Cycle Threshold
- DALY: Disability-Adjusted Life Years
- DPC: Days Post-Contact
- EI: Equine Influenza
- EIV: Equine Influenza Virus(es)
- FMDV: Foot-and-Mouth Disease Virus
- HA: Haemagglutinin
- HAART: Highly-Active Anti-Retroviral Therapy
- HIV: Human Immunodeficiency Virus
- HPAI: High-pathogenicity Avian Influenza
- IAV: Influenza A Virus(es)
- LDDT: Local Distance Difference Test
- LFV: Low-frequency Variant
- LPAI: Low-pathogenicity Avian Influenza
- PA: Polymerase Acidic protein
- PB1: Polymerase Basic protein 1
- PB2: Polymerase Basic protein 2
- MCC: Maximum Clade Credibility
- MCMC: Monte-Carlo Markov Chain
- MP: Matrix Protein
- NA: Neuraminidase
- NEP: Nuclear Export Protein
- NP: Nucleoprotein
- NS1: Non-structural protein 1
- OAS: Original Antigenic Sin
- OIE: Office International des Epizooties (now WOAH)
- qPCR: Quantitative Polymerase Chain Reaction
- RNAP2: RNA Polymerase II
- RdRp: RNA-dependent RNA Polymerase
- RNP: Ribonucleoprotein
- SARS-CoV-1: Severe Acute Respiratory Syndrome Coronavirus 1
- SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
- SRH: Serial Radial Haemolysis
- ssRNA: Single-stranded Ribonucleic Acid
- VCT: Variant Call Tool
- VSV: Vesicular Stomatitis Virus
- WOAH: World Organisation for Animal Health

1 Introduction

1.1 Impact & Importance of Influenza Viruses

IAVs generally possess high transmissibility and moderate lethality and so these viruses can result in large losses of life and/or Disability-Adjusted Life Years (DALY) in a variety of species (Hicks et al., 2020). Global human influenza mortality was estimated at 290,000–645,000 (Iuliano et al., 2018), however this figure was calculated prior to the COVID-19 pandemic which is assumed to have drastically altered the landscape of infectious respiratory diseases. Accounting for the burden of illness without mortality, the European Centre for Disease Prevention and Control estimated 52.6 DALYs per 100,000 inhabitants of the European Union population (Cassini et al., 2018). This dwarfs the 15.5 DALYs estimated for the impacts of COVID-19 on the size-scaled population of Scotland by Wyper et al. (2022).

IAVs cause disease through viral cytotoxicity and/or immunopathology (Belser et al., 2020). Innate immune involvement causes most of the commonly observed symptoms (fever, myalgia, malaise, rhinitis and dry cough) though seasonal influenza is generally self-limiting in the immunocompetent (Nicholson, 1992; Ryu & Cowling, 2021). Mortality from seasonal influenza most commonly results from secondary bacterial infections (Cullinane & Newton, 2013; Klenerman & Zinkernagel, 1998; Wood & Grenfell, 2009). Emergence of novel influenza viruses (i.e. pandemic 'flu') is often associated with higher morbidity and mortality due to a lack of adaptive immune memory in the population. IAVs expressing a substantially novel antigenic type can lead to the over-activation of innate immune cells and molecules causing severe, potentially life-threatening, immunopathology referred to as a Cytokine Storm. Additionally, IAV in domestic chickens is often classified based on its virulence into low- or high-pathogenicity avian influenza [L/HPAI] (Abdel-Moneim et al., 2010; Ganti et al., 2021; Monne et al., 2014).

1.2 Influenza A Viruses

1.2.1 Virus Structure

Inside the virion, the eight genomic segments are complexed with the viral polymerase and surrounded by nucleoprotein. The envelope membrane is acquired from the host cell on exiting and is studded with three viral proteins: haemagglutinin, neuraminidase and matrix (M2) channels. While Figure 1.1 represents the virion spherically, it should be noted that the morphology of influenza viruses is somewhat variable, ranging from spherical to bacilliform to filamentous (Chlanda et al., 2015; Seladi-Schulman et al., 2014).

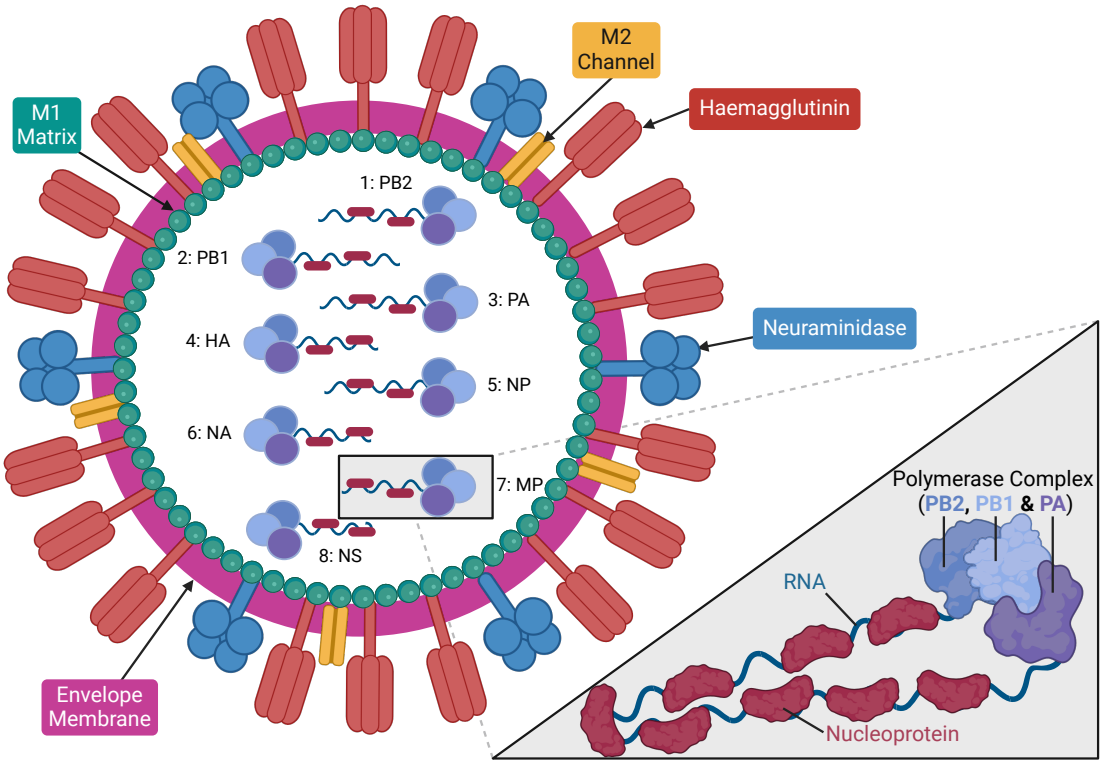


Figure 1.1: Cartoon representation of an Influenza A virion, with an expanded view of a genomic segment. Adapted from Grant et al. (2014) and produced using BioRender.com

Influenza A Viruses are differentiated through the subtyping of their two surface proteins: haemagglutinin (HA), with eighteen subtypes (H1-H18), and neuraminidase (NA), with eleven subtypes (N1-N11). As these proteins are exposed on the extracellular surface of the virion, they are the major antigens of the influenza virus. They facilitate viral entry into (HA), and release from (NA) the host cell and so their activities bracket the replication cycle. Classically, HA and NA are the targets of adaptive immunity and, consequently, vaccine development. Both proteins allow viral sub-classification based on sequence diversity (Figure 1.2): haemagglutinin sequences may be divided into two groups, while neuraminidase sequences form three groups, one of which comprises the subtypes N10 and N11, which are substantially different from the other subtypes (Ekiert et al., 2011; Wu et al., 2014).

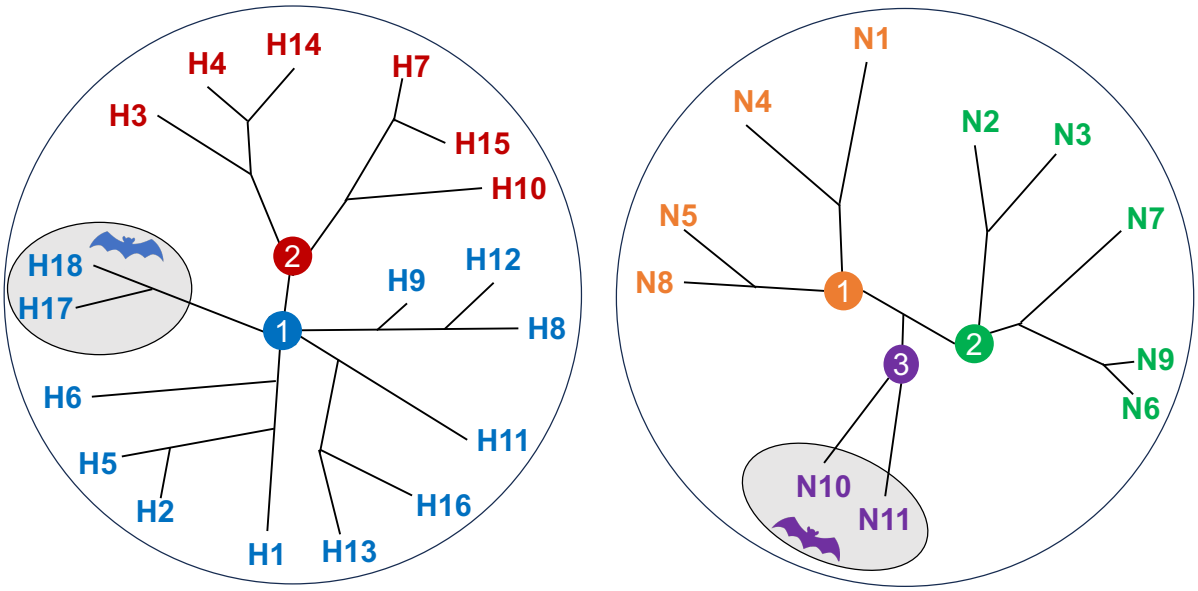


Figure 1.2: A) Haemagglutinin and B) Neuraminidase sequences are divided into 2 and 3 subtypes respectively. Both trees have shaded areas representing subtypes found exclusively in bats, to emphasise their stark divergence from other influenza viruses. Adapted from Wu et al. 2014

1.2.2 Viral Genome Organisation

The genomes of IAV are organised into eight distinct negative-sense RNA segments numbered in decreasing order of length in nucleotides (Figure 1.3). Their negative-sense, single stranded genome taxonomically places them within the family *Orthomyxoviridae* of the *Negarnaviricota* phyla, alternatively allocated within Group IV of the Baltimore classification system (Lefkowitz et al., 2018; Walker et al., 2022). Each RNA segment is tightly packed by the nucleoprotein , which in turn folds into a hairpin structure with the polymerase complex (Bera et al., 2017) holding both the 3' and 5' ends; overall this is known as the ribonucleoprotein (RNP). Noncoding sequences at the ends of each genome segment are conserved between all segments in all influenza viruses.

The heterotrimeric polymerase complex is comprised of Polymerase Basic protein 2 (PB2), Polymerase Basic protein 1 (PB1) and Polymerase Acidic protein (PA). The PB2 protein binds the cap of host pre-mRNA molecules in order to begin the cap-snatching process (Gocnikova & Russ, 2007) while PB1 conjugates RNA bases together during replication of the viral genome. PA is integral to the replication cycle of IAV within the cell and acts by cleaving the cap from host mRNA (Rash et al., 2014). Additionally, the other proteins present in the polymerase complex are structurally secured by PA.

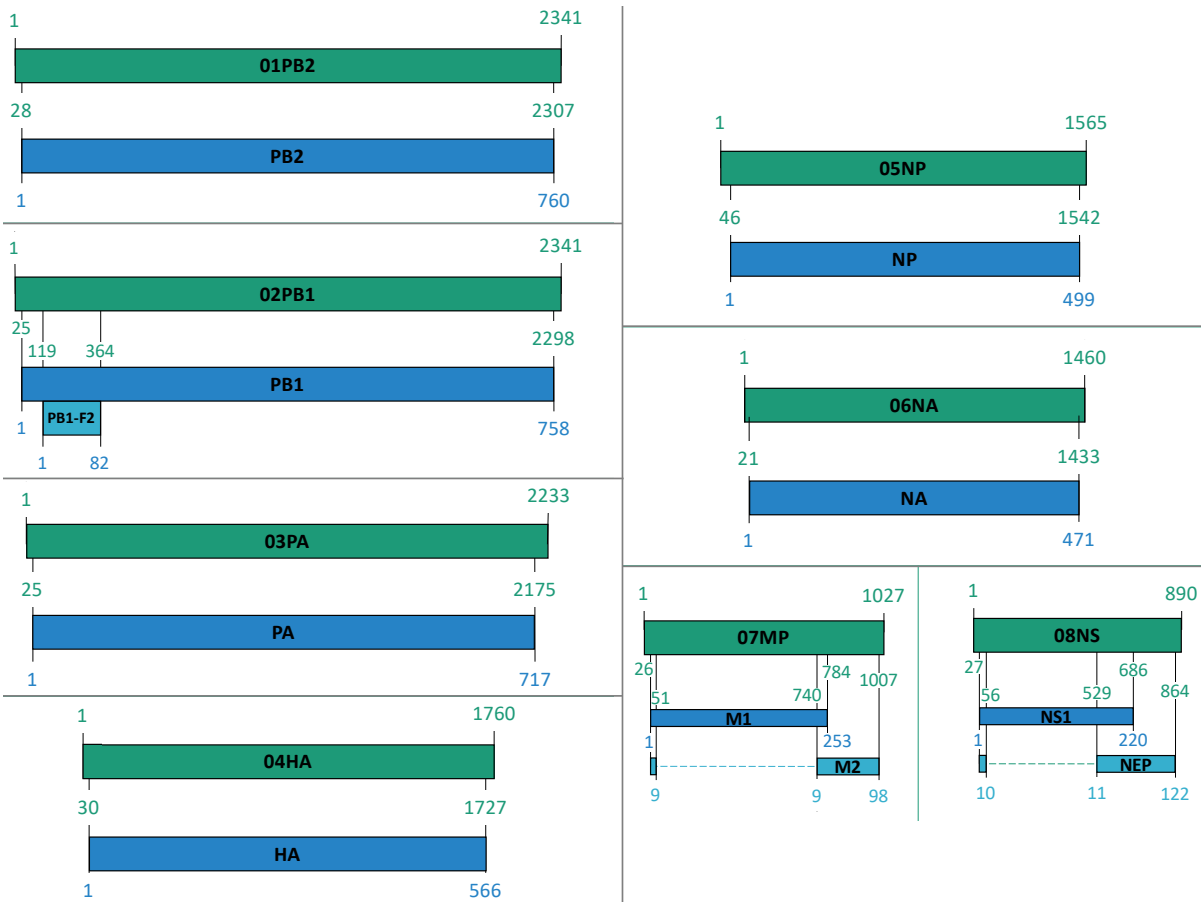


Figure 1.3: Equine H3N8 Influenza A genomic organisation. Nucleotides of the eight genomic segments are annotated in green with corresponding amino acids on the coding regions of each segment.

Haemagglutinin (HA) is a homotrimeric surface protein that mediates cell entry by binding to sugar moieties called sialic acids and initiating receptor-mediated endocytosis via endosome-acidification (Russell, 2021; Schotsaert & García-Sastre, 2014).

Nucleoprotein (NP) is involved in nuclear import & viral packaging (Abdel-Moneim et al., 2011). All viral RNA synthesis occurs in the nucleus, where trafficking of large molecules is tightly regulated by the cell. Viral RNPs are too large for passive diffusion across the nuclear membrane and thus rely on an active nuclear import mechanism. All proteins in the RNP complex possess nuclear localisation signals (NLSs). The transport of proteins across the nuclear membrane is an active process initiated by karyopherin α upon recognition of a protein presenting an NLS.

Neuraminidase (NA), the second antigenic surface protein, is a homotetramer responsible for disassociation from the host cell during viral exit (von Itzstein, 2007). This carbohydrase cleaves sialic acids from host cell surface proteins so that emigrating virus does not get re-attached to a previously infected cell.

Segment 7 encodes both Matrix 1 (M1) and Matrix 2 (M2) proteins, via splicing of primary transcripts (explored below: 2.3.4). M1 is the most abundant protein in the virion. It is situated just beneath the viral envelope where it binds both the cytoplasmic tails of membrane glycoproteins and RNPs, thus connecting inner core

components to surface proteins (Selzer et al., 2020). M1 interacts with both RNP and Nuclear Export Protein (NEP) and the cytoplasmic tail of M2. M1 may therefore also assist with packaging by recruiting virion components to the assembly site at the host cell's plasma membrane. The M2 tetramer mediates viral unpackaging once the virion is endocytosed by enabling a proton gradient sufficient to cause membrane conformational changes that in turn allows membrane fusion and viral escape from the endosome (Ito et al., 1999).

Segment 8 encodes 2 distinct proteins, again with the help of mRNA splicing. Non-Structural protein 1 (NS1) down-regulates host RNA translation and instead causes the cell to favour production of viral proteins (Chauché, 2017; Clark et al., 2017). It also modulates host cell innate immunity, most notably as an antagonist of cellular interferon-mediated responses to viral infection. Nuclear Export Protein interacts with nuclear transport proteins (nucleoporins) of the host cell, enabling viral genomes to cross the nuclear membrane.

1.2.3 Influenza A Virus Replication

Like other Orthomyxoviruses, EIV utilises a fast, error-prone RNA polymerase throughout genomic replication (Lauring, 2020). The viral life-cycle can be sub-divided into a number of processes, beginning with viral attachment and ending with the budding of new virions.

1.2.3.1 Attachment

Influenza viruses bind sugar moieties called sialic acids on the surface of epithelial cells to initiate infection. Viruses adapted to different species show specificity in the sialic acids to which their HA binds (Kuiken et al., 2006). Haemagglutinin does not exclusively bind a single type of sialic acid; yet preferential binding to certain sialic acid moieties can determine viral tropism and host range (von Itzstein, 2007). Mammalian and avian epithelial cells can present multiple forms of sialic acids in various proportions, across different tissues (Feng et al., 2015; Yang et al., 2022). Mammalian IAV most often has the greatest affinity for sialic acids which are attached to host cell surface carbohydrates by an α2,6 linkage (SAα-2,6-Gal). Epithelial cells lining the upper respiratory tract of mammals usually have higher proportions of SAα-2,6-Gal moieties than cells deeper in the respiratory tree – hence shaping tropism of IAV infections.

Conversely, avian viruses bind to sialic acids with an α2,3 linkage (SAα-2,3-Gal), more commonly found through the gastrointestinal tract of waterfowl than in the respiratory tract (Abdel-Moneim et al., 2010). Due to this, IAV infections in birds lead to GI symptoms & pathology. Moreover, this adaptation to strongly bind α2,3-linked sialic acids presents additional risks if/when avian IAV jump species barriers (Lipsitch et al., 2016). As mentioned above, epithelial cells in mammalian upper respiratory tracts present SAα-2,6-Gal; however, cells deeper in the bronchi and lungs do have α2,3-linked sialic acids. For this reason, infection of mammals with

avian-adapted IAV can lead to more severe lower respiratory disease (Yan & Chen, 2012).

1.2.3.2 Fusion and Uncoating

After haemagglutinin mediates binding to the cell surface, the virion is endocytosed. The low pH within the endosome activates fusion of the viral membrane with that of the endosome in order to remove the coat of the virus. Viral envelope fusion is induced by a structural change in haemagglutinin. Inside the acidic environment of the endosome, HA is cleaved into two proteins: HA1 and HA2. This exposes the fusion peptide at the N-terminus of HA2 which is able to insert itself into the endosome membrane, joining it to the viral envelope. Remaining haemagglutinin subunits then enter the endosomal membrane forcing open a channel, which releases viral RNPs (vRNP) into the host cell cytoplasm. M2 is located sparsely throughout the viral envelope enabling ion channel activity in acidic environs. An influx of protons from the acidic endosome into the virion denatures protein interactions, causing the release of RNP from the M1 matrix layer within the virion.

1.2.3.3 Transcription

After uncoating, genomic segments complexed with NP and the polymerase (viral ribonucleoproteins [vRNPs]) are actively transported into the nucleus by nucleoporins. Incoming negative-sense genomic segments are transcribed within the host cell nucleus. Frame-shifting during the transcription of segment 2 enables access to two alternate open reading frames, leading to the creation of PB1-F2 and PB1-N40 mRNA rather than the mature PB1 transcript.

1.2.3.4 Splicing

Orthomyxoviruses can increase the efficiency of their genomes by encoding multiple proteins from a single gene via an alternative splicing mechanism. Segments 7 and 8 translate proteins from both spliced and unspliced mRNA transcripts. They, however, lack the efficiency of cellular splicing, and must express proteins from both spliced and unspliced mRNA transcripts simultaneously. Controlling the proportion of spliced to unspliced transcripts must be balanced, and there are limited ways in which the virus itself can regulate this. Transcripts can only be spliced inside the nucleus, so increasing the rate at which the unspliced mRNA is exported from the nucleus reduces the rate at which the transcripts are spliced.

1.2.3.5 Regulating Gene Expression

IAV does not need to express every protein at all stages of the replication cycle. Proteins can be produced at different proportions throughout the cellular replication cycle and can mark transitions between stages of replication. Much of the regulation of gene expression is controlled at the translational level and, in some cases, is

partly responsible for cytopathic effects of infection. IAVs can modulate translation of their own genes and suppress host cell protein synthesis.

1.2.3.6 Translation

Translation of viral proteins utilises cellular ribosomes, unlike replication of genome segments, and thus requires viral mRNA to be adapted for cellular translation processes. The heterotrimeric viral RNA-dependent RNA polymerase (RdRp, referred to as RNApol and also replicase) is composed of PB2, PB1 and PA proteins bound to the 5` and 3` ends of the genome-nucleoprotein complex (Dias et al., 2009).

As vRNP are present within the host cell nucleus, the RdRp can bind nearby cellular transcripts. These host transcripts, produced by cellular DNA-dependent RNA polymerase II (RNAP2) have short (10-13 nucleotides) primers attached which is then cleaved by endonuclease activity in the PA portion of viral RdRp (De Vlugt et al., 2018). Primers are then attached to the viral mRNA, creating hybrid virus-host transcripts which are exported from the nucleus and passed to cellular translation machinery. A by-product of this is a suppression of host cell metabolic processes; as cellular transcripts lack the primers necessary for translation, host proteins become less likely to be produced than viral proteins.

1.2.3.7 Genomic Replication

Genomic RNA is replicated from the negative-sense ssRNA genome of infecting viruses. Within the nucleus of an infected cell, RdRp replicates each genomic segment. Complementary RNA (cRNA), a positive-sense ssRNA strand, is then transcribed which complements the original RNA of the genome segment. This cRNA then acts as the template strand for both generating transcripts and replication of the genome.

1.2.3.8 Packaging

Components of the new progeny virions congregate at the apical surface of infected epithelial cells. Transmembrane proteins associate with the cellular membrane in what will become the viral envelope. Non-structural viral proteins and vRNP complexes assemble near the cell membrane and are incorporated into the budding virion.

1.2.3.9 Budding

Once assembled, the virion pushes through the cell lipid bilayer, taking part of the latter to form the viral envelope. While the factors determining viral morphology are still not fully understood, the cellular cytoskeleton, actin filaments and viral M1 and M2 proteins are all known to be implicated in influencing how spherical or filamentous each particle is.

1.3 Influenza Evolution

Influenza virus evolution and diversity is underpinned by a number of features. In addition to a relatively high nucleotide substitution rate (Zhao et al., 2019), the segmented genome structure enables large-scale genomic reassortment. Interestingly, the evolutionary rates of the influenza A genome, as explored by clock models, is not consistent between sub-populations, over the course of infection or even between different segments of the same overall virus (Kühnert et al., 2011). Genomic segments range in size and in the number of proteins they encode. For example, non-synonymous nucleotide changes may be less well tolerated in regions encoding active sites in proteins than those encoding purely structural regions. Additionally, different Influenza proteins experience different selective pressures, the result of which is internal non-structural proteins being more conserved than surface-exposed antigenic proteins. It should also be noted that due to the unusual architecture of IAV genomes, mutations in the overlapping regions of coding sequences could potentially impact two separate proteins.

1.4 Mechanisms of Viral Evolution

1.4.1 Nucleotide Substitutions

After entry into a host cell, IAV begins its replication cycle. The lack of proof-reading capabilities in the viral RNA polymerase is an important source of genomic variation. Substitution rates of IAV vary substantially across hosts and viral strains with rates ranging from 1.35×10^{-3} substitutions/site/year in equine influenza viruses (Murcia et al., 2011), 2.70×10^{-3} in swine IAV (Dunham et al., 2009) to 3.66×10^{-3} in human IAV (Smith et al., 2009). The rapid, error-prone replication of IAV genome segments permits the misincorporation of nucleotides into genes by the viral RNA polymerase. In coding sequence, these point mutations can either be synonymous (with no amino acid change, due to codon redundancy), nonsynonymous (inducing an amino acid change) or nonsense (encoding premature stop codons). As they are more likely to have a minimal (if any) effect on viral fitness, synonymous mutations are less liable to be subject to selective pressures, and therefore less likely to be removed from the population. Nonsynonymous and nonsense mutations may lead to lethal mutations where proteins are made so inefficient that the virus cannot function competitively. The gradual accumulation of selectively neutral point mutations in a population contributes to the phenomenon known as *genetic drift*.

These conclusions are however based the presupposition that mutations act independently; more complex interactions between multiple genomic mutations (genetic linkage) can affect the fate of mutations in unexpected ways. These epistatic relationships between mutations have been observed in both IAV nucleoprotein (Gong et al., 2013) and neuraminidase (Pedruzzi & Rouzine, 2021)

genes as well as across the entire genome (Lyons & Lauring, 2018); applying hitherto understudied constraints and/or allowances upon IAV evolution. Especially in regions subject to strong selective pressures (e.g. antigenic epitopes), mutations that may be otherwise detrimental can be "saved" by such epistasis and maintained in the genome without a decline in fitness (Kryazhimskiy et al., 2011; Lee et al., 2023). The complexity of separating mutations that happen to co-occur from those that display some level of interactivity is difficult but could unlock understanding of viral attempts to escape antiviral-therapeutics or to sustain them while they cross fitness valleys associated with cross-species transmission.

1.4.2 Reassortment

Another source of diversity in an IAV population is the capability for *genetic shift*, which is defined as the reassortment of influenza genomic information (Bountouri et al., 2011). Viral reassortment, the incorporation of genomic segments originating from a different parental viruses into a single progeny virus, allows for a substantial amount of genetic diversity to be generated very rapidly. Reassortment is viable in viruses with segmented genomes. Genetically distinct viruses co-infecting a host cell can mispackage genome segments from one variant into another creating a composite viral genome with one or more non-native genomic segment(s) (Marshall et al., 2013; Vijaykrishna et al., 2015). Reassortment allows viruses to surpass host adaptive immunity much faster than can be done by relying on the accumulation of point mutations alone (Ding et al., 2021). When reassortment leads to changes in haemagglutinin and/or neuraminidase activity of the virus, it is termed an 'antigenic shift'. Reassortment accounts for a large part of IAV pandemic potential; spillover into novel host populations may be possible with the incorporation of proteins with binding affinity to novel hosts (Lindstrom et al., 1998). However, observations of both experimental and natural infections show that diversifying reassortment events are rare (Rabadan et al., 2008). Furthermore, the viruses within a single host are usually sufficiently genetically similar that reassortment events may not lead to gross genomic change; even if, during the assembly and packaging process, gene segments from a non-parental virus can be incorporated into a progeny virus' segments being genetically identical are overwhelmingly high (Lauring, 2020).

Direct recombination of information between unrelated viral segments is possible and may be facilitated by the interruption of viral RNA polymerase during replication and the switch to an alternate template strand on resumption of transcription (Vijaykrishna et al., 2015). However, homologous (the polymerase switches to the same site on both templates) and non-homologous (the polymerase resumes transcription at a different site on the secondary template strand) recombination occurs rarely *in vivo*, if at all. In fact, reports of experimental IAV recombination are generally understood to instead be caused by laboratory contamination (Lefeuvre et al., 2009; Pérez-Losada et al., 2015). Due to lack of evidence both *in vivo* (De et al., 2016) and *in vitro* (Han & Worobey, 2011),

recombination in IAV is expected to have little to no impact on IAV evolution (Boni et al., 2010; Lauring, 2020).

Both recombination and reassortment are predicated upon co-infection of a single host cell by multiple viral particles (Marshall et al., 2013), further decreasing the likelihood of *in situ* occurrences influencing influenza population evolution. As these evolutionary mechanisms depend strongly on co-infection, the viral population size, seeding dose and timing of infections may be considered risk factors for such dramatic evolutionary changes.

1.4.3 Selection

Each segment of the influenza genome is capable of mutational plasticity (that is, the ability to tolerate mutations with limited fitness consequences), and each step of the replication cycle provides an opportunity for nucleotide substitution.

This broad range of plasticity is not without boundaries; genomes with excessive mutations run the risk of encoding unstable, or even wholly ineffectual, proteins leading to the evolutionary fitness of the virion plummeting. The principles of Müller's Ratchet (Chao, 1990; Muller, 1932), displayed in Figure 1.4, show the necessary balance between genomic stability and plasticity. These limits on potential mutational plasticity, termed the error threshold (Domingo & Perales, 2019), nudge mutational capacity into a classic gaussian distribution.

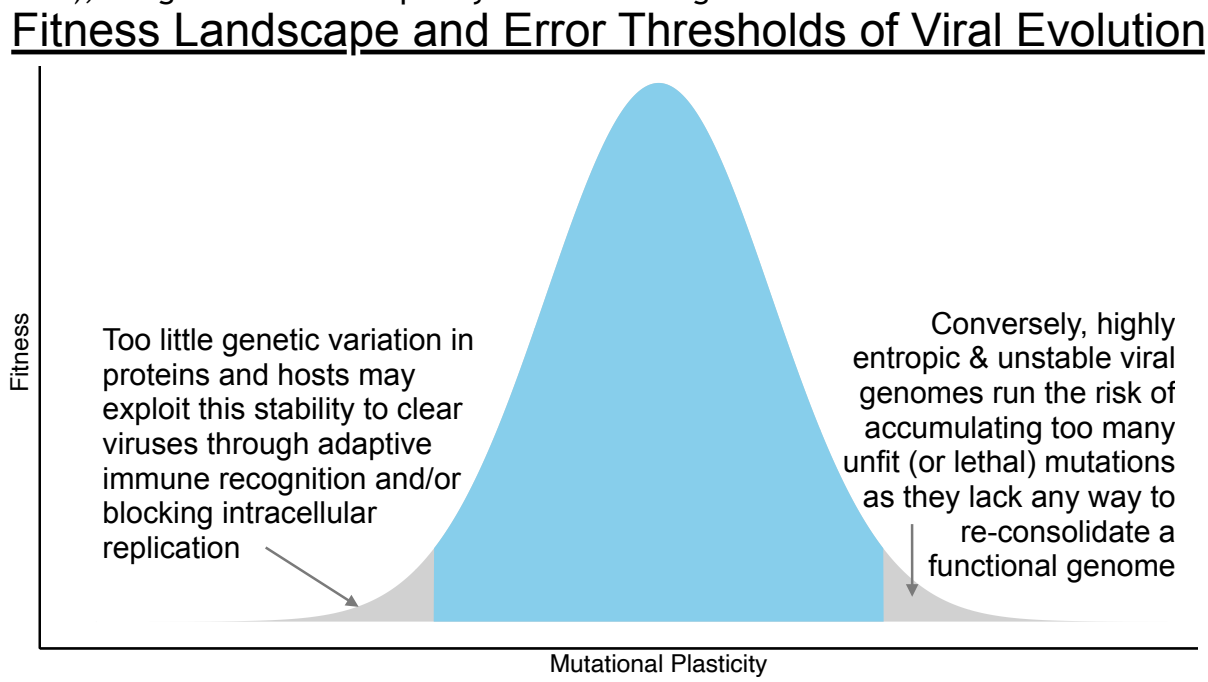


Figure 1.4: A simple explanation of a fitness landscape, wherein fitness of the hypothetical viral feature in question is distributed normally. The blue area shows where in the landscape the virus can compete successfully, the grey tails show where Müller's ratchet begins to remove viruses from the population.

1.5 Antigenic Escape

Influenza viruses most often cause acute infections of their hosts. A population of founder viruses colonises the host, establishes an infection of the respiratory tract and is then transmitted by the individual. Usually, IAV is cleared within approximately 14 days, the time taken to mount an adaptive immune response (Bonilla & Oettgen, 2010). Once B cells have undergone somatic hypermutation and affinity maturation of the B cell receptor, clonal expansion begins; this process usually eradicates all IAV particles from the host. Influenza A must overcome these selective pressures in order to survive.

1.5.1 Protein Structure and Immune Recognition

Influenza A viruses have three transmembrane proteins embedded within the lipid bilayer envelope acquired from the host cell, namely haemagglutinin, neuraminidase and the M2 ion channel (Woodward et al., 2015). Due to its small size (Virmani et al., 2011), relatively conserved sequence (Ito et al., 1991), and positioning (as a transmembrane channel, spatially M2 barely reaches beyond the height of the envelope itself) M2 will be disregarded in the present discussion of antigenic extra-virion proteins. Both HA and NA have distinct 'stalk' domains embedding them within the viral envelope together with 'head' domains, which hold the active sites of both of these cleavage enzymes (DuBois et al., 2011). Classically, these head domains are the targets of cells and molecules of the host's adaptive immune system (Tusche et al., 2012). Consequently, they are under strong selective pressures to change structurally in order to evade targeted neutralisation and removal by the host (Neverov et al., 2015). However, researchers are now seeking to develop immunogens targeting the more conserved stalk domains of these proteins (Arevalo et al., 2020). This approach is aimed at maintaining vaccine efficacy for multiple years, in contrast to the current vaccines which are updated annually in order to account for frequent structural changes in the HA and NA head domains (Flannery et al., 2016).

As discussed above (1.3 Influenza A Virus Replication), HA begins the process of viral entry while NA facilitates release from the host cell. Mature haemagglutinin trimers bind sialic acids on the surface of epithelial cells (Boukharta et al., 2014). The distribution of cells with these carbohydrates differs between hosts. Classically, avian influenza presents as an enteric disease in wild birds since α-2,3-sialic acids are found in the highest concentration in the digestive tract (Lazniewski et al., 2018). Mammalian infections are, instead, localised in the airways due to the abundance of α-2,6-sialic acids on epithelial cells of the upper respiratory tract (Righetto & Filippini, 2018). However, cells presenting α-2,3-sialic acids reside in the lower respiratory tract of mammals and for this reason a persistent IAV infection may broach deeper in the lungs and cause a viral pneumonia.

The neuraminidase tetramer also cleaves sialic acid moieties, though at the terminals of the carbohydrate (Saito et al., 1993). This prevents the virus sticking

to the host cell; as the virion buds, HA will naturally begin to bind and attempt to re-enter the cell it just left. As NA is far less abundant than HA, and is organised with some polarity (Vahey & Fletcher, 2019), sialic acid removal is focused on the part of the virus closest to the progenitor cell. This ensures that the virus is likely to move away from the cell from which it has just budded without interfering with the next viral entry and replication cycle (Chen et al., 2018).

1.5.1.1 Antigenic Drift

The gradual accumulation of point mutations specifically in the epitopes of surface proteins is termed antigenic drift, as the epitope targets to which adaptive immunity was previously protective have now sufficiently changed in conformation as to make newly circulating strains unrecognisable to previously protective antibodies (Rouzine & Rozhnova, 2018). Antigenic drift is facilitated largely by the huge array of genetic variation in a viral population (Poon et al., 2016; Righetto & Filippini, 2018). Mutations within antigen genes encoding proteins such as HA or NA may cause epitopes to change to such a degree that they are unrecognisable, or at least far less liable to binding, by immune cells and molecules (Lee et al., 2016). Strong selective pressures from host adaptive immunity are exerted on HA and NA, leading to a higher evolutionary rate in these genes and particularly within the epitope regions compared to the other six genomic segments (Pauly et al., 2017).
Antigenic drift is also responsible for seasonal influenza. Nonsynonymous mutations to surface proteins occur so rapidly that the binding affinity of adaptive immune cells and molecules originally developed against virus circulating in the previous year is weakened or entirely negated. This necessitates vulnerable populations to receive IAV vaccines annually, as there is no assurance that antibodies from the previous winter will be able to bind IAV sufficiently strongly to grant protective immunity.

1.5.1.2 Antigenic Shift

Antigenic shift is a particular form of reassortment, wherein segments four (HA) and/or six (NA) of the influenza A genome are contributed by a genetically distinct virus and are incorporated into a nascent virion. Thus, a reassortant virus is created with antigenic surface proteins distinctly different from those of the parental virion. If the changes to proteins are sufficiently different, the resulting virus could be completely unrecognisable to the immune memory of ensuing hosts. As the host in which the reassortant virus developed (sometimes referred to as the mixing host) must necessarily be co-infected by two or more distinct IAV subtypes, we would expect them to develop immune responses to both parental strains. On transmission to another host however, the presentation of unrecognisable surface proteins may require the development of completely novel adaptive immune responses; additionally, potentially contributing to the immunopathology and cytokine storm syndromes often reported in infections with newly-emergent IAV.

Though antigenic shift events are rare, when they do occur it often involves a collision of host reservoir populations. For example, the 1918 H1N1 pandemic was

the result of reassortment of segments from avian-adapted IAV into human-adapted IAV, possibly within a swine intermediary host. Swine are susceptible to a range of influenza A viruses and so can act as 'mixing vessels' (Canini et al., 2020; Lewis et al., 2016). Following this reassortment event, the virus remained adapted to human hosts but now contained avian surface proteins that were unrecognisable by any previously existing adaptive immunity in the human population.

Most of the four human influenza pandemics which have occurred since the turn of the 20th century ("Spanish Flu (H1N1)": 1918, "Asian Flu (H2N2)": 1957, "Hong Kong Flu (H3N2)": 1968 and "Swine Flu ($H1N1_{pdm09}$)": 2009) have been as a consequence, at least in part, of antigenic shift (Figure 1.5). A novel H1N1 virus emerged in humans in the US around 1918 which contained avian genomic segments. Reassortment of this virus with other avian IAV resulted in virus with novel segments 4 and 6 (H2N2) leading to the pandemic of 1957, which later shifted in 1968 with another novel haemagglutinin gene (H3N2). Finally, the 2009 "swine flu" outbreak originated with an entirely novel H1N1 reassortant virus composed of swine, avian and human genomic segments.

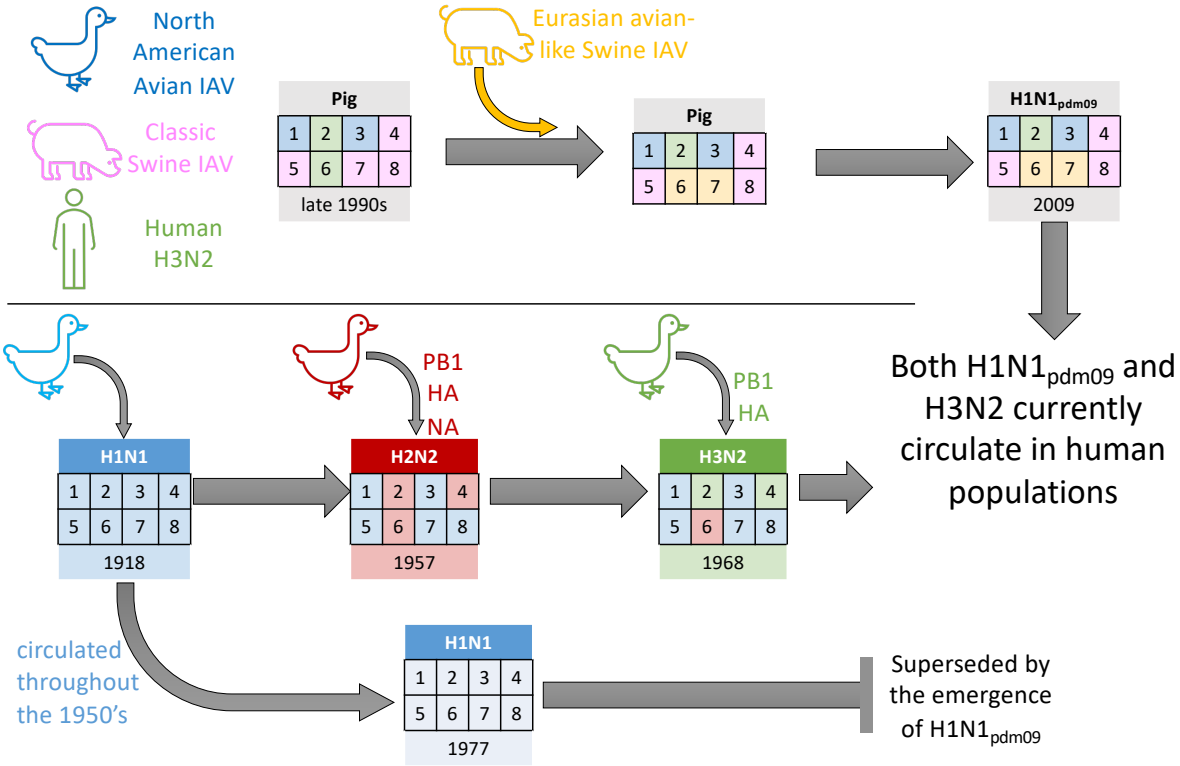


Figure 1.5: Viruses from human influenza pandemics since the beginning of the 20th Century. To highlight the importance of reassortment, viral genomes are shown as grids; with genomic segments 1-8 coloured according to the host from which they originated.

1.5.1.3 Original Antigenic Sin

The anti-IAV host immune response may also be hampered by previous exposure history, according to the theory of Original Antigenic Sin (OAS). This theory, put forward initially by Francis et al. (1960), concerns adaptive immune recognition of influenza A virus strains. When first exposed to IAV, immunocompetent individuals

will mount innate and subsequently adaptive immune cascades resulting in the development of (among others) memory B cells with corresponding epitope-binding antibodies specific to the infecting influenza strain. These memory responses enable rapid re-activation of adaptive immunity should the immunogen appear in the body again. Original Antigenic Sin hypothesis then contrasts the over-simplification that novel memory responses are generated for each new pathogen encountered. Rather than undergoing new clonal selection processes, OAS holds that the "good enough" binding of previous influenza memory cells will forego generation of a novel B cell repertoire and instead reactivate an adaptive memory cascade reusing existing memory cells (Monto et al., 2017). This set of imperfectly matched antibodies would then be capable of binding to pathogen epitopes, but at a reduced efficiency compared to antibodies generated during the primary exposure to influenza. Humoral immune responses, such as opsonisation and neutralisation, would be unable to bind altered viral proteins with the same strong affinity as they did to the Original Antigen.

OAS is sometimes referred to as "antigenic seniority' (Henry et al., 2018), indicating the bias of the immune system towards the first IAV strain encountered. Original Antigenic Sin theory predicts that the strength of an adaptive immune response to a completely novel influenza strain may in fact be stronger and more protective than the response to an IAV strain that only moderately differs from one to which the individual has pre-existing immunity. The theory has been contentious since its proposal, but evidence by Rioux (2020), Gostic (2016, 2019) and Simonsen et al. (2004), among others, sought to associate first influenza exposure (the eponymous *Original Antigen*) with weakened responses to related but distinct influenza strains. The effects of OAS also apply to vaccine-mediated immunity; hosts primed with a vaccine immunogen may be granted protection from that specific strain, but may have a weaker response to similar IAV.

1.5.2 Within-Host Evolution

The viral population infecting a single host is rarely genetically homogeneous (Rozek et al., 2021). Though an overall consensus genomic sequence may be established, virions containing variant sequences will likely be present, be they replication-competent or not. In an individual host, the viral diversity may be purely entropic or be biased towards certain genotypes. As the viral population diversifies in the course of an infection, variants will be subjected to competition (Bessière & Volmer, 2021) and must be able to either outcompete or survive alongside other IAVs descended from the original donor population. Within-host variance of pathogens (Duxbury et al., 2019; Grubaugh et al., 2019) generates genetic plasticity of the virus, and host-pathogen interactions shape this plasticity to influence viral population demographics (De Fine Licht, 2018).

Influenza viruses must surpass multiple host barriers in order to establish an infection; these in turn shape viral evolutionary patterns by providing selective pressures (Balasuriya, 2020; Diskin et al., 2020). Initially, the virus must find a

suitable environment in which to replicate. Beyond the spatial elements of establishing an infection, viruses must also counteract or avoid the host immune responses (Xue et al., 2017, 2018). The resultant pressure which is exerted on the virus represents a significant driver of selection for antigenic escape of viral surface proteins. Except in cases of cross-species spillover, viruses generally demonstrate at least moderate adaptation to their hosts and specificities for tissues, which is mediated by the range of competent cells (Mendenhall et al., 2019; Moustafa et al., 2017). This enforces a spatial structure within the host; influenza viruses adapted to mammalian hosts often have a tropism for the respiratory tract in contrast to many avian influenza viruses which may instead infect cells of the waterfowl digestive tract (Kratsch et al., 2016). As different viruses which have established successful infections of their hosts are in the same spatial environment, they may experience evolutionary processes such as gene reassortment (Wasik et al., 2019).

One of the biggest selective pressures acting on populations of viruses within an infected host is that of immune responses, classically those of the adaptive cell-mediated and humoral responses, but also potentially from innate immune cells (Oxburgh & Klingeborn, 1999). Individuals vaccinated against IAV may still be capable of hosting asymptomatic infections, as reported for humans in the FluWatch study which recorded almost 75% of infected persons display no symptoms (Hayward et al., 2014), and the viral population within these hosts is subject to immune pressures that are likely to drive antigenic escape. Additionally, antiviral therapies often attempt to interrupt viral replication cycles, and vaccines are designed to stimulate immune responses much faster than natural immune activation cascades (Spielman et al., 2019). Antivirals place powerful selective pressures on viral communities, stimulating them to evolve evasion mechanisms (Sunayana, 2019). Most IAV antivirals, such as zanamivir and oseltamivir, are competitive inhibitors acting on neuraminidase (Das et al., 2010). Like the selective pressures caused by host immunity, conformational changes to viral proteins can arise as viruses attempt to evade impediment by antiviral drugs (Lazniewski et al., 2018; Magori & Park, 2014; von Itzstein, 2007).

1.5.3 Between-Host Evolution & Transmission Bottlenecks

With the development of high-throughput sequencing and metapopulation genetics, the level of genetic diversity of both intra- and inter-host pathogen populations can be more clearly determined than with previous sequencing techniques. The genetic diversity generated in multiple hosts is conducive for global antigenic drift among other potentially beneficial mutations (Rodríguez-Nevado et al., 2018; Simmonds et al., 2019). Selective pressures experienced by viruses undergoing transmission bottlenecks shape the overall epidemic viral population and help determine which mutations become fixed in the broader viral population. Work on vesicular stomatitis virus (VSV) (Elena et al., 2001), however, has shown that though VSV population size increases with the number of susceptible hosts in the environment, the size of bottlenecks in each transmission event remains relatively

consistent. Characteristics of transmission bottlenecks are shaped by both genetic and ecological host-pathogen interactions including, but not limited to, host contact patterns, mode of transmission and the presence of a competing microbiome (Armero et al. 2021, Bendall et al. 2023).

Though the mutational spectrum within an infected host is broad, inter-host diversity is highly dependent on the transmission bottleneck. A donor host will shed a finite quantity of viral particles and, even in directly transmitted infections, only a limited number of these particles establish infection in a recipient host (Poon et al., 2016). Maintaining a fully representative picture of population diversity through this bottleneck is difficult, but the mutational spectra from a donor host and in a recipient host can be observed and compared to understand the viral genomes that survived transmission intact and infer characteristics of the bottleneck itself (Sobel Leonard et al., 2017). Comparing the population diversity before and after a transmission event can highlight the challenges that viruses must overcome in order to establish new infections. If the genomes of viruses in the donor and recipient share significant levels of identity, this implies that only a few viral particles were able to survive the transmission event. This is further complicated, however, by the fact that the viral seeder population can be unrepresentative of the viral population within the donor host.

Furthermore, the size of transmission bottlenecks can have strong influences on the forces of evolution acting upon viral populations. Smaller viral populations are much more susceptible to stochastic changes than larger populations, which may maintain some of their diversity through a transmission event (Lauring, 2020). Studies have shown that while bottlenecks can preserve transient variants (Stack et al., 2013), transmission bottlenecks themselves are unlikely to drive viral evolution, unless the transmission event itself applies strong selective pressures (i.e. encountering vastly different host environments, as in host jumps) (Varble et al., 2014). Instead, selection is proposed to occur in infected recipients. The impacts of severe bottleneck restrictions mirror Müller's ratchet, wherein the stochastic loss of IAV virions is most likely to remove the most virulent genomes from the viral population (Bergstrom et al., 1999).

To conclude, viruses face within-host challenges, such as competition and immune evasion, punctuated by population re-structuring caused by (possibly unrepresentative) sub-sampling during transmission events.

1.5.4 Mutant Spectra

Next-generation sequencing technologies and genome assembly bioinformatic processes are increasingly sensitive, able to detect and exclude the majority of sequencing errors; this enables sub-consensus mutations to be recognised with sufficient confidence that the variation detected is not generated by erroneous sampling (McCrone et al., 2020). However, distinguishing viral mutations present at very low frequencies from sequencing errors will likely remain problematic until a 100% accurate sequencing methodology is developed. Variant genomes present at

only 1% proportion within a single host can be reliably detected and analysed (Xue et al., 2018). Following these variants throughout the course of disease in an infected individual can help infer transmission trees (Campbell et al., 2018; De Maio et al., 2018). As genetic sequencing technologies further improve, the ability to explore the dynamics of viral diversity within hosts is expanding, constrained only by the capacity to distinguish technical errors from true mutations. Methods to track the evolution of viral populations in a single infected host primarily rely on serial sampling and sequencing of genetic material (Watson et al., 2011).

A mutant spectra, alternatively termed a 'viral cloud', describes the total range of genetic variants within a particular viral population. Note, this is distinct from the 'Pan Genome' concept of bacterial genetics which details a shared genetic structure with additional 'disposable' genes that are not present in all individuals across a species (Rouli et al., 2015; Tettelin et al., 2005). However, in rapidly evolving viral populations (usually, but not limited to, RNA viruses) an array of point mutations can emerge. While many of these may be neutral or even deleterious, some have the chance to be beneficial.

The mutant spectra present in an infected individual can have a range of clinical and public health repercussions. The genetic diversity generated in multiple hosts provides the tools for global antigenic drift among mutations causing other potential phenotypic changes that can then be selected for/against (Rodríguez-Nevado et al., 2018; Simmonds et al., 2019). Selective pressures enacted upon viruses undergoing transmission bottlenecks shape the overall epidemic viral population and determine which mutations become fixed in the broader viral population.

Pathogenicity is one of the key issues to consider when discussing viral evolutionary and population dynamics (Oakeson et al., 2017). To quote Holland et al. (1992), "The acute effects, and subtle chronic effects, of infection will differ not only because we all vary genetically, physiologically and immunologically, but also as we all experience a different array of quasispecies challenges". The emergence of bacterial anti-microbial resistance (AMR) provides a clear example of the clinical impact of broad mutant spectra in a pathogen population. AMR and other related phenomena, such as anthelminthic resistance, originates when a challenge (i.e. an antimicrobial) is applied to a pathogen population. The two-fold effects of placing pathogen populations under such strong selective pressures and simultaneously eradicating the majority of competing strains creates an 'easy to exploit' ecological niche for any mutant strains able to resist the antimicrobial compound. Examples of this are perhaps best displayed in the field of HIV and the Highly-Active Anti-Retroviral Treatment (HAART) required to combat the emergence of drug-resistant strains. Though unlike influenza, HIV causes chronic infections, the breadth of diversity generated in sub-consensus mutants within both viral populations presents a range of variants with possible drug-resistance phenotypes capable of emerging.

1.5.5 The Ever-Elusive Quasispecies

Quasispecies theory in viral population dynamics contends that the broad array of genotypes that comprise the overall pathogen population within an infected individual work in concert to generate genomic plasticity (Domingo & Perales, 2019). Rather than considering the range of co-infecting viral genomes as being independent entities, quasispecies theory dictates that all these genomes are the subject upon which the mechanisms of selection act (Gregori et al., 2016) and that intra-host genomic diversity is necessary for virus survival and evolution. Quasispecies theory depends on the infecting population behaving as a singularly evolving unit, a concept which has hitherto been difficult to prove. To date, no empirical evidence has been found that the mutational spectrum, that is to say the range of replication-competent genomes within a system (whether in a single host or a group of epidemiologically-linked hosts, as long as the viruses are able to interact and compete with one another), substantially impacts the fitness of a viral population (Geoghegan & Holmes, 2018) and so quasispecies dynamics in IAV infection remains a conceptual notion. The influenza viruses within a host contain naturally stochastic mutations which are subject to selection; this does not mean that the population as a whole is experiencing selective pressures as a single evolutionary unit. Sub-consensus variants of IAV reflect and provide evidence for within-host diversity, which in turn facilitates the creation of further diversity thereby shaping the overall viral population. Importantly, these impacts upon viral populations are likely caused by evolutionary forces acting independently on viral genomes rather than cohesive forces acting upon the population in its entirety.

The literature on the concept of quasispecies is ever-expanding as sequencing technologies improve in terms of read length and depth of coverage. We have come to recognise that a single consensus sequence is often unrepresentative of a measurably evolving population of pathogens (Biek et al., 2015; Meinel et al., 2018). Though helpful for observing epidemic-scale dynamics of pathogens, the simplifying assumptions of a consensus sequence approach prohibits us from comprehensively evaluating population dynamics on both an inter- and intra-host scale (Hapuarachchi et al., 2016). While a diverse viral genetic composition can now easily be observed within-host, the causal relationship between this mutational spectra and population-level selection remains to be demonstrated.

Though the quasispecies theory was first put forward in the late 1990s (Domingo et al., 2017; Kim et al., 2016), the scope of research on this issue was limited until Next-Generation Sequencing enabled deep-sequencing strategies to reliably detect a range of minority variants, including single nucleotide polymorphisms (SNPs), within a sample (Baele et al., 2016; Jones & Good, 2016). As viral evolution has come to be better understood on both a small, within-host scale and a large, epidemic scale, it has become apparent that the evolutionary dynamics that act upon mutant spectra can shape the pathogenicity of viral populations (De Maio et al., 2018; Hidano & Gates, 2019).

Influenza A viruses are renowned for their relentless evolution together with their proficiency for host immune evasion. However, the potential for cross-species transmission drives the fearsome reputation of these viruses; already broadly disseminated through avian and mammalian host species, the homogenising effect of globalisation opens new opportunities for different hosts to mix in close quarters thereby creating a conducive environment for novel host adaptations.

1.6 Viral Ecology

Influenza A Viruses have a known propensity for their wide host range and cross-species transmission potential. Their ubiquitous association with a broad range of birds and mammals is well documented (Chen et al., 2009). Understanding the consequences of mutations is challenging, however, as even apparently 'neutral' mutations that fix in the population may have some unknown property that encourages their maintenance in a population. As hinted at in the discussion of virus sub-typing (Figure 1.2), although highly distinct IAV can be found in bats, host-specialisation may be observed in a range of species. Although almost all non-chiropteran influenza strains have been recorded in, and are believed to have originated within, waterfowl, influenza viruses have been found in many other endotherms.

Birds may be the source of many IAV strains and are expected to be involved in maintaining the virus as a viral reservoir (Cleaveland et al., 2007; Haydon et al., 2002). While seasonal influenza of humans does not require constant re-introduction from avian hosts, a spillover event from any non-human host may be considered a risk factor for pandemic emergence of influenza strains novel to humans. The interconnectedness of IAV host populations is likely to be even more complicated than currently understood; for example, H3N8 viruses circulate in avian, equine and canine hosts. Sometimes viruses transmit between these hosts, such as equine-canine transmission, while at other times they circulate exclusively in a single host. It should be noted that, in addition to the above-mentioned hosts, H3N8 viruses have also been detected spuriously in swine, phocine and human populations, suggesting the possibility of viral spillover, distinct and unrelated to transmission cycles to those of endemic equine or canine H3N8.

1.7 Viral Ecological Interactions

Importantly, like all biological processes, the viral evolutionary mechanics discussed above do not happen in a vacuum. IAV is not only interacting with host cells but also with other microbes present in the host respiratory system. Spatially, IAV spread within hosts is localised, meaning that virions are in constant communication with the rest of the local influenza A population (Gallagher et al., 2018), thus enabling co-infection at the single-cell level. Experiments with H3N2 influenza variants showed that highly distinct variants benefitted from virions more closely related to the population consensus. If a rare variant infects a host cell, the secondary, co-infecting virion is likely to be distantly related to the rare variant and

thus, any defects or less-competitive mutations in the rare variant are “rescued” by the fitter, more conserved secondary superinfecting virion (Leeks et al., 2018). This negative frequency dependence can actually facilitate the maintenance of high levels of diversity and even the persistence of unfit variants in the population.

A study of swine influenza observed nonsense mutations of IAV genotypes within pigs which could still be transmitted from animal to animal (Murcia et al., 2012). The authors suggested this maintenance of presumably deleterious mutations was possible through trans-complementation. Replication of viral RNA begins with complementary positive-strand RNA (cRNA) which serves as the template strand in genome replication. Experiments have shown that non-parental polymerases can replicate genomic RNA *in trans* and become incorporated into progeny vRNPs, however transcription was only reported *in cis* (Jorba et al., 2009). Effectively, during co-infections viral genomic material can be replicated by the polymerase complexes of any other IAV – but the capping and polyadenylation processes (and therefore transcription) can only be carried out by polymerases closely resembling (or originating from) those of parental virions.

Hence, otherwise deleterious mutations, which would severely impede viral fitness, could arise if complementary proteins from co-infecting viruses are present. This, however, would only occur with any regularity if co-infection of single host cells during infection was a frequent event.

Hosts co-infected with IAV and other pathogens can support a range of interactions. Virus-virus relations can be competitive; for example, Dee et al. (2021, 2022) showed that IAV inhibits SARS-CoV-2 replication. Influenza infections can also suppress common cold viruses by activating host immune systems. This has even been implicated in disconnecting the seasonal circulation of rhinovirus from that of IAV. There is also emerging evidence of the hybridisation of viral particles during IAV and Respiratory Syncytial Virus co-infections (Haney et al., 2022). This mutually beneficial relationship shows very different viruses interacting not just ecologically, but molecularly and structurally.

Virus-bacterial interactions can also be mutualistic. As discussed above, much of the mortality associated with influenza A infections in immunocompetent people is caused by secondary bacterial pneumonia. In samples from hosts infected with *Streptococcus pneumoniae*, those individuals co-infected with a respiratory virus had consistently higher *S. pneumoniae* loads (Shrestha et al., 2013). This is not to suggest any symbiosis, but to highlight the multiple viral, bacterial and host players in the ecological system.

A phenomenon enabled by population-level evolution is that of the maintenance of unfit, or even ‘lethal’, mutations. Whereas a single viral particle may suffer due to these detrimental mutations, or even be replication-incompetent, piggybacking on fitter viruses within the same infection locus may allow defective viruses to fulfil replication cycles. Additionally, not all ‘fatal’ mutations actually prevent the virus from replicating, since these otherwise nonsense stop codons may be substituted by correct, functioning copies of the genes from co-infecting viral

particles (McCrone et al., 2018; Schönherz et al., 2016). Clearly the full complexity of understanding the genetic diversity of even a small viral outbreak can reveal a great deal of information about the interplay of viral populations within and across hosts (and potentially vector) populations.

Even potentially deleterious mutations in antigenic regions may be complemented by replication competent viruses, enabling otherwise lethal mutations to persist in the population and grant additional immune escape functionality to the mutant spectra. In this way detrimental, uncompetitive mutations can be maintained in a population without being purged – they effectively escape selection. Fatal mutations prevent the virus from replicating independently, however co-infecting viruses, may provide correct, functioning copies of the genes which defective viral particles can use to substitute their own fatally-flawed proteins. (McCrone et al., 2018; Murcia et al., 2012; Schönherz et al., 2016).

1.7.1 Impacts of Transmission Bottlenecks

Transmission bottlenecks of acute viral diseases can vary greatly in size and composition; this may impact epidemic and clinical outcomes across a range of scales. At the individual scale transmission bottlenecks can be a large determinant of whether the recipient becomes infected or not. A key determinant of the size of a transmission bottleneck is the transmission route of the pathogen. Aerosol transmission for example, is associated with more stringent bottlenecks whereas pathogen spread through direct contact via blood generally allows for a greater number of viral particles to pass to the recipient host (Varble et al., 2013, 2014). Vaccines place significant selective pressure on pathogens and therefore the transient bottleneck population can be shaped or distorted through this specificity funnel (Bessière & Volmer, 2021).

When observing bottlenecks at epidemic scales, SNV that appear transiently can be used to reconstruct transmission chains (Klinkenberg et al., 2017; Skums et al., 2018). Conversely, if viral populations in two hosts both developed the same point mutation despite no epidemiological contact, we infer that the site of this mutation is hypervariable and/or phenotypically relevant (Biek & Real, 2010). Furthermore, with descriptions of the sub-consensus variants present across cohorts, features of the transmission bottleneck such as size (how many viral particles pass through) and stringency (how diverse are the viral particles that pass through) can be characterised (Ghafari et al., 2020).

Anthropogenic behaviour surrounding host movements may disconnect epidemic network associations from geographic networks, establishing global transmission chains as seen in detail during the SARS-CoV-1 epidemic (Riley et al., 2003). At local levels, transmission bottlenecks are at least partly shaped by the viral population and the density of virions in infected hosts (Zwart & Elena, 2015). A greater number of viruses present in tissues, especially tissues related to transmission (such as nasal mucosa for droplet transmission), simply increases the chances that any transmission event will include more infectious viruses and thus

broadens the size of bottleneck populations. Interference from human activities can also establish contact links that were otherwise impossible, enabling unexpected IAV spread. Alternatively, globally-reaching organisations such as WHO (World Health Organisation) or WOAH (World Organisation for Animal Health) make concerted efforts to manage diseases in humans and animals, providing similar challenges to viruses around the world.

1.7.2 Transmission Phenotypes

In addition to the array of mutations generated entropically through replication, some viruses display phenotypic shifts before, during and after transmission events. Recorded in some highly host-specified viruses, such as HIV, the transmission population differs from the general phenotype present through the rest of the course of infection (Kariuki et al., 2017; Maeda et al., 2020). Once a new infection has been seeded, the transmission phenotype of the virus may not be suitable for continued infection of the host (Domingo et al., 2017). Indeed, the recipient host may present novel selective pressures to the virus that weren't present in the donor host (Illingworth et al., 2020; Theys et al., 2018); potentially causing viral populations between donor and recipient hosts to diverge drastically (Yu et al., 2018).

Studies of HIV spread between hosts have shown that the viruses involved in the transmission event have a significantly different demographic composition (Lazarus et al., 2016). The biased transmission populations have an increased resistance to type-I interferons and preferentially bind CCR5 receptors on host cells, both adaptations to initiating an infection which are downregulated later in the infection process (McCrone & Lauring, 2018). Though such stark adaptations have not yet been discovered in influenza viruses, the density-dependent spread of IAV (opposed to HIV's frequency-dependence) could understandably be assumed to drive similarly selective processes for phenotypic differences depending on the stage of influenza infection.

Though a distinct transmission phenotype has not been recorded for EIV (or any other IAV) studies show (Domingo, 2020) that the genotype composition of a viral population may distinctly specialise around transmission events. Assumptions of consensus sequences can often ignore mutations that occurred within an individual host but also neglects the non-random emphasis of certain phenotypes to be chosen in transmission processes, as seen in HIV and parvoviruses (Voorhees et al., 2019). Transmission-specific phenotypes have not been reliably observed in acute IAV infections. However, studies into the morphology of influenza virions have suggested structural pleomorphism to coincide with different stages of tissue colonisation, tissue infection and viral replication (Seladi-Schulman et al., 2014; Vahey & Fletcher, 2019).

1.7.3 Genomic Memory

Memory genomes have also been reported alongside quasispecies in viral populations (Domingo et al., 2017). Experimentally passaged FMDV will non-stochastically revert from a diverse genomic population back to a population more closely resembling the initial inoculating population (Morelli et al., 2013). Some have proposed a hitherto unseen selective force driving the quasispecies away from endlessly divergent mutations (Firestone et al., 2020). The statistically significant proclivity of a FMDV quasispecies to converge back towards the genomes of previous lineages prompts further investigation into the causes and effects of this biased evolution (Xue & Bloom, 2020). It is however yet to be seen in influenza viruses. Further, as discussed above, IAV antigenic proteins change so rapidly that hosts are able to be re-infected on an annual basis – a function to continually revert back to previously circulating genomes would deprive influenza viruses of one of their greatest adaptations, and so would be expected to be quickly purged from the population.

1.8 Equine Influenza

Equine influenza is a veterinary disease of global importance. Economic losses, mainly from racehorses & thoroughbred breeding, can be dramatic (Yongfeng et al., 2020). As an example, the initial detection & isolation of H3N8 EIV in Australia, 2007 was estimated to cost AUD$3.35 million per day (Callinan, 2008); adjusted for inflation, AUD$5.26 million daily at time of writing. Globally, horses fulfil a variety of economic purposes, an example of this is shown in Figure 1.6, where the net import-export numbers of horses are shown (as a proxy for economic importance) for each country. Outbreaks in other equids can also burden communities; donkeys are working animals with critical socio-economic roles in West and Central Africa so reports of EIV spreading across the region are alarming for many (Adeyefa et al., 1996; Diallo et al., 2021).

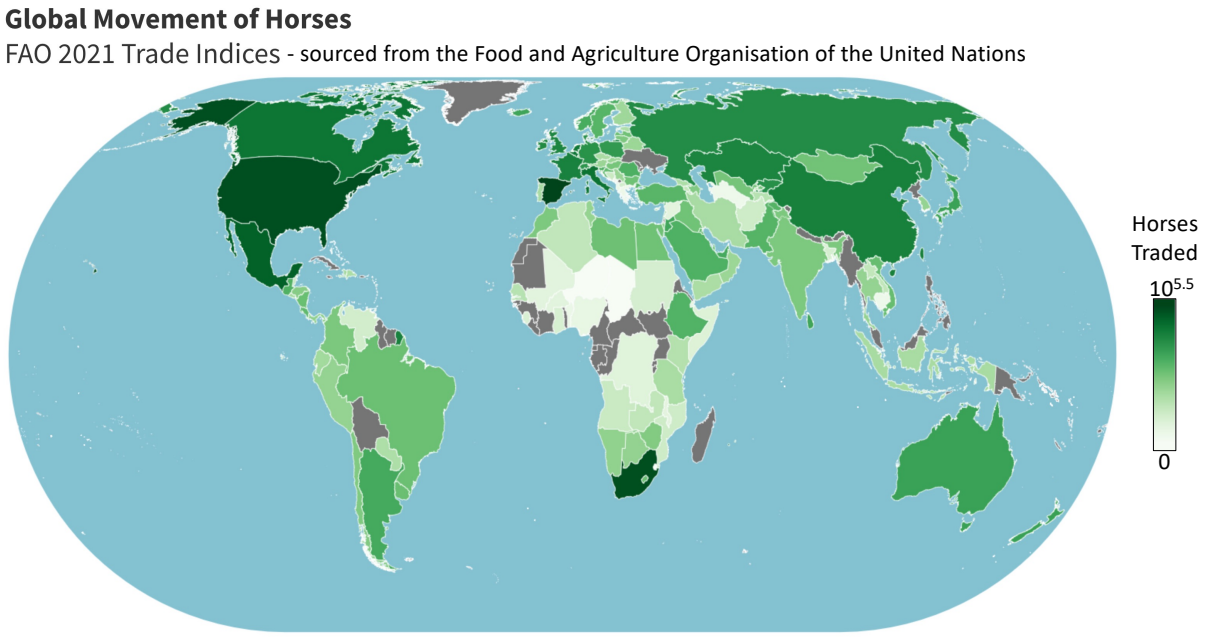


Figure 1.6: Distribution of horse imports and exports in 2021. These trade records approximate the economic importance of horses globally (www.fao.org/faostat/en/#data/TCL).

1.8.1 Impacts of Equine Influenza

Much like human influenza, seasonal endemic EIV occasionally breaks out to epidemic or pandemic proportions (Yondon et al., 2013). In 2019, a large EIV epidemic occurred in the UK, causing millions of lost income (Oladunni et al., 2021). Furthermore, the infection of vaccinated horses indicates that vaccine efficacy is insufficient to wholly prevent infection.

Equine Influenza Virus (EIV) is also capable of jumping into other species, most notably canines but may have additional cross-species potential (Zhu et al., 2019). As seen in other epizootics of livestock, human-mediated transport and events can facilitate and streamline the spread of pathogens (Biek et al., 2015; Theys et al., 2018). EIV displays many characteristics of IAVs in other mammals, and to date the dynamics seen at all scales of equine influenza pathogenesis are broadly applicable to IAVs in other mammalian hosts.

Clinically, equine influenza presents similarly to human infection including fever & respiratory difficulties (Toh et al., 2019), characterised by a high morbidity but a low mortality rate which is driven almost exclusively by secondary bacterial pneumonia (Dunning et al., 2020). Transmission is droplet-mediated and in close-quarter stables, EIV can easily spread between horses. Numerous outbreaks have documented infection of vaccinated horses, implicating insufficiently protected horses as potential spreaders, even when asymptomatic (Back et al., 2016). Important to acknowledge are the sampling and recording procedures around reporting EIV outbreaks. Symptomatic horses are over representative of current EIV sequence samples. Like many sub-clinical, acute infections overcoming this sampling bias is currently unrealistic; regular collection of high-quality viral genomes from horse populations would require intense manual labour and constant sequencing

procedures. However, semi-regular sampling of a small but representative subpopulation from a herd is theoretically possible though may again be biased towards higher income farms where non-emergency veterinary work can be afforded.

1.8.2 History of Equine Influenza

Equine influenza has a surprisingly long association with humanity since the domestication of the horse. Though the medical historiography is sparse, records of outbreaks of disease in military, economic or agricultural horse populations are consistent throughout European chroniclers (as illustrated in Figure 1.7); influenza-like illness in horses has been noted since antiquity (Khan et al., 2021). Additionally, the links between horse outbreaks preceding spread of human influenza were remarked upon as early as 1688 (Morens & Taubenberger, 2010) with chroniclers noting "In October an influenza began among horses and then attacked men as usual''. The previously proximity of horses and humans in rural European life may have explained these now uncommon zoonotic transmission dynamics. Despite the recognition of co-occurrence of equine and human influenza outbreaks in many instances across European history, these outbreaks never display the sociological impact or lasting memory observed with many other epidemics through history (Cohn, 2020; Rosenberg, 1992).

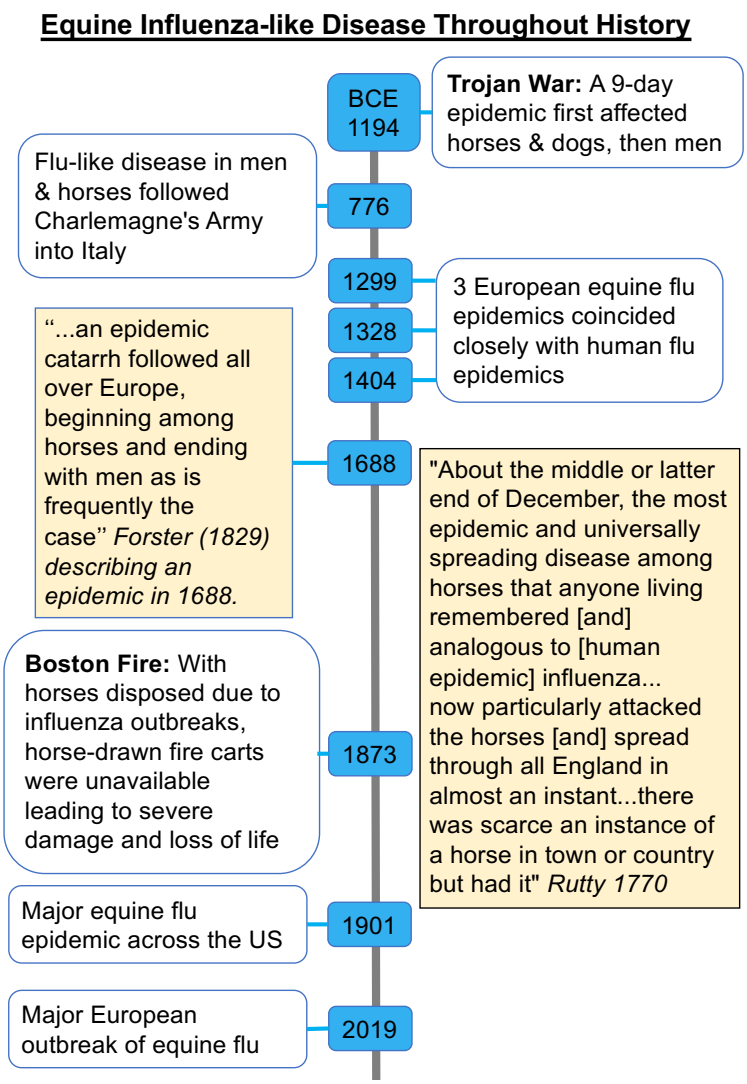


Figure 1.7: Some notable EI-like outbreaks through history - data sourced from Morens (2010). Quotes from Forster (1829) and Rutty (1770) describe recorded associations between outbreaks of equine and human respiratory disease.

1.8.3 EIV Evolution

EIV was first isolated in the mid-20th century in Prague, then the 4th Czechoslovak Republic, after notice in the equine population (Sovinova et al., 1957, 1958). Three IAV subtypes have been transmissible between equid populations:

H7N7, H3N8 and a divergent H3N8 subclade localised to Jilin, China (Daly et al., 1996; Lai et al., 2001; Lewis et al., 2011). However only H3N8 has been seen in natural environments since the 1970s; H7N7 presumed to be extinct (Harvey et al., 2016). Hence, unless otherwise stated, throughout this review all mentions of EIV will refer to the global H3N8 subtype only. Indeed, H3N8 has endemically seeded almost every country with equine populations except for Iceland and New Zealand; with Australia having cleared EIV after a brief introduction (Olguin-Perglione & Barrandeguy, 2021).

Modern, currently circulating EIV is assumed to be an avian-origin virus and the MRCA is estimated to have emerged in the middle of the 20th Century (Chambers, 2020; Murcia et al., 2011). Though unproven, this is a sturdy assumption; almost all IAVs can be phylogenetically traced back to avian influenza viruses (Yoon et al., 2014). From initial detection in 1963, H3N8 fell into either a Eurasian or American clades. Globally, EIV has now split into 4 clades: South American, Kentucky, Florida 1, and Florida 2 (Nemoto et al., 2021). Florida clade 2 has now further diverged into European and Asian subclades (Landolt, 2014; Legrand et al., 2015; Müller et al., 2009). Readers are encouraged to look to the explanatory figure in Chambers' (2020) discussion of Equine Influenza evolution, adapted here as Figure 1.8.

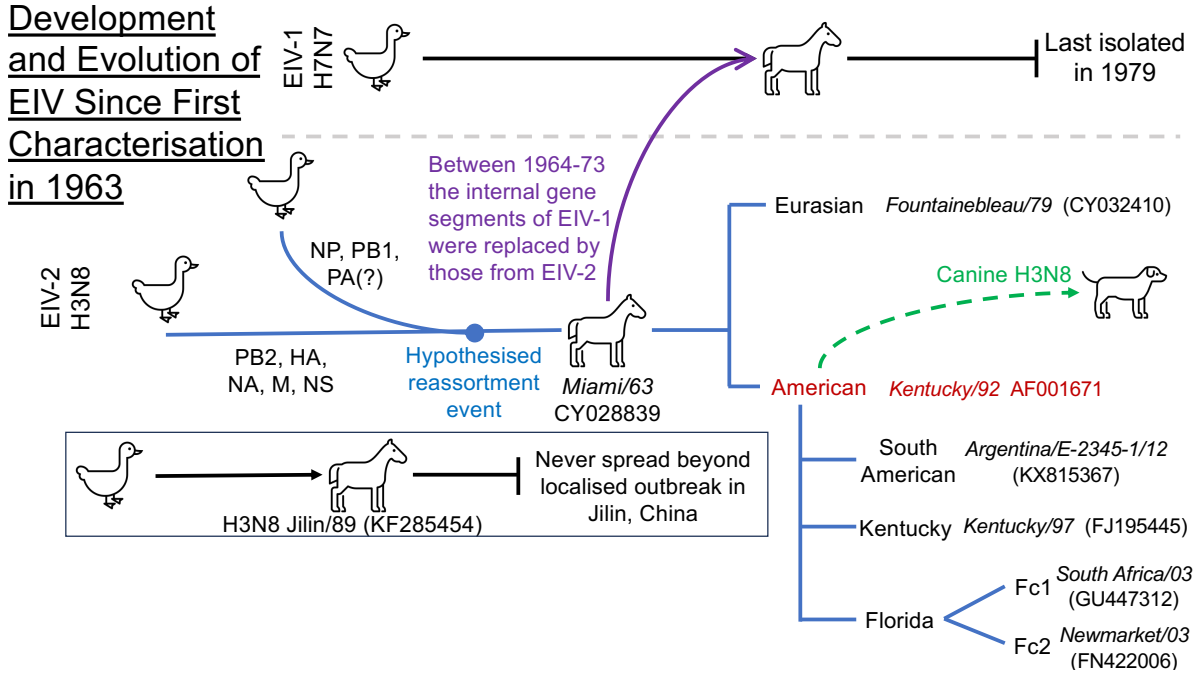


Figure 1.8: Proposed ancestry of current EIV strains. Strains representing H3N8 divergence events are given in italics with their associated GenBank accession numbers.

Much like the human IAV vaccine, the World Organisation for Animal Health (https://www.oie.int/en/disease/equine-influenza-2) regularly updates the antigenic composition of commercially available EIV vaccines. Currently, the World Organisation for Animal Health (WOAH) recommends that the EIV vaccine contains representative strains from both Eurasian (Florida 1) and North American (Florida 2) subclades (Olguin-Perglione & Barrandeguy, 2021). Vaccines are recommended to be administered routinely and boosted every 6-12 months, though in outbreak scenarios boosters may be given pre-emptively. Much like human IAV vaccines, the exact

composition of epitopes needed to elicit contemporarily protective antibodies is decided by the OIE's Expert Surveillance Panel on Equine Influenza Vaccine (OIE-ESP) on an annual basis (Bryant et al., 2011). The constant escape of vaccine-strain antigens by EIV evolution *in vivo* means that vaccines need to be regularly updated to maintain efficacy.

However, antigenic drift and epitope structural changes to epitopes do not fully explain EIV evolution. The role of internal viral proteins may have been previously overlooked in understanding the emergence of endemic EIV and circulating EI epidemics. For instance, some determinants of host-range are situated in internal or non-structural proteins of the virus, like the polymerase complex (Cauldwell et al., 2014; Min et al., 2013).

1.8.3.1 Evolution Between Infected Cohorts

Within a cohort of horses, usually a single herd or farm unit, EIV can spread rapidly through droplets or fomites, making control of the disease difficult once an infected horse has been introduced (Adeyefa et al., 1996; Karamendin et al., 2014). The transfer of horses globally, for racing and breeding, provides critical opportunities for pathogen transmission and introduction into non-endemic areas (Daversa et al., 2017). Additionally, global movement of horses is not random but is often concentrated and anthropogenic which grants opportunities for many horses from across the world to be in proximity for a brief period before returning to their original cohort. On a regional or national scale, individuals can be moved between premises or cohorts frequently – especially for breeding and competition purposes.

Sporadic EIV outbreaks can be seeded by a small number of horses after their travel (Newton et al., 2006; Whitlock et al., 2018). Additionally, the individuals most likely to be moved (racehorses and studs) usually also have the best veterinary care and closest observation. Horses moved for breeding or competitive purposes are more likely to be up to date with vaccine regimens than other non-special horses, or indeed even obliged by regulatory bodies, and so are likely to provide heavy selective pressures on incoming viruses from vaccine-boosted immune challenge.

As expected, the probability of a successful transmission event increases dramatically as the spatial distance between donor and recipient hosts decreases (Ostfeld et al., 2005). Managed athletic and breeding horses are provided with unusual population structures due to their high exposure to numerous other horses and their wide, sometimes intercontinental, travel. The evolutionary and ecological structure of pathogen populations can provide clues as to the factors required to maintain the populations of that pathogen (Parker et al. 2015). This maintenance includes both pathogen (replicative speed, immune avoidance strategies) and host (immunity, contact networks) factors.

1.8.3.2 What Drives Global Evolution?

Like all pathogens, the evolution of EIV is driven by the interplay of deterministic (selective pressures) and stochastic (genetic drift) processes (Lauring,

2020). The international movement of subclinically-infected horses is the most likely predictor of viral dissemination, much as that observed in swine influenza A viruses (Lee et al., 2021; Nelson & Hughes, 2015). Not only does this movement provide opportunities for spread over a further geographic range, but it increases the possibilities for exchange with endemically circulating EIV. This introduction of new viruses into a population may co-opt endemic viruses, using them as stepping-stones to better adapt to the new range of hosts. Surveillance of EIV epidemics show that global outbreaks occur roughly every 2-8 years, mostly associated with the emergence of new antigenic variants (Koelle et al., 2010).

Insufficiently protected hosts (i.e. immunocompromised or unvaccinated) can provide a fitness landscape that is easier to traverse for viruses. This in turn can allow viruses to diversify in the face of lower selective pressures, potentially even adapting into a strain better able to infect. This is seen in the dissociation of viral evolutionary relationships to geographic locations (Lai et al., 2001, 2004). Equine influenza provides an interesting case of heterogeneous host populations: individuals that travel internationally (sport horses) are also the most likely to be vaccinated and closely observed; meaning that a traditional risk factor for "super-spreader" status is potentially avoided due to the heightened immunity and medical observation. Conversely, horses that are more geographically stationary and limited to national or local travel may receive less frequent veterinary visits and are not subject to the same vaccine requirements as sport horses. Understanding how the immune status and vaccination history of these opposing equid populations can affect the evolution of EIV is thus key avenue for further animal health practices.

Additionally, though not unique to EIV, outbreaks of influenza are seen to propagate even in individuals with up-to-date vaccination routines. Antigenically similar viruses are also able to spread through populations with limited prior immunity (Lumby et al., 2020). These weaknesses in individual immunity may explain the occasionally low presentation of symptomatic individuals seen in most outbreaks. Indeed, the proclivity of asymptomatic infections may be responsible for dramatic underreporting of EIV.

Analysis in the US showed a regular fixation of amino acid substitutions distancing circulating wild strains from vaccine strains (Lee et al., 2021). Antigenic shift, at least inter-subtype, has being not detected or associated to equine outbreaks. The origin of the currently circulating virus is unknown, though with cross-species transmission from avian donors and into a wide range of canine recipients since 2004 (Rivailler et al., 2010), H3N8 EIV clearly has the potential for cross-species transmission. Despite multiple historical records recognising links between outbreaks of equine and human influenza-like disease, there is very limited evidence of H3N8 transmission to humans from horses. Infection is possible however, Alford et al. (1967) tested the responses of 33 humans to EIV A/Miami/1/63 (H3N8) and found moderate influenza-like symptoms in four patients. Though notably virus could be recovered from 21 of the 33 participants, so sub-clinical infections with H3N8 are viable in humans.

Whether direct equid-human transmission of IAV was previously possible with contemporarily circulating influenza strains or both populations were infected by a third reservoir population (e.g. waterfowl), equine influenza has coincided with outbreaks of human influenza for centuries (Forster, 2021; Rutty, 1770). Additionally, due to the difficulty in sampling both wild and domesticated horses, most genomic sequences publicly available for EIV are of segment 4; of these a considerable number focus exclusively on the HA1 chain of the segment 4 coding gene (Russell, 2021). This can skew the overall view of EIV evolution, though evolution in antigenically available HA1 epitopes is undoubtedly important.

1.8.3.3 Frozen Evolution

Through longitudinal observation of EIV, detection of decade-old virus strains infecting hosts can confound the reputation of rapid influenza A evolution. After having explored the myriad ways in which diversity can be generated and maintained in fitness equilibria, the preservation of a viral lineage over decades contradicts everything I have discussed so far. Records, mostly in Western Europe, have detailed the collection of EIV samples that much more closely resemble strains that circulated in previous years before being supposedly outcompeted (Lindstrom et al., 1998; Manuguerra et al., 2000). Termed ‘frozen evolution’ by Endo et al. (1992) viruses that had circulated over a 25-year period were recognised with a shockingly low amount of antigenic or genomic change. Viral samples from German EI outbreaks (Borchers et al., 2005) in 2002 resembled genomes isolated from viruses circulating in Europe in the early 1990s. This lack of diversification has been documented in many outbreaks (Manuguerra et al., 2000). As to how stagnant genomes can compete with viruses circulating in subsequent years, experiments to compare the reproductive fitness of these frozen genomes could elucidate whether these viruses are able to replicate independently or not.

These papers evidencing frozen evolution however display some notable oversights. Primarily, analyses from Endo (1992), Manuguerra (2000) and Borchers (2005) rely solely on the short HA1 coding sequence (939 bases) for their claims of abnormally slow evolution; the first two utilise amino acid sequences (328 residues) exclusively. This alongside the small sample size of these studies (never more than 15 sequences) does detract from the validity of their findings. Finally, and I admit speculatively, the context of these evolutionary dynamics cannot be overlooked; all three of these studies explored why unusually old EIV sequences appeared in Europe through the mid-to-late 1990s. I cannot help but note that the fall of the Berlin wall was concurrent with these findings; shifting trade policies across Europe may be an unexciting explanation, but the simple facts of the environs in which this ‘frozen evolution’ took place can easily be understood using the context of larger-scale ecological changes surrounding viral spread.

1.8.4 EIV as a Model of Influenza Phylodynamics

EIV displays many characteristics of IAVs in other mammals, and to date the dynamics seen in equine influenza pathogenesis are broadly comparable to IAVs in other mammalian hosts. Hence, I use H3N8 EIV here to model the evolutionary processes within- and between-hosts as well as the larger epidemic population dynamics of influenza A viruses. Research on the phylodynamics of influenza in horse populations is used here as a model system to investigate key drivers of mammalian influenza A virus evolution, and the ways in which evolution between- and within-hosts can lead to vaccine escape, seasonally-recurrent outbreaks and even cross-species adaptation or pandemic potential.

Understanding the causes and consequences of viral mutant spectra is obviously important for virologists, public health workers and clinicians; but how do we detect and observe them? By definition, variant viruses are a small minority of the overall population and so conventional genome amplification, and sequencing techniques cannot necessarily be relied on. Many bioinformatic procedures were designed specifically to exclude spurious outliers, so how then do we obtain this information from a viral sample such as a clinical specimen (e.g., nasal swab or sputum sample)? Further developing EIV surveillance techniques will be key as globalisation increases; horses are already the most internationally moved domestic animal (Oladunni et al., 2021) and so having up-to-date records of viruses causing symptomatic and asymptomatic infections will help track the evolution of the globally intertwined EIV population.

The trends of EIV evolution mirror those of other IAVs, especially with the anthropocentric movement of horses around the globe for sports. Racehorses may become super-spreaders and seed infections acquired from hyper-mixing populations into home pastures on return from competitions. At large-scale epidemiologic levels, this shows the clear “ignition spark” introduction à la SARS-CoV-2 emergence. As international travel only accelerates, the future of EIV is sure to have ample opportunities for further dissemination.

1.9 Study Aims

Hence the importance in understanding evolutionary processes within- and between-hosts as well as the larger epidemic or pandemic population dynamics. In examining the impact of host heterogeneities on viral populations, I sought to understand:

- The role of prior exposure to influenza viruses in affecting viral population size and evolution at within-host and outbreak scales
- Differences in viral load between vaccinated and naïve hosts; whether a primed immune system causes reduced viral shedding and consequently lowers the infectivity of vaccinated hosts
- The spread of EIV in transmission chains comprised of hosts with differing histories of immunological exposure, as seen in real EIV outbreaks

- The fate of consensus-level mutations and whether they are impacted by the vaccination status of a host; if so, does immunological history matter
- Putative impacts of nonsynonymous mutations on 3D protein structures based on *in silico* modelling and experimentation
- The role of sub-consensus diversity in shaping viral populations over an outbreak
- What drives diversification and selection of viral variants in a host

Characteristics of viral bottlenecks in tightly controlled transmission chains; their size and how much variation they permit passing from one host to the next. Epidemiology and pathogen evolution influence each other, yet overwhelmingly, research focuses on the ways in which evolution alters epidemic dynamics. Obtaining pathogen sequence data at high resolution is still relatively rare across whole viral genomes.

The use of mixed, vaccine-exposed and immunologically naïve individuals in transmission experiments aims to represent real populations of horses with differing levels of immune exposure to naturally-circulating influenza viruses. Where EIV is endemic, horse populations are regularly exposed to IAV and thus many individuals will have some level of prior exposure; the majority of individuals without any prior IAV exposure will be young. However, populations will be heterogeneous in respect to their levels of immunity and their history of exposure to pathogens; older individuals are more likely to have encountered multiple different strains of IAV during their lifetime. Including heterogeneity in host immune statuses enables us to explore questions of viral evolution from many different angles, under varying situations.

Additionally, despite the plethora of work on IAV proteins, comparatively little has been done explicitly on EIV proteins. Consequently, many of the specifics relating to EIV proteins, from numbering to regional annotations, are inferred from work on other (often avian) viruses. As discussed above, EIV is a direct descendant of avian IAV and we can therefore assume that many of the protein characteristics are shared. Comparative analysis of orthologous IAV proteins may therefore inform on the structure and function of equine influenza viruses.

2 Methodology

2.1 Experimental Design

To investigate the impact of prior host immunity on IAV viral diversity and population size, data collected from two transmission experiments were obtained for the present study. Both experiments involved the use of naïve and vaccinated Welsh Mountain Ponies challenged with the A/equine/Newmarket/5/2003 strain of IAV (txid:568375). This study was carried out following animal care guidelines of the Animal Health Trust Ethical Review Committee, under Home Office project licence 80/1871. Each transmission chain included six pairs of animals (individuals A and B in pairs 1-6). Pairs 2, 3 and 4 had previously been immunised with H3N8 inactivated (non-adjuvanted, formalin-inactivated egg grown) virus prior to the experiment to

allow the development of adaptive immunity; finally, pairs 5 and 6 in the chain of transmission were immunologically naïve. The first pair of each transmission chain (seeders, which had been experimentally inoculated) were excluded from further analyses as they did not represent infection by the natural route of transmission.

Week	Vaccine Dose	Vaccine Antigen			
		Multivalent	NCBI txid	Monovalent	NCBI txid
0	1	A/Equine/Miami/63	387223	A/Equine/Newmarket/3/05	568375
4	2	A/Equine/Miami/63		A/Equine/Newmarket/3/05	
16	3	A/Equine/Newmarket/79	1334814	A/Equine/Newmarket/3/05	
28	4	A/Equine/Newmarket/1/93	159470	A/Equine/Newmarket/3/05	
40	5	A/Equine/Newmarket/3/05	568375	A/Equine/Newmarket/3/05	
		Transmission experiment ran from Week 68 to Week 71		Transmission experiment ran from Week 80 to Week 83	

Table 2.1: Vaccination schedules of each transmission chain.

The difference between transmission chains was the vaccination schedule of the horses (Table 2.1): horses in the multivalent vaccine group received five doses of four different antigens, while horses in the monovalent vaccine group received five doses of the same antigen. The antibody levels of vaccinated horses were measured using single radial haemolysis (SRH) until they reached a value low enough (<60 mm^2) to allow natural infection as previously described (Murcia et al., 2010,

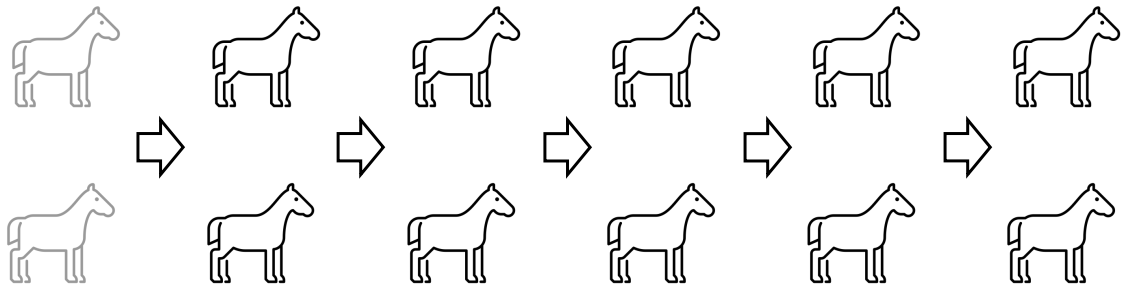


Figure 2.1: Schematic representation of transmission chains. Both experiments had the same structure, the only difference being the use of two different exposure regimes. Pair 1 horses were infected by inoculation rather than by natural transmission; for this reason they are not included in the analysis and so are greyed-out.

2013). To initiate each transmission chain, a pair of immunologically naive "seeder" horses were experimentally inoculated via nebulised aerosol (20 ml of $\log_{10}$ $10^{6.5}$/ml of 50% egg infective dose [EID_{50}]) of A/equine/Newmarket/5/2003 (H3N8). This challenge virus is also the strain used in vaccines to which vaccinated horses were exposed. To allow natural transmission by direct contact, seeder horses were cohoused in the same stable with a pair of recipient horses until at least one of the latter started shedding virus, which was confirmed using an enzyme-linked immunosorbent assay (ELISA)-based assay. At this point seeder horses were removed and replaced by a new pair of recipient horses, i.e. 'Pair 2'. This procedure of sequential mixing of pairs was repeated a further four times down to 'Pair 6' (Figure 2.1). At no time were more than two pairs of horses sharing a stable. Nasal swabs were collected daily in 5ml of virus transport media (VTM) and stored at -80°C until further processing. Samples that contained >2960 viral copies/µl of transport media

as measured by a PCR assay (see section 2.3) were subject to full genome PCR amplification and sequencing using the Illumina platform (see section 2.4).

The multivalent vaccination schedule was planned to simulate individuals with a life history of multiple previous exposures to a range of antigenically distinct viruses. Alternatively, horses that received the univalent vaccine were expected to model an individual that has developed specific immunity to a currently circulating strain via vaccination. The schedules of the whole experiment, including the exposure history of vaccinated hosts, are shown below in Figure 2.2, using abbreviated names for the virus strains that match those in Table 2.1. Horses within this vaccinated class received one dose of inactivated virus at each vaccination point (V1-5).

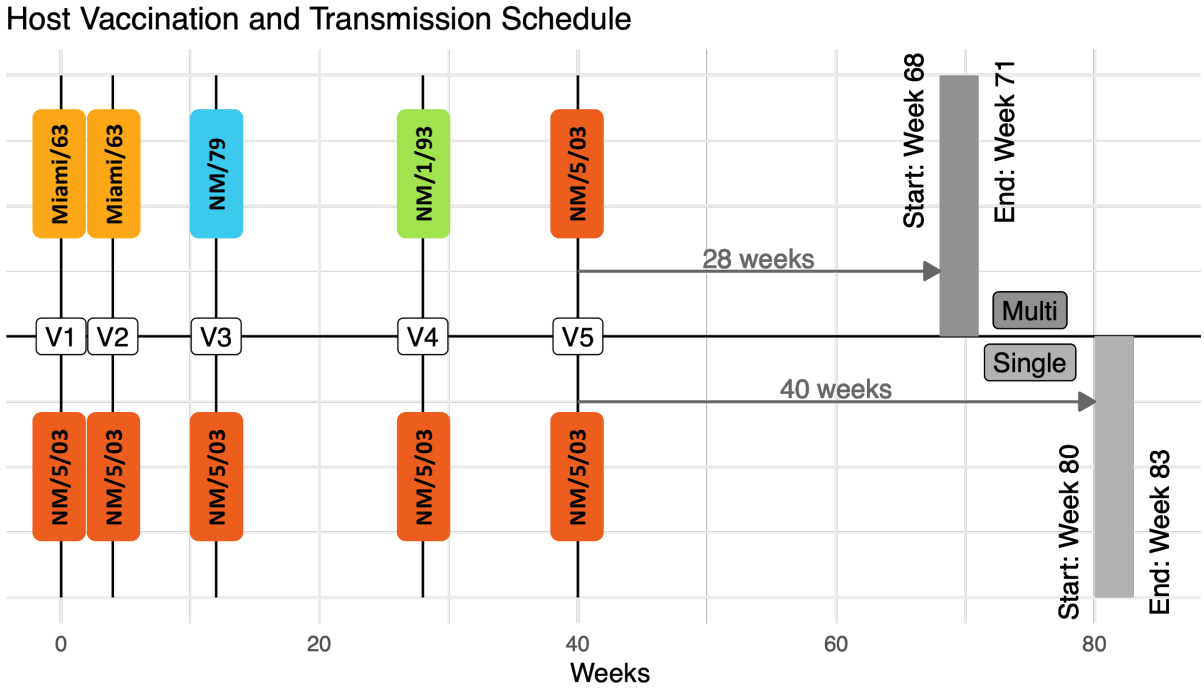


Figure 2.2: Diagram of the exposure regimen, and the time each experimental transmission chain began. Inactivated viruses were administered at dates V1-5, referencing the Table 1 schedule.

After each vaccination point, horses were bled in order to test serum adaptive immune responses to the inactivated viruses. Exposures were staggered to allow for a return to sera norms before exposure to the next immunogen.

The two transmission chains were used to observe whether differences could be observed between a specific and a generalised adaptive immune response in terms of viral load, virus diversity and/or viral phylodynamic processes. Hence, two transmission chains were studied, each containing five pairs which fell into one of four immunological statuses: vaccinated or naïve, in transmission chain one (vaccinate-multivalent chain V_M or naïve-multivalent chain N_M) or two (vaccinate-univalent chain V_S or naïve-univalent chain N_S).

2.1.1 Transmission Experiment

Seeder horses were experimentally infected with $10^{6.5}$ egg infectious doses 50 (EID_{50}) of A/equine/Newmarket/5/2003. Nasopharyngeal swabs were collected on a daily basis and virus shedding was detected using a rapid nucleoprotein (NP) enzyme-linked immunosorbent assay (ELISA) test. If the assay was positive for at least one

of the two recipient ponies, these ponies would become the new donor ponies; two new vaccinated recipients would be co-housed in the transmission room and the previous donors would be removed.

Once at least one of the new recipient ponies were positive for virus shedding, the original donors would be removed to the recovery room, where they would continue to be swabbed for up to seven days, or as long as they were positive by NP-ELISA, whichever was the longest. The transmission room would be thoroughly cleaned and disinfected between movements of pairs to avoid environmental and fomite transmission of virus.

2.2 Data Collection

2.2.1 Viruses and vaccines.

A/equine/Newmarket/5/2003 was passaged *in ovo* to generate a stock of challenge virus. Vaccine viruses were cultivated in embryonated chicken's eggs, followed by clarification, sucrose purification and inactivation using 0.02% formaldehyde. Vaccines were tested by passaging in embryonated chicken's eggs (two passages) to ensure they were no longer infectious.

2.2.2 Nasal Swabs

Nasopharyngeal swabs were collected for up to eight days after a horse tested positive with a rapid Nucleoprotein ELISA test. Swab tips were immersed in viral transport medium (5ml) and stored at -80°C. Daily nasal swabs were used to quantify viral loads, 137 swabs giving positive qPCR values were collected: 68 from the Single group (42 from vaccinates [V_S], 26 from naïves [N_S]) and 69 from the Multi group (41 from vaccinates [V_M], 28 from naïves [N_M]).

2.2.3 Virus Quantification via qPCR

RNA was extracted from nasal swabs by the team that carried out the original transmission study in order to quantify the amount of virus present. The team then used qPCR as described in Murcia et al. (2010, 2013), with full multi-segment reverse transcription-PCR (M-RTPCR), the details of which are available in Deng et al. (2009). Viral RNA from nasal swabs was isolated from 280µl aliquots using the QIAamp viral RNA minikit (Qiagen) according to the manufacturer's instructions, eluting in a volume of 50µl.

To calculate the number of virus genome copies present in each sample, cDNA was generated using Superscript III (Invitrogen) and primer Bm-M-1[5]. Reverse Transcription was performed at 55°C for 90 min, followed by incubation at 70°C for 10 min. Viral copy numbers were estimated by qPCR, performed using the QuantiTect Probe PCR kit (Qiagen) according to the manufacturer's instructions and using the same primers and probe as in both Murcia and Hughes (2012; 2010, 2013), which had been designed using Beacon designer (Premier Biosoft). Standard curves were generated using 10-fold dilutions of a plasmid containing the matrix segment (cloned from an egg-grown Equine/Newmarket/1/1993 isolate), ranging from 1×10^2 to 1×10^8copies/µl. For each run, all samples, no-template controls, plasmid standards, and positive and negative controls were run in triplicate and expressed as the mean number of viral RNA (vRNA) copies of cDNA per µl.

Samples that exhibited >2960 viral copies/µl of transport media were subject to full genome PCR amplification as described by Zhou et al (2009). Swabs with fewer genomes were unable to be amplified without the introduction of stochasticity in sequences. PCR amplification was performed using Platinum Pfx DNA polymerase (Invitrogen) and segment non-specific primers as designed by Zhou et al (2009): MBTuni-12 [5′-ACGCGTGATCAGCAAAAGCAGG] and MBTuni-13 [5′-ACGCGTGATCAGTAGAAACAAGG]. As IAV genomic segments have conserved 12nt and 13nt sequences at the 3' and 5' ends respectively, these universal primers can be used to amplify genomes irrespective of virus subtype. PCR amplification was performed for 40 cycles (94°C for 30s, 55°C for 1 min, and 68°C for 1 min), followed by a final extension at 68°C for 10 minutes.

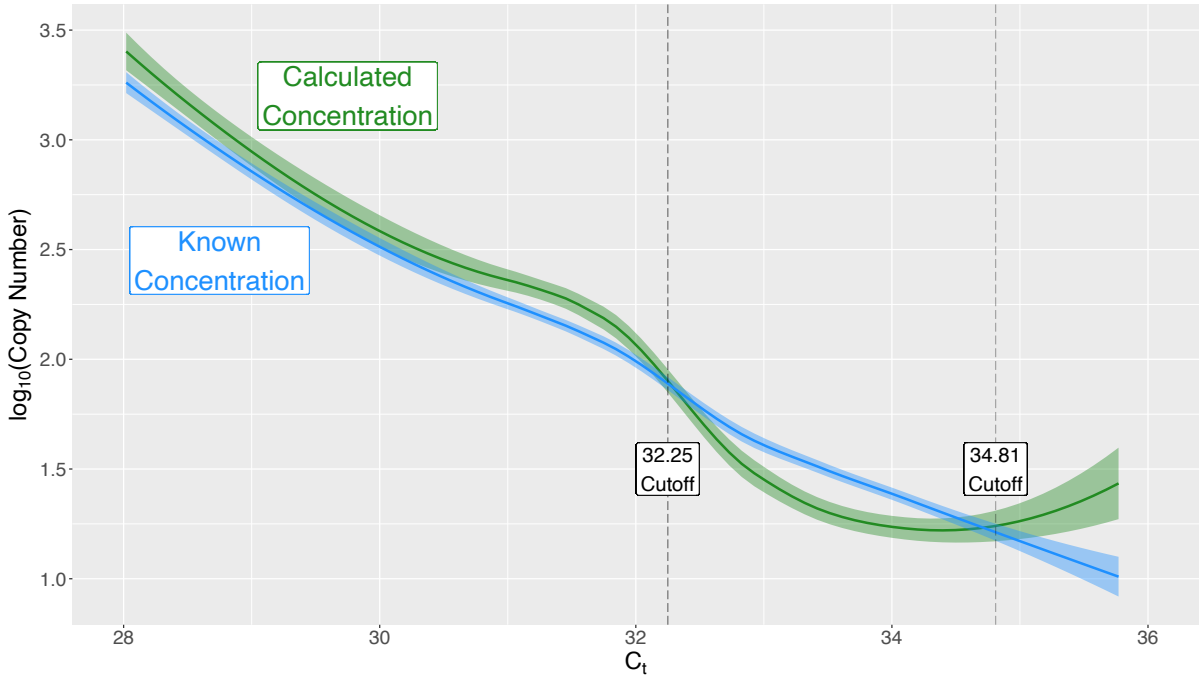


Figure 2.3: The mean copy numbers of plasmid standards used to generate standard curves for qPCR validation. Known numbers of plasmids are input for amplification (blue). The resulting output (green), gives the number of cDNA copies counted after the amplification. When the two figures mismatch, the threshold of detection is reached.

With PCR amplification outputs, known and calculated concentrations of plasmid standards were compared to validate and determine a cut-off point. This point denotes the upper limits of detection, and represent a concentration too low to correctly match the known input concentration (Figure 2.3). This was initially determined at a Cycle threshold (C_t) of 34.81, matching that reported in Murcia et al. 2010. However, on re-analysis, a more conservative C_t was declared at 32.25 as this is the first introduction of stochasticity in amplification of known concentrations of plasmid standards.

2.2.4 Sequencing & Sequence Assembly

DNA was diluted to a concentration of 175ng in 50µl of sample prep then acoustically sheared using a Covaris S220 sonicator. Sequencing was carried out at The Genome Analysis Centre (TGAC) in Norwich, UK. Illumina GA2x sequencing was

then performed in one lane with 100bp paired-end reads. Illumina libraries were then constructed from 200-300 bp fragments.

53 sequence libraries were generated from forward and reverse reads: 24 from the Single group (11 from vaccinates [VS], 13 from naïves [NS]) plus 29 from the Multi group (13 from vaccinates [VM], 16 from naïves [NM]). Forward and Reverse reads were aligned to the genome sequence of the virus challenge strain (A/equine/Newmarket/5/2003, NCBI Taxonomy ID: 568375) using the Burroughs-Wheeler Algorithm (Li et al. 2009). Adaptor sequences were removed with Trimmomatic v0.4 (Bolger et al., 2014). Sequences with a mean quality score <30 were also removed. Functions within the 'samtools' package array v1.12 (Danecek et al., 2021) sorted, indexed and compiled the reads into useable fastq & fasta files, the code for which is shown below.

Box 1: Example of the code used for adaptor trimming, read compilation, genome assembly and finally creation of FASTQ files for subsequent genomic analyses.

```
prinseq-lite.pl -trim_left 7 -trim_right 7 -min_qual_mean 30 -ns_max_n 0
-lc_method dust -lc_threshold 7 -fastq File1_R1.fastq -fastq2 File1_R2.f
astq -out_good File1_good_reads -out_bad File1_bad_reads;
bwa mem Reference.fa File1_R1_001.fastq File1_R2_001.fastq > File1.sam;
samtools sort -@10 File1.sam -o File1.bam;
samtools index File1.bam;
samtools idxstats File1.bam;
samtools mpileup -uf Reference.fa File1.bam | bcftools call -c | vcfutil
s.pl vcf2fq > File1_clean.fastq
```

2.2.5 Variant Calling

Utilising the 53 consensus genomes, sub-consensus variants were called from BAM files using iVar v1.4.2 (Grubaugh et al., 2019), LoFreq v2.1.5 (Wilm et al., 2012), vSensus (Orton, 2022) and FreeBayes v1.3.6 (Garrison & Marth, 2012). Each tool employs different algorithms, processing requirements and filtering procedures. Datasets detailing the frequency of sub-consensus mutations at each nucleotide position were then associated with sample metadata.

Following the creation of consensus sequences, sub-consensus variants may be detected and analysed. Due to the sheer volume of viral genomes in most viral samples, most minority variants fall below a set threshold and thus are excluded from analysis. This threshold varies depending on the efficacy of the genome amplification technique and the specificity of the sequencing procedure, but as a standard, most laboratories (Koel et al., 2020; McCrone & Lauring, 2016) place a cut-off value at genomes that constitute less than 1% of the total viral population meta-genome after amplification. With genomes annotated and aligned, the mutant spectra can then be analysed as with any phylogenetic dataset, the only exception being the scale of both time and relatedness are much smaller than in traditionally multi-species/multi-strain phylogenetic trees.

Where variant calling tools did not present a conclusive list of variants (as in Diversitools), outputs were manually filtered. Any site in which fewer than 99% of reads were congruous (i.e. 1% or greater variant frequency threshold) was flagged as a site of low-frequency variation. Filtering was performed using R scripts.

2.2.5.1 Variant Caller Selection

The variant calling tool on which to base subsequent analyses was selected on the basis of performance benchmarks, including specificity and sensitivity. Ten tools were compared in total (Table 2.2A): those mentioned above together with VarScan (Koboldt et al., 2012), DeepSNV (Gerstung et al., 2012) and Diversitools (Hughes, 2016). Box 2.2 shows the pipeline used to produce each low-frequency variant array.

Table 2.2: A) Bioinformatic tools selected for comparative analysis and B) the datasets containing the sequences which were used to compare and assess them.

A)

Tool	Reference
DeepSNV	Gerstung et al., 2012
DiversiTools	Hughes, 2016
FreeBayes v1.3.6	Garrison & Marth, 2012
iVar v1.4.2	Grubaugh et al. 2019
LoFreq v2.1.5	Wilm et al., 2012
VarScan	Koboldt et al., 2012
vSensus	Orton, 2022
V-Phaser2	Yang et al., 2013
QuasiBAM	Manso et al., 2017
VirTools	Verbist et al., 2014

B)

Dataset	Repository
SimData	N/A
McCrone 2016	PRJNA317621
McCrone 2018	PRJNA412631
Han 2021	PRJNA722099
Poelvoorde 2022	PRJNA692424

Testing began with previously published sequence data, with corresponding records of the sub-consensus variant frequencies found by the original authors (Table 2.2B). These five control datasets were then used to compare each in terms of True-Positive Rates (sensitivity) and True-Negative Rates (specificity). Thus, the known, published results were taken as a gold standard, and assumed to be the absolute truth - as the aim of this testing was deciding upon a reliable, repeatable bioinformatic tool the actual values are less important than aligning with the performance of the tool. Processing times for each sample were also recorded, for measuring performance efficiency. Ultimately, a combination of LoFreq and FreeBayes provided the most comprehensive results.

2.2.5.2 Variant Call Analysis

Variants from each sample's BAM assemblage were called using LoFreq v2.1.5 with reference to the consensus of that sample, rather than using the consensus of the entire dataset for each variant call. The use of this dynamic consensus, specific to the read library in question, was chosen to avoid spurious variants; a base in disagreement with the consensus of the entire dataset may have become fixed at the consensus level in one host and thus by definition would no longer be a sub-consensus variant. Low-frequency variant arrays were then associated with sample metadata.

```
for f in *.bam
do echo $f ",iVar" >> time.txt;
{time (samtools-1.12/samtools mpileup -aa -x -B -d 0 -A -q 0 -Q 0 -C 0 -
f ref.fasta $f | ivar variants -p ivar -q 0 -t 0 -m 0 -r ref.fasta -g
Ref.gff) ; } 2>> time.txt;
mv ivar.tsv $f.tsv;
echo "\n" $f ", Diversitools" >> time.txt;
{time (diversiutils_macosx -bam $f -ref ref.fasta -orfs CodingRegions.txt
-stub $f) ; } 2>> time.txt
echo "\n" $f ",LoFreq" >> time.txt;
{time (lofreq call -f ref.fasta -o $f.vcf $f);} 2>> time.txt;
echo "\n" $f ", FreeBayes" >> time.txt;
{time (freebayes -f ref.fasta $f > $f-FreeBayes.vcf);} 2>> time.txt
done

for i in *.mpile.txt;
do echo "\n" $i ",vSensus" >> time.txt;
{ time (java -jar VSENSUS.jar $i > $i_log.txt);} 2>> time.txt;
echo "\n" $i ", VarScan" >> time.txt;
{time (java -jar VarScan.v2.3.9.jar mpileup2snp $i --min-var-freq 0.02 -
-p-value 0.05 > $i-VarScan.tsv);} 2>> time.txt;
done
```

2.3 Data Analyses

2.3.1 Analyses of Viral Shedding

Throughout the thesis, various statistical methods were used to explore the patterns and impacts of variables, all of which were performed in R v4.2.0 (2022). Correlations between viral copy numbers and numeric variables, such as sequence diversity, were assessed using a Spearman correlation test. Tests for the normality of shedding values relied on Shapiro-Wilk tests. Comparisons between stratified datasets were implemented with non-parametric Kruskal-Wallis and Wilcoxon Rank-Sum tests; this included assessing whether host variables, i.e. transmission group and/or vaccination status, significantly differentiated viral copy numbers. All of the above tests incorporated a Bonferroni correction for multiple testing.

Finally, the impact on viral shedding of variables such as transmission group or days post-contact was quantified using an array of linear and additive general models, all of which were performed under a Bayesian prior-parameterisation process. These models were created and estimated by the *rstan* package (2022). MCMC chains were examined for 50,000 samples to ensure proper mixing of posterior values and sufficient sample sizes from which to draw inferences. Mean posterior-

predictive density ($mean_{PPD}$) was used to qualify certainty of coefficient estimations and was deemed informative when representing over half of the trialled models.

To note, in creating statistical models, vaccinated hosts were nested within each transmission chain to which the host belonged. Care was taken to uncouple these variables when making inferences and so models were created in triplicate, initially to observe host vaccination status and their transmission chain (m_1: $Vacc_S$ + $Vacc_M$ + Naïves), then to compare with models that only account for transmission chains (m_2: X_{Single} + Y_{Multi}) and a totally null model (m_0: Host). Individual animal was not included as a random effect. Further, as samples were collected daily, the data are time-serial and so individual animals were incorporated as random factors in order to account for non-independence of variables. Model regression tables and raw data are presented in appendices (Supplementary Figure 2.1). To incorporate time-serial samples, Generalised Additive Models were created with the use of the 'Day Post-contact' variable (abbreviated to DPC), opting to smooth days 0-8 over an eight-fold kernel. This enables flexibility of predictions without the constraints of assuming linearity between variables. Overall, best-fit models were constructed with the following foundation:

$$log_{10}(mean\ copy\ number) \sim Status + Group + s(DPC, k = 8)$$

2.3.2 Phylogenetic analysis

2.3.2.1 Sequence Alignment

Labelled with their corresponding metadata: individual sampled, day of sampling, transmission group and vaccination status, fasta sequence files were then imported into Geneious Prime v2023.1.2 where they were aligned using the Clustal Omega multiple sequence aligner (Sievers et al., 2011, 2020; Sievers & Higgins, 2018). Mutations were called from the consensus of this alignment. As a convention throughout this thesis, when discussing mutations, a lowercase letter indicates nucleotides (e.g. a101g) while an uppercase letter indicates amino acids (e.g. Ser101Asp).

2.3.2.2 Substitution Model

The most parsimonious evolutionary model was selected by opening sequence alignments with ModelFinder, embedded in the IQTree2 package (Kalyaanamoorthy et al., 2017). Substitution models were assessed and chosen based on Akaike and Bayesian Information Criteria (AIC and BIC). Finally, a model with unequal transition/transversion rates and unequal base frequencies was selected (HKY) (Hasegawa et al., 1985) with the additional assumption of empirical base frequencies (+F); ultimately an HKY+F substitution model was declared the best-fit for this alignment. Though reassortment is undoubtedly a feature of IAV evolution, it was ignored in the ensuing analyses since the high homogeneity at the consensus level precluded its examination. At the sub-consensus level, detecting reassortment would be even more of a challenge to detect and would involve using bioinformatic programmes at the forefront of development, beyond the scope of this project.

2.3.2.3 Maximum Likelihood

Maximum Likelihood (ML) trees were estimated using IQTree2 (Minh et al., 2020) with the eight genomic segments being concatenated to make a single 13kb sequence for each sample. Trees were generated for each individual segment and for all segments concatenated together. All trees were validated using 1000 bootstrap replicates.

2.3.2.4 Maximum Clade Credibility

Trees were estimated in Beast v10.4 with the help of the BEAST suite and auxiliary programmes such as BEAUti v10.4 (Drummond & Rambaut, 2007; Suchard et al., 2018) and Tracer v1.7.1 (Rambaut et al., 2018). Two monophyletic trees were estimated, as both shared the ancestral Newmarket/5/03 strain as the initial challenge inoculum. Each of the eight genomic segments had independent locally random clock models and HKY substitution models giving a total of 17 evolutionary models to analyse, including the shared tree model.

MCMC chains for tree estimation ran for 100 million iterations, with 10% burn-in, on the CIPRES (https://www.phylo.org/portal2) server. Constant and SkyRide coalescent models were tested, but for such a small, homogeneous population the differences between final MCC estimations proved negligible. The tree sampling process was repeated four times independently in order to ensure proper mixing and convergence of MCMC chains. After BEAST runs concluded, model parameters and goodness-of-fit were assessed in Tracer via Effective Size Sampling before finally TreeAnnotator was used to compile tree estimations into a single parsimonious Newick file.

2.3.2.5 Phylogenetic Trees

Tree visualisations were created in FigTree v1.4.4 (Rambaut, 2018) or R with the ‘ggtree’ package (Xu et al., 2022). Tree topologies and other properties were analysed using R packages such as ape (Paradis & Schliep, 2018), PopGenome (Pfeifer et al., 2014) and, specifically for MCC trees, Tracer (Rambaut et al., 2018). Mean substitution rates of each genomic segment were calculated in BEAST using the median Rate statistic.

2.3.3 Analyses of Sequence Diversity

At both the consensus and sub-consensus levels, sequence diversity was measured with multiple metrics (Gregori et al., 2016), ranging in complexity and representativeness. Most of these metrics are here applied at the consensus and per-site, sub-consensus scales granting an insight to the diversity of sequences as a whole and on a site-by-site basis. All calculations of within-host diversity utilised the variant call data from LoFreq.

2.3.3.1 Mutation Abundance

Mutation frequency (M_f) is an estimation of diversity based on comparing all haplotypes to the most frequent haplotype in a population. It is the average number of mutations observed in all haplotype sequences relative to the most frequent

haplotype: $M_f = \frac{\text{mutations}}{\text{reads} \times \text{nucleotides}}$. Simply, it is the proportion of sequences/reads that do not match the consensus sequence/nucleotide.

At the sub-consensus level, the number of different mutations together with their respective frequencies is referred to as the site frequency spectrum and can be used as a measure of evenness of the population. A population can be said to be very even if all mutations or haplotypes have a similar prevalence in the population. Finally, we considered richness of mutant sequences/reads, which is the number of polymorphisms per alignment/kilobase.

2.3.3.2 Simpson's Index

Having been adapted from ecological studies, Simpson's Index gives the likelihood that two sequences, randomly selected from a viral population, are identical (Gregori et al., 2016). With the mutational frequency p_k, the probability of two randomly sampled sequences having identical nucleotides at a given position (k) is given by: $P_S = \sum_k p_k^2$. This is then averaged across the entire genome. Simpson's index is therefore bound between 0 (no chance of finding identical sequences) and 1 (lack of diversity in the population). This was carried out in R by summing the mean mutational frequencies (p_k) of each genomic segment.

To note, Simpson's index is strongly skewed by the most abundant sequence in the population. Further, squaring the mutation frequencies (p_k^2) means that rarer mutations become quickly lost in the analysis. This can have the detrimental effect of biasing results, by enriching the mutations that are already present in high abundance; whereas our aim in this study needed observation of mutations present in very small proportions of genomes.

2.3.3.3 Shannon Entropy

Shannon Entropy (H_S) is another diversity metric used commonly in ecological studies which has been adapted for use in virology. Shannon Entropy is known to be sensitive to the size of the sample under study. Shannon's Entropy (Shannon, 1948) of consensus sequences was calculated using the entropy function of Bios2cor (Taddese et al., 2022).

To compare the genetic diversity between multiple samples, the mean entropy across all sites is used. Shannon entropy can be computed as:

$$H_S = -\sum_{\alpha \in \{A,C,T,G\}} p_{i\alpha} \times log(p_{i\alpha})$$

In this expression, i represents each base position and $p_{i\alpha}$ is the proportion of nucleotide α at position i. This was carried out in R using the 'entropy.Dirichlet' function from the R package 'entropy' (Hausser & Strimmer, 2008, 2021).

2.3.3.4 Tajima's D

Tajima's D is a population genetic test computed as the difference between two measures of genetic diversity (Tajima, 1989), the mean number of pairwise differences between sequences and the number of polymorphic sites. This was calculated using the PoPoolation package (Kofler et al., 2011) which was ran locally in perl for each read library.

Tajima's D test aims to distinguish between a genetic sequence evolving randomly (neutrally) and one evolving non-randomly. A randomly evolving genetic

sequence is expected to contain mutations with no effect on fitness and survival. The purpose of Tajima's test is to identify sequences which do not fit the neutral theory model at equilibrium between mutation and genetic drift. Tajima's statistic measures the total number of polymorphic sites in the sampled genome and the average number of mutations between pairs in the sample, both of which are estimates of the population genetic parameter θ. If the difference between these two parameters (θ_1 and θ_2) could be reasonably explained by chance, then the null hypothesis (H_0 = neutrality) cannot be rejected. Otherwise, the null hypothesis of neutrality is rejected.

Under the theory of neutral evolution, for a population of constant size at equilibrium, the following equation is applicable:

$$E\left[S \div \sum_{i=1}^{n-1} \frac{1}{i}\right] = 2\mu N_{eff} = \theta$$

In this expression, S is the number of segregating sites, n is the number of samples, N_{eff} is the effective population size, μ is the mutation rate at the locus in question, and i is the index of summation.

d_{Tajima} is calculated by taking the difference between the population genetics parameter θ of two samples ($d = \theta_1 - \theta_2$). D is then calculated by dividing d_{Tajima} by the square root of its variance $\sqrt{\mathrm{Var}(d)}$ (its standard deviation, by definition). Thus,

$$D = \frac{d}{\sqrt{\mathrm{Var}(d)}}$$

Tajima demonstrated *in silico* that D could be modelled using a ß distribution (Tajima 1989), work which was then built upon by Kim et al. (2016) in their exploration of chicken and human IAV diversity. If the D value for a sample of sequences lies outside the confidence interval of this distribution, then the null hypothesis, i.e. neutral evolution, is rejected for that sequence. However, in real world uses, past population changes, such as a population bottleneck, can bias the value of D.

2.3.3.5 Pairwise Distance Indices

Nucleotide diversity (π) is used to quantify the distance between two samples through the proportion of sites that they do not have in common. Population nucleotide diversity, or index π, measures the average number of nucleotide differences between any two genomes of the quasispecies (Nei & Gojobori, 1986). Pairwise differences have been traditionally evaluated using the Hamming distance, which is the number of mutations that distinguish a pair of sequences, although any substitution model (JC69, K80, F81, etc) or subsets of differences (transitions or transversions, synonymous or non-synonymous mutations) may be considered. Index π provides more valuable information than M_f because it takes into account the differences between any two genomes in the population.

Consensus π Diversity

Nucleotide π diversity measures genetic variation within a population. Overall, π diversity counts the net number of nucleotide differences between sequences, ultimately giving the average number of differences between two randomly selected sequences from the dataset. In the present study, it was calculated using the 'diversity.stats' function, which is based on original methods from Nei (1988),

Hudson (1992) and Wakeley (1996), within the PopGenome R package (Pfeifer et al., 2014). π is calculated by:

$$\hat{\pi} = \frac{n}{n\text{-}1} \sum_{ij} x_i \, x_j \, \pi_{ij}$$

where x_i and x_j are the respective frequencies of the i^{th} and j^{th} sequences, π_{ij} is the number of nucleotide differences per nucleotide site between the i^{th} and j^{th} sequences, and n is the number of sequences in the sample. The term in front of the sum $\left(\frac{n}{n-1}\right)$ guarantees an unbiased estimator, making the π value comparable across any dataset, regardless of the number of sequences.

Sub-consensus Nucleotide Diversity

Nucleotide π diversity quantifies the distance between two samples through the proportion of sites that they do not have in common. The π diversity is calculated with the SAMFIRE tool by Illingworth (2016). To summarise, π may be calculated as the probability of two random sequences (*sequence$_i$* and *sequence$_j$*) having different nucleotides at a specific position (*d*), averaged over all positions throughout the entire genome sequence (*n*).

$$\pi = \sum_{i<j} \frac{2d_{ij}}{n(n\text{-}1)}$$

This is then normalised against the population size (in this case the number of genome copies) in much the same way as Shannon Entropy, giving: Effective π Diversity (π_e) = $\frac{\pi}{\text{copy numbers}}$

2.4 Evolutionary Selection Analysis

The following algorithms within the HyPhy package were used to examine evidence of selection or directional evolution: 1) Mixed Effects Model of Evolution (MEME); 2) Fixed Effects Likelihood (FEL); 3) Single Likelihood Ancestor Counting (SLAC); 4) Fast Unconstrained Bayesian Approximation for Inferring Selection (FUBAR); 5) Branch-site Unrestricted Statistical Test of Episodic Diversification (BUSTED) and 6) Adaptive Branch-Site Random Effects Likelihood (aBS-REL). To note, segments 7 and 8 were excluded from the analysis as they did not have enough diversity to measure any kind of evolution at the consensus level. Evolutionary process calculations were carried out by the HyPhy software package wrapped in DataMonkey's web server (Delport et al., 2010; Pond et al., 2005).

2.5 Protein Structure Analysis

Consensus nucleotide sequences were translated into protein sequences using the 'ape' package in R, following which protein properties were estimated via ProtParam tools (Duvaud et al., 2021; Gasteiger et al., 2005). These tools allow for the estimation of a range of physiochemical properties such as weight, hydrophobicity and charge. Surface accessibility and protein localisation, though already well-understood for influenza A viruses, were confirmed in EIV using online tools such as the Deep-learning Transmembrane Hidden Markov Model (Hallgren et al., 2022) and the Emini surface accessibility scale (Emini et al., 1985).

2.5.1 Surface Accessibility

For a given amino acid sequence, the accessibility score of residue n is a normalised product of the surface probabilities of amino acids in positions n-2 to n+3, using experimentally qualified amino acid accessible surface probabilities (Janin et al. 1978). A surface residue is defined as one with >2.0nm^2 of water-accessible surface. Utilising the surface probabilities for amino acids, a surface probability (*S*) at residue *n* is defined as:

$$S_n = \left(\prod_{i=1}^{6} \delta_{n+4-i} \right) \times 0.37^{-6}$$

where δ_n is the fractional surface probability for the amino acid at position *n*. S_n probabilities greater than one indicate an increased likelihood that the residue and its immediate surroundings are accessible.

2.6 Estimation of Immunogenic Sites

The Immune Epitope Database (Vita et al., 2019) hosts the Kolaskar-Tongaonkar Antigenicity (Kolaskar & Tongaonkar, 1990) scale, which is used to estimate possible immunogenic sites from protein sequences. This is coupled with the Linear Epitope Prediction 2.0 tool (Jespersen et al., 2017) in order to highlight putative epitopes based on the properties of amino acids in a sliding window of seven residues. EIV haemagglutinin and neuraminidase protein sequences were examined in the current work in order to assess whether any non-synonymous mutations would affect protein antigenicity.

The semi-empirical method by Kolaskar & Tongaonkar predicts the antigenicity of heptapeptide strings across whole proteins, making use of physio-chemical properties of amino acids and their abundance in experimentally-qualified epitopes to estimate and score how antigenic a sequence is. Each residue is scored from 0.77-1.41, and the average of this score together with that of the three residues before and after, gives the probability of a heptapeptide being recognisably antigenic by cells and molecules of the adaptive immune system.

Parameters such as hydrophilicity, flexibility, exposed surface and polarity of polypeptide chains have been correlated with the location of continuous epitopes. This has led to a search for empirical rules that would allow the position of continuous epitopes to be predicted from certain features of the protein sequence. All prediction calculations are based on propensity scales for each of the 20 amino acids. Each scale consists of 20 values assigned to each of the amino acid residues on the basis of their relative propensity to possess the property described by the scale.

When computing the score for a given residue *i*, the amino acids in an interval of the chosen length, centred around residue *i*, are considered. In other words, for a window size *n*, the $i - \frac{n-1}{2}$ neighbouring residues on each side of *i* were used to compute the score for residue *i*. Unless specified, the score for residue *i* is the average of the scale values for these seven amino acids. In general, a window size of five to seven residues is appropriate for finding regions that may potentially be antigenic.

Application to a variety of proteins has shown that this method can predict per-residue antigenicity with about 75% accuracy (Kolaskar and Tongaonkar 1990).

Indeed, this tool has been referenced in 46 publications (at time of writing) regarding SARS-CoV2, highlighting its relevance and ease-of-use. In the present study both consensus (denoted haplotype A) and mutant forms of EIV haemagglutinin (haplotypes E and G) and neuraminidase (haplotypes H, L and M) proteins were analysed to observe any potential differences non-synonymous mutations have on protein antigenicity.

2.6.1 Epitope Prediction

BepiPred Linear Epitope Prediction 2.0 tool was applied to known surface proteins, with B-cell epitopes predicted from putative antigen sequences (Jespersen et al. 2017). B-cell epitopes are predicted from a protein sequence using a Random Forest algorithm trained on amino acids from both epitopes and non-epitopes, as determined from crystal structures; sequential prediction smoothing is performed afterwards.

This epitope prediction tool has been used for research of MPox, SARS-CoV2 and many other pathogens; since published in 2017 it has been cited 904 times (at time of writing). Simply, it qualifies the probability of a residue to be the centre of a heptapeptide with epitope presentation; a score above 50% indicates likely epitope.

2.7 Structural Modelling

Protein structural analysis was undertaken as part of the present study. The first step in this process was finding analogous structures in the Protein Data Bank (Berman et al., 2000; Gore et al., 2017) by searching with the EIV protein fasta sequences. Putative matches were then uploaded to ChimeraX suite (Pettersen et al., 2021) for structural and spatial examination. Initially, non-synonymous mutations detected at the consensus level were simulated *in silico* on homologs (with the 'swapaa' command) to observe possible impacts of mutant amino acids on local protein topography, as measured by changes to surrounding Ramachandran angles. Additionally, the likelihood of amino acid replacements by point mutations was quantified by a general PAM250 matrix to roughly estimate how costly and unlikely non-synonymous mutations would be.

Beyond using homologous protein crystal structures, protein sequences were used to predict the structures of each EIV protein. As the HA trimer is the only resolved structure for equine IAV, creating models of the other major proteins of EIV allowed observation of changes caused by non-synonymous mutations detected throughout our experimental transmission chain, as well as comparative analyses between predicted EIV proteins and those of other IAVs. Protein sequences were inputted to AlphaFold (Abbas et al., 2023; Evans et al., 2021; Varadi et al., 2022) both locally and on the online API provided by Google Code (Mirdita et al., 2022) which resulted in the return of predicted structures. These predictions were then post-processed in AlphaPickle v1.4.1 (Arnold, 2021) for quality control and prediction confidence.

2.7.1 Validation of Structural Predictions

The accessory tool AlphaPickle (Arnold 2021) allows for structural model validation with two main statistics: Predicted Aligned Error (PAE) and the Local Distance Difference Test (LDDT) (Mariani et al. 2013) which measures the local

distance differences between all atoms in a structure estimating confidence in the predicted model.

AlphaFold2 reports the quality of a structural model as a per-residue pLDDT score, which assesses prediction confidence. pLDDT ranges from 0 to 100, with higher scores indicating higher quality predictions. Accuracy of AlphaFold predictions is generally allocated into one of four confidence levels based on the pLDDT scores: high (pLDDT ≥ 90), medium (pLDDT < 90), low (pLDDT < 70) or very low (pLDDT < 50) (Abbas et al. 2023; Varadi et al. 2022). AlphaFold2 calculates the pLDDT score by comparing the distance between pairs of atoms in the predicted structural model with the corresponding distances reported in experiments using actual protein structures. This comparison of distances is performed for each individual residue, giving a final score reflecting the similarity between the predicted and experimental reference structures at each residue (Tejera-Nevado et al. 2023).

2.7.2 Comparing and Analysing Structures

In silico modelling of proteins with both consensus and mutant residues was used to explore the impact of nonsynonymous mutations on proteins. At the per-residue scale, changes to the polypeptide backbone can show the impacts of a specific point-mutation and so rotational changes may be quantified through Ramachandran or Dihedral angles.

Two torsion angles in the polypeptide chain (Sobolev et al., 2020) describe the rotations of the polypeptide backbone around the bonds between alpha carbons (C_α) and the amino group (N-C_α) called Phi, φ and secondly, the carboxyl group (C_α-C) called Psi, ψ. These φ and ψ angles are shown as green and blue respectively in Figure 2.4, adapted from Lennox et al. (2009). The range of the Phi & Psi Ramachandran angles accessible to a polypeptide chain defines the flexibility of the backbone and its ability to adopt a certain fold.

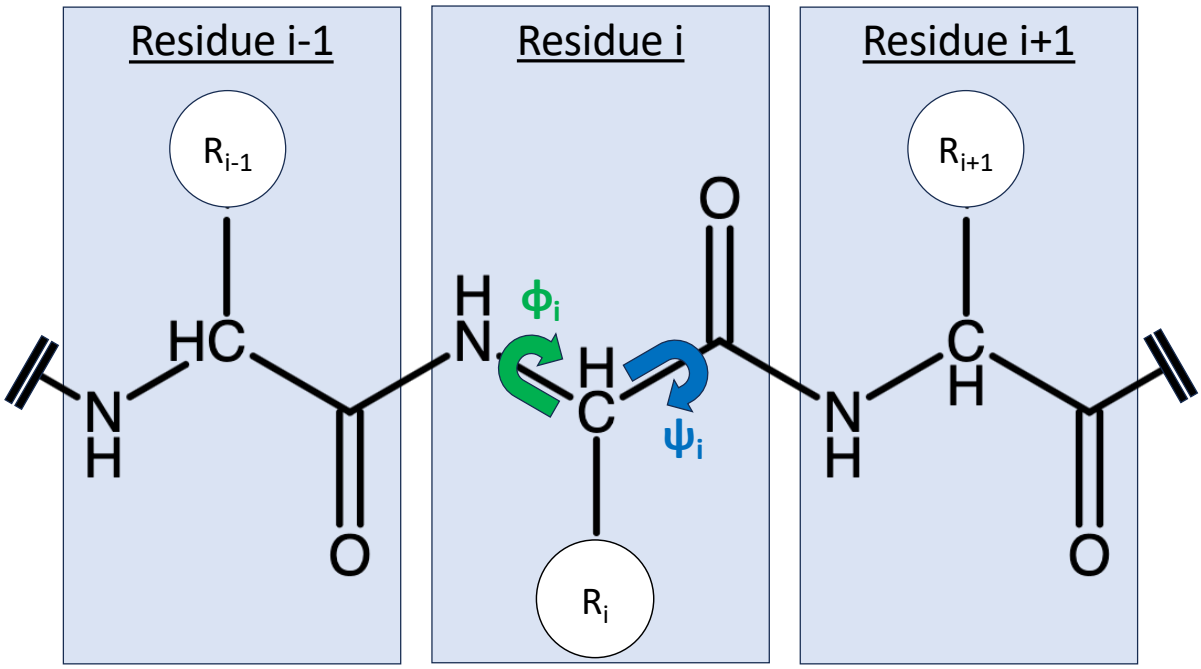


Figure 2.4: Protein backbone with labelled Ramachandran angles (ψ and φ) around a dihedral bond. White circles represent amino acid side chains. Adapted from Figure 1 of Lennox et al. (2009) and created using the Chemical Sketch Tool hosts by PDB.

Each simulated mutation gives a φ, ψ and chi (x) value associated with rotamer properties based on the Dunbrack rotamer library (Shapovalov and Dunbrack 2011). Both φ and ψ describe residue angles in relation to the protein backbone while x describes torsion of amino acid side chains (readers are encouraged to explore Ramachandran angles with the interactive tool on Proteopedia). As x reflects the orientation of side chains around the residue backbone, it has minimal impact on structural phenotype and so is not further analysed here. To note, glycine has no associated angles as it does not possess side chains.

Mutations were simulated on both structures with the intention of not only repeating *in silico* experiments for validity but to observe any differences in structural changes between the two models, assessing potential strengths or weaknesses of exclusively using either nearest-homolog crystal structures or simulated structures alone.

2.8 Transmission Bottleneck Estimation

The sizes of transmission bottlenecks were assessed using the exact Beta-Binomial sampling model proposed by Sobel Leonard et al. (2017). Code from the authors' supplemental materials was incorporated into R functions for ease-of-use. Transmission pairs were decided upon based on the known dates of co-housing between hosts.

The equation is composed of two probability distributions evaluated for each possible bottleneck value (N_b) as specified at the start of the function.

$$L(N_b)_i = \sum_{k=0}^{N_b} \beta bin\left(R_{var,\,i} \mid R_{tot,\,i},\, k, N_b\text{-}k\right) \times bin\left(k \mid N_b,\, v_{D,\,i}\right)$$

The first probability function draws from a ß-binomial distribution where, given the number of variant reads and the total number of reads, the probability of drawing a variant is defined by the variant frequency at that site.

The number of sub-consensus variants (R_{var}) is measured against the total number of reads (R_{tot}) at site *i*. The remaining terms mark the probability of success, i.e. the threshold at which a variant can differentiated from mechanical error which is usually 1%, and the number of trials. This likelihood value then populates a binomial distribution of probable bottleneck sizes for the number of successes in *k* trials, where the *probability of success* is given by the observed frequencies in the donor ($v_{D,\,i}$).

This matrix is then evaluated by a binomial distribution where each successful draw indicates an N_b value suitable for explaining the variants distributed across both donor and recipient hosts. Repeating this estimation for each possible value of N_b (set between 0-200 in our initial trial) then gives the probability for the bottleneck size most likely to lead to the viral population observed in the recipient host. This maximum N_b was then incrementally adjusted in steps of 200 for samples that showed estimates higher than 200. This incremental increase of the allowed maximum continued up to an N_b of 1000, accurately estimating all but one sample.

3 The Impact of Prior Immunity on Virus Shedding

The quantity of virus that an infected host releases into the environment over the duration of the infectious period can vary greatly depending on host population structure, environmental conditions and individual host factors. Host populations are rarely homogeneous and these differences can be reflected in the amount, duration and method of viral dissemination. Here, horses were infected with influenza virus in a transmission experiment and swabbed nasally to collect viral genomic material. Quantified by qPCR, the resulting data provided us with an insight into the amount of virus present in each host on each of the eight-day observation period. Predictably, vaccinated horses had smaller viral populations than hosts with no vaccine-mediated immunity. Vaccine composition differed however, and horses that received vaccines that matched the challenge virus to which they were exposed had substantially lower viral populations than hosts that received vaccines made from multiple different influenza strains. From this, we observe that even hosts with prior exposure to the infecting virus can be infected and furthermore, that hosts with vaccine-conferred adaptive immune memory to the infecting virus had smaller viral populations when they were infected.

3.1 Introduction

Influenza A viruses (IAV) infect a broad range of mammalian and avian species. Their negative-sense segmented RNA genome is comprised of eight reassortment-capable RNP complexes. Two of these segments, numbers four (haemagglutinin-coding, HA) and six (neuraminidase-coding, NA), are major determinants of antigenicity and immunogenicity; 18 HA and 16 NA subtypes have been detected globally, across a broad range of species. Like many Orthomyxoviruses, IAV have a rapid replication cycle and poor genomic-proofreading capabilities (Aeschbacher et al., 2015; Alves Beuttemmüller et al., 2016; Khan et al., 2021). Both of these factors, along with the lower fidelity of RNA-dependent RNA polymerases (RdRp) compared to DNA, result in high mutation rates of IAV (Laabassi et al., 2015; Lai et al., 2004; Landolt, 2014).

It is useful to consider the range host responses that affect virus load, including mechanisms that either limit viral infection of cells or destroy infected cells before they have the opportunity to sustain viral replication. Such defences include a) Intrinsic Immunity, via the anti-viral responses of somatic cells (McKellar et al., 2021; Yan & Chen, 2012); b) Innate Immunity mediated, for example, by natural killer cells and macrophages (Hartshorn, 2020; Hemmink et al., 2016); c) Adaptive Immunity mediated by B- and T-lymphocytes (Paillot et al., 2016). Our experiment is designed in such a way that the only independent variable under consideration is the adaptive immunity of the hosts, though variation is of course introduced by host heterogeneities we could not control for, such as innate immunity. More specifically, we are looking at differences caused by the adaptive immune response conferred by vaccination, following the work of Murcia et al. (2013) and Oladunni et al. (2021).

The viral load of a host is dependent upon multiple virus-host interactions. Studies by Amat et al. (2021) however show that over the course of an infection, H3N8 viruses attenuate slightly, causing less severe tissue damage in favour of greater cell-to-cell spread. This may also impact the overall viral population size as well as directly influence the amount of free virus able to be shed in mucus or droplet. As viral load is mediated by the interplay of virus and host factors (Ganti et al., 2021), our transmission experiment (detailed in the Methodology Chapter 2 Section 1.1) accounts for some viral genetic and phenotypic variability; all seeders were challenged with inocula developed from the same lab-grown strain and thus

should perform with roughly the same fitness. Though as the parallel transmission experiments were carried out some weeks apart the two inocula were not from the same batch; sequencing of the two inocula however showed homogeny between batches. Hence, changes to load must have either developed *de novo* in the virus during the outbreak or, more likely, be mediated by host factors.

Viral load represents the amount of virus present within a host and, while a helpful figure, it is often difficult to quantify other than in *in vitro* studies. Shedding, rather, gives the amount of virus a host expels into the environment; this is an important metric when estimating direct and indirect transmission and can be used as a proxy for viral load. Hence, to better understand these relationships between host factors and the amounts of virus shed, we performed a transmission experiment which involved infecting horses with an equine influenza virus (EIV). Using natural transmissions between vaccinated and unvaccinated horses, we assess the impact that host adaptive immune response has on the amount of EIV they shed on a daily basis.

An important concept to appreciate when considering the host immune response in this form of experiment is the theory of the Original Antigenic Sin. This theory, put forward initially by Francis (1960), concerns adaptive immune recognition of closely related virus strains. When first exposed to IAV, immunocompetent individuals will mount innate, and subsequently, adaptive immune cascades resulting in the generation of memory B cells with corresponding epitope-binding antibodies. These memory responses enable rapid re-activation of adaptive immunity should the immunogen appear in the body again. The Original Antigenic Sin hypotheses deviates from the classical concept of adaptive immunology, which holds that novel memory responses are generated for each new pathogen encountered. OAS instead posits that newly encountered antigens sufficiently similar to ones already responsible for generating the memory B cell repertoire will trigger reactivation of the existing adaptive response rather than stimulating a novel clonal expansion cascade. Thusfar OAS has mostly been studied in mammalian influenza systems, ranging from *in vivo* mouse studies by Kim et al. (2009) to *in silico* ordinary differential equations constructed by Pan (2011) with a review of 23 human and animal experiments presented by Yewdell & Santos (2021). However, older work on rodent-borne arenaviruses (lymphocytic choriomeningitis virus) from Klenerman & Zinkernagel (1998) shows the same OAS dynamics seen in IAV infections.

Rather than undergoing new clonal selection processes, OAS hypothesises that the “good enough” binding of previous influenza memory cells will forego generation of a novel B cell repertoire and instead reuse the existing memory cells to stimulate an adaptive immunity reactivation cascade (Monto et al., 2017). This set of imperfectly matched antibodies (Figure 3.1) are then capable of binding to pathogen epitopes, but importantly at a reduced efficiency compared to antibodies generated to the primary influenza exposure. Original Antigenic Sin theory proposes that the strength of an adaptive immune response to a completely novel influenza strain may in fact be greater than the response to an IAV strain that only moderately differs from one to which the individual has pre-existing immunity and, consequently, OAS is sometimes also referred to as ’antigenic seniority‘ (Henry et al., 2018). The theory has been contentious since its proposal, but evidence generated by Simonsen (2004) among others, showed that influenza exposure early in life grants lifelong immune protection to hosts against strains similar to the eponymous “Original Antigen” with which they were first infected.

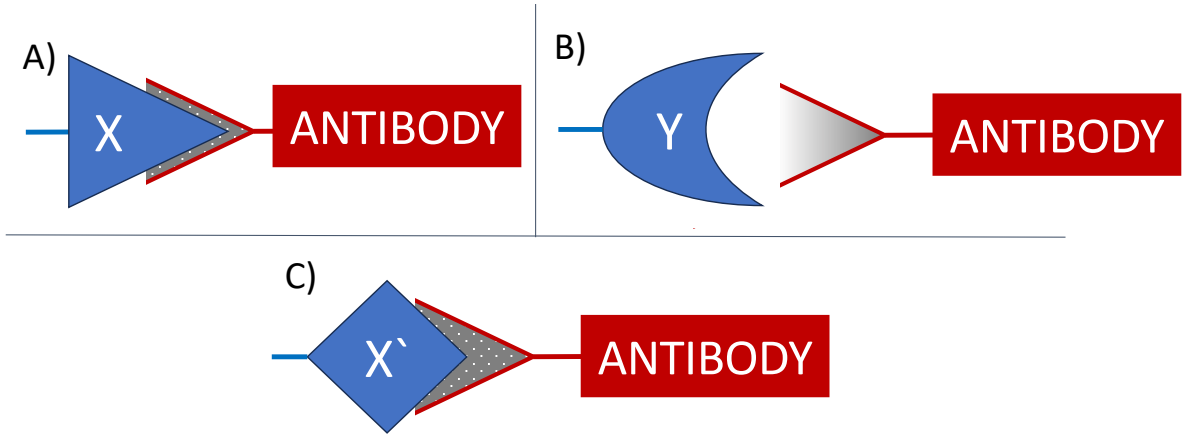


Figure 3.1: Cartoon representation of Original Antigenic Sin. A) On first exposure, antibodies bind to epitope X. B) Antigenic shift presents an entirely new viral epitope for the host to respond to. C) However, on exposure to a similar, but antigenically distinct virus, antibodies matching epitope X can still bind the novel epitope X`, but imperfectly.

3.1.1 Why is shedding important?

The viral load of a host likely affects the amount of infectious virus that individual sheds into its' surrounding environment (Perglione et al., 2016). We can assume that the longer an individual sheds, the longer they can transmit virus. In a compartmental epidemiological SEIR (Susceptible-Exposed-Infected-Removed) model (Figure 3.2) the viral load will impact the recovery rate (γ) of hosts, assuming that the recovery rate is a proxy for the rate at which hosts cease shedding virus and increase the size of the infected pool. Therefore, we would expect that longer durations of shedding infectious virus will have epidemiological consequences. However, a host that sheds a greater quantity over the same average time period, has an increased transmission rate (β), representing higher infectivity (Heesterbeek, 2002; Matthews & Woolhouse, 2005). We would therefore expect hosts that shed a greater quantity of virus and/or shed for a longer duration to have elevated force of infection (β), essentially meaning that they have a higher chance to infect secondary hosts. Thus, the quantity as well as duration of a host's shedding, when scaled up to the population-level, can influence the overall dynamics of an epidemic. Finally, regarding the evolution of the pathogen itself, shedding indicates sustained infection, so a longer period of shedding shows that the virus is actively replicating for longer. Therefore, with more replication cycles within the host, the more likely stochastic mutations will appear in the viral genome thereby creating a greater pool of diversity upon which selective processes can act.

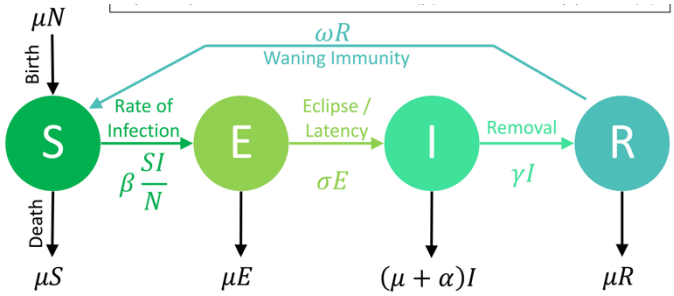


Figure 3.2: A standard SEIR model, showing four groups and the interacting dynamics between them. Viral shedding of hosts especially affects transmission rates (β) and the latency period (σ).

3.1.2 What is known about IAV shedding?

Epidemiological implications of viral shedding have long been explored, and since the 1918 H1N1 pandemic a great deal of attention has been given to influenza viruses. Understanding the shedding of IAV infected hosts has given insight into pandemic preparedness models (Ferguson et al., 2005, 2006), critical community thresholds for widespread vaccine coverage (Aoki & Boivin, 2009; Ip et al., 2017) and guided public health decisions (Lau et al., 2010; Liao et al., 2010). These data are however understudied in non-human influenza epidemic dynamics.

Values of other EIV and IAV studies, displayed in Figure 3.3:

- Equine Influenza
  - $10e^6$ copies/µl at peak of shedding (72 hours post contact) in experimental transmission of naïve hosts (Murcia et al., 2010)
  - in vaccinated horses experimentally exposed, shedding averaged $8.4e^4$ copies/µl, though on most days was around $10e^5$ copies/µl and peaked at $10e^6$ copies/µl in one host (Murcia et al., 2013)
  - viral shedding from horses across yards in a UK outbreak averaged at $6.37e^3$ copies/µl (Hughes et al., 2012)
- In other hosts:
  - during transmission experiments, vaccinated pigs shed less virus than naïve ones (an average of 71 compared to 281 copies/µl in the naïve pigs) (Lloyd et al., 2011).
  - clinical samples from human patients averaged 6.25, or 5.02 $\log_{10}$ copies per ml depending on whether the sample tested ELISA positive or negative respectively (Ward et al., 2004).
  - in testing oseltamivir treatment for humans, To et al. (2010) recorded loads of $1.84e^8$ copies/ml in $H1N1_{pdm2009}$ infections and $3.28e^8$ copies/ml in patients with seasonal IAV strains
  - observing swine and barns in southern Minnesota, Neira et al. (2016) reported $4.03e^7$ copies/ml in saliva samples, $4.16e^7$ copies/ml on railing surfaces and $1.25e^6$ copies/m^3 of sampled indoor air

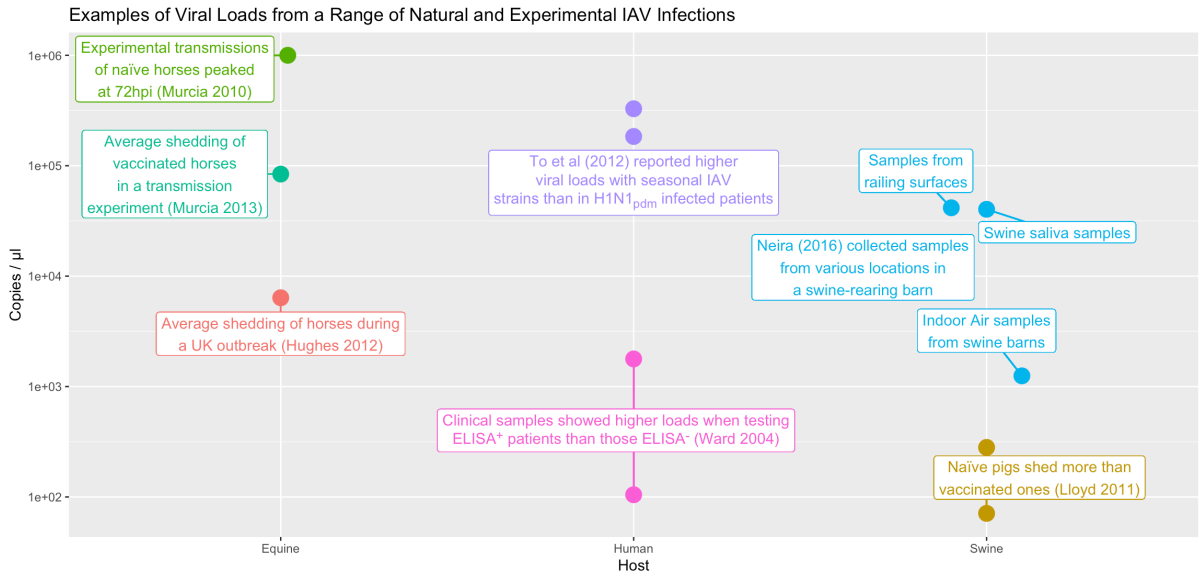


Figure 3.3: Examples of influenza viral loads of hosts under various conditions

3.2 Results

3.2.1 Viral Shedding

Finding the relationship between C_t Values and the concentration of the standard curve, we can examine at which point the calculated qPCR values become unreliable. Here, the ability of Ct values to explain the copy numbers present really breaks down at a C_t of 32.25, corresponding to around 80 copies/µl (detailed in Chapter 2 - Methodology). Thus, we will use this cut-off point going forward. This conservative estimate of when there is so little virus present in the sample that the qPCR become stochastic also indicates a sparsity of viral genomes to with which to establish further infections.

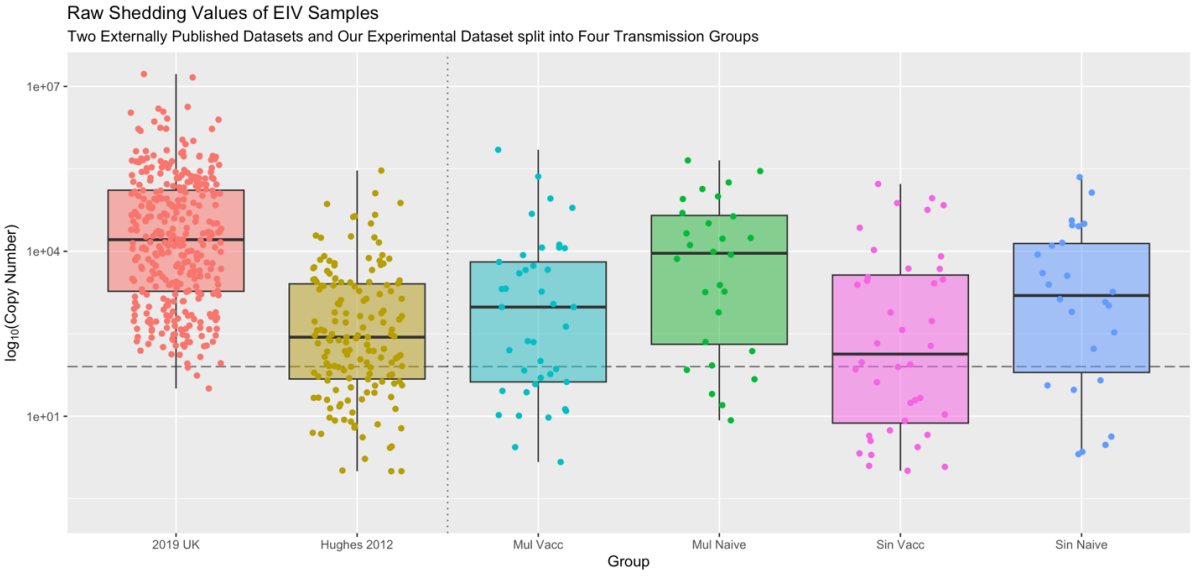


Figure 3.4: Copy numbers/µl of EIV in naturally infected hosts, our transmission study lies to the right of the dotted line. Trajectories of shedding from individual hosts from the experiment are presented in further detail in Supplementary Figure 3.2

3.2.1.1 Viral Loads

Averaging the copy numbers of hosts into sets based on their transmission chain and immune status, we have four distinct host classes: Vaccinates in the multi (V_M) and single (V_S) transmission chains and Naïves in the multi (N_M) and single (N_S) transmission chains. The raw qPCR values are shown in Figure 3.4 alongside other collated EIV datasets. Nasal swabs were used to quantify viral loads, $\frac{137}{160}$ swabs gave positive qPCR values (68 from the Single group: 42 vaccinates [V_S] & 26 naïve [N_S], 69 from Multi: 41 vaccinates [V_M] & 28 naïve [N_M]) were collected.

The samples show substantial variation between hosts and even within hosts day-to-day. Within an individual, this variation in viral shedding is expected as the population exhibits well-described growth kinetics. To better observe impacts that host factors may have on viral shedding, qPCR values were averaged into an epidemiological class showing the average viral load on each day of observation plus the total area under the curve (AUC) in Figure 3.5. The peak shedding in most groups occurs two days after contact with infected individuals. As shown graphically and in

the table of average population sizes (Table 2) on the day of peak viral load, shedding in the multi group is marginally higher than that of hosts in the single group regardless of whether the host was vaccinated or naïve.

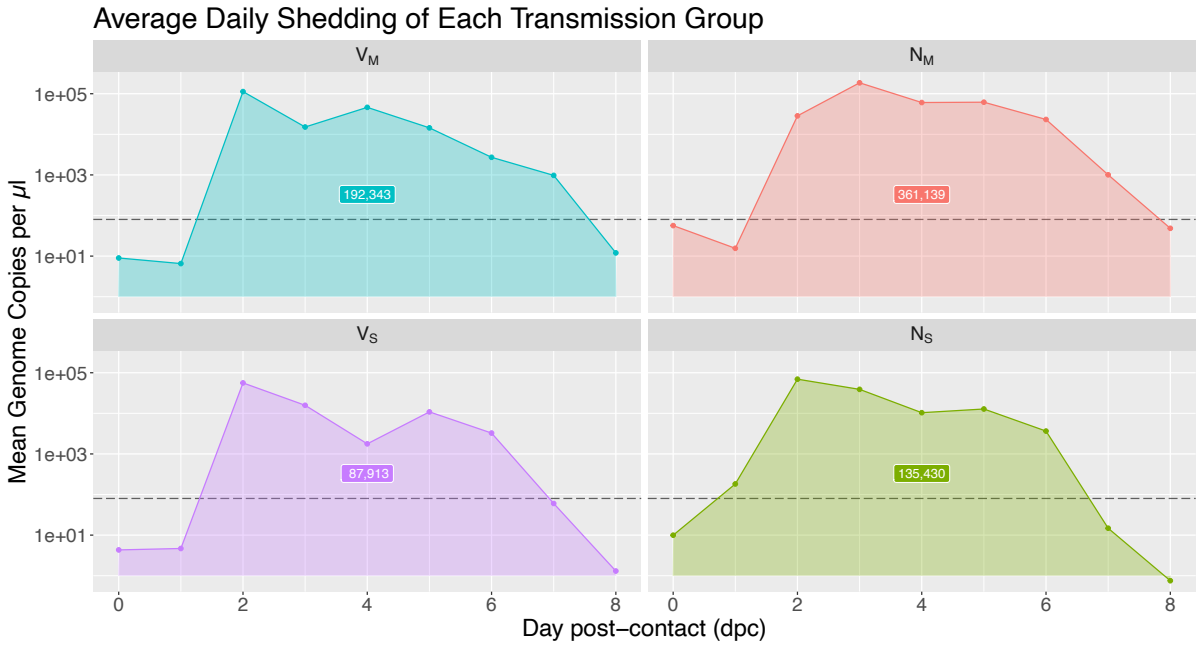


Figure 3.5: Shedding values averaged across epidemiological groups for each day post-contact with an infected individual. Annotations show the mean population size of each group, measured as copy numbers per µl of transport media.

Notably the V_M group sheds more than the N_S; this suggests that previous exposure to multiple viral strains offers less protection to the host than no vaccination at all, on the condition that preceding hosts in the transmission chain have been immunised to the specific challenge strain. However, to note, by transmission into naïve hosts the two viral populations were not identical across the Multi and Single groups and diverged at the consensus genome level.

The duration of shedding is important to observe; we would expect a greater total quantity of shed viruses to correlate with infectivity, but hosts shedding the same quantity over a longer period may have slightly different epidemiological implications. Especially for an acute virus spread in a density-dependent manner, such as IAV, a longer infectious period increases the number of potential contacts an individual may encounter.

Most hosts, in both transmission chains, shed for at least 3 consecutive days, with naïve hosts usually shedding for 5 days. There are a few occasions where a host will stop shedding for a day and then bounce back, such as Multi 4A and 4B. Hosts 4A & 4B in both transmission chains both have very low viral loads and tend to barely broach the threshold before dipping back below it again (Figure 3.6).

The second set of graphs show the specifics of transmission events just between co-housed hosts, the solid line representing the donor hosts and the dashed always the recipient horses. The grey box in the background represents the period in which the recipient

Table 3.1: Shedding of each group on the day of peak, and the day that shedding peaked, plus the total viral loads of each group.

Group	Day	$\log_{10}$ Copies/µl	
		Peak	Total
V_M	2	5.61	6.16
N_M	3	5.81	6.06
V_S	2	5.38	5.73
N_S	2	5.21	5.72

pair were infected by the donor, based on viral loads, known dates of sample collection and sequence analysis.

In the multi-chain, the 3-4 event shows a slightly lower load in 4A and 4B and then the 4-5 event shows the population rebounding as it enters naïve hosts. But in the single chain, the 3-4 event almost crashes the population and event 4-5 looks like transmission almost sputters out. Onward transmission is only possible because pair 5 is naïve rather than being another pair of vaccinates, thus mirroring the same "rescue" of transmission chain by naïve hosts seen in both Jiao (2021) and Parsons et al. (2024). This suggests that the adaptive immune response in horses that received a multivalent vaccine, those representing hosts with a life history of multiple exposures to different IAVs, inhibits viral growth to a lesser extent than in hosts with specific immunity to the challenge strain.

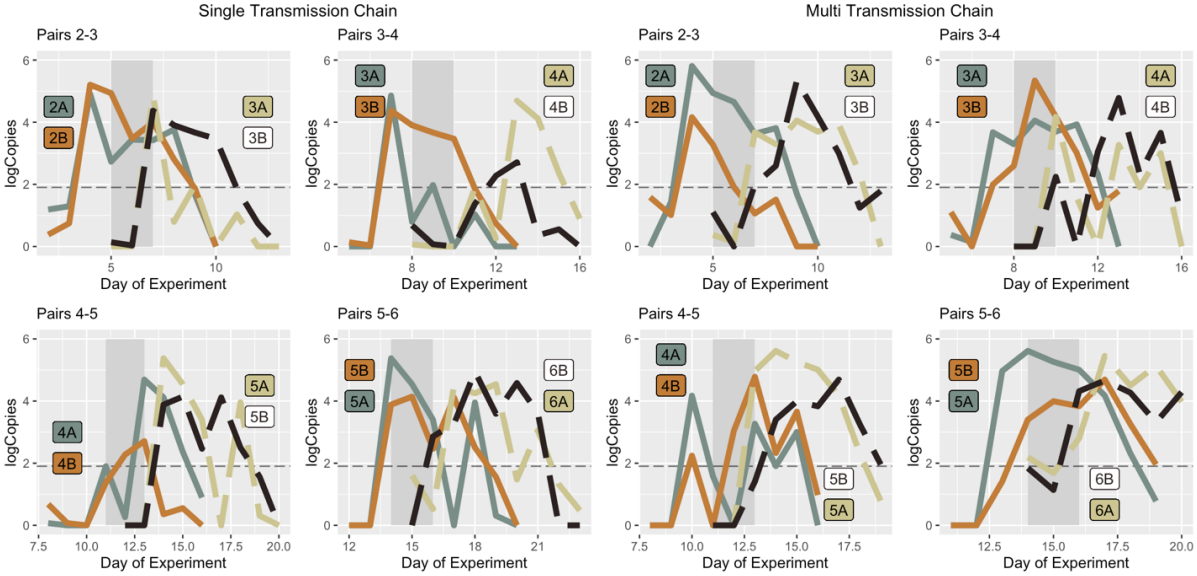


Figure 3.6: Focus on the transmission events between each pair of hosts. Shaded areas indicate the period in which recipient pairs were assumed to be infected.

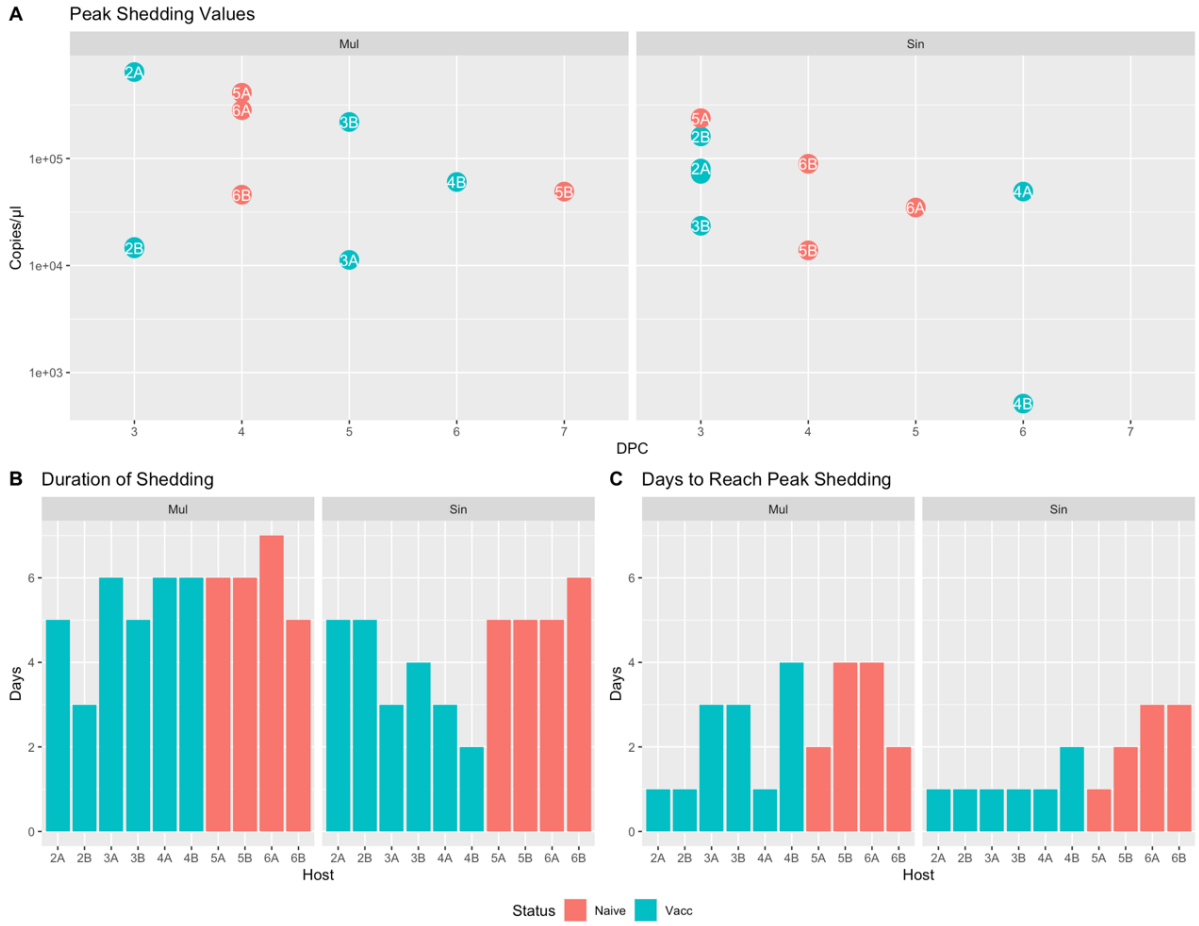


Figure 3.7: A) The day at which a host peaks in their shedding, and the copies/µl of that peak. Most of the hosts in a pair peak on the same day. Shown in B) the number of days a host is positively shedding and C) the number of days post-contact until a host becomes shedding-positive.

3.2.1.2 Non-Parametric Tests

Data from all 137 samples did not appear normally distributed with either a density histogram or a qq-plot. To verify this, a Shapiro-Wilk test was used to assess whether the copy numbers were actually normally distributed. A p-value of $2.2e^{-16}$ provided a strong indication that the residuals of the mean copy numbers did not follow a Gaussian distribution (W=041014). On this basis, non-parametric hypothesis tests were used in the ensuing analyses of copy numbers (further information in Supplemental 3.3). A mono-sampled Kolmogorov-Smirnov test was used to compare the data first with a continuous normal distribution and then with a continuous Cauchy distribution, to examine whether the mean copy numbers align with either of these distributions. The mean copy numbers we observe did not adequately fit either probability distribution and so were definitely non-normal.

To understand whether host factors impact shedding, differences in the means in shedding quantities between experimental groups (Multi or Single) and Exposure history (Vaccinated or Naïve) were tested. As data were non-normally distributed, the tests applied included the Kruskal-Wallis and Pairwise Wilcoxon Rank Sum Tests. The Kruskal–Wallis test is a non-parametric method for testing whether samples originate from the same probability distribution and for the present dataset it was used to ask whether copy numbers were associated with particular host factors. A significant Kruskal–Wallis test would indicate that at least one value in the dataset

is associated with the host factor in question. The Wilcoxon Rank Sum (Mann-Whitney U) Test compares the probability that a sample drawn from one group is greater than a sample from the alternate group, in this case 'is a randomly drawn copy number *more* likely to come from a vaccinated host than from a naïve host?'.

- Transmission Chain
  - Both tests were unable to state a significant difference in shedding between individuals in the two transmission groups (Kruskal-Wallis chi^2= 3.4371, df = 1, p-value = 0.06375).
- Vaccination Status
  - Wilcoxon Rank and the Kruskal-Wallis testing determined that shedding from vaccinated individuals was consistently different to that of the naive hosts (Kruskal-Wallis chi^2= 5.8987, df = 1, p-value = 0.01515).
- Group
  - Continuing to use the Wilcoxon paired rank tests, differences between hosts according to their transmission group and immune status were compared and displayed in Figure 3.8. The only groups that displayed significant differences in shedding were the naives and vaccinates in the multi group (N_M: V_M, p-value = 0.044) and the vaccinates in the single group compared to the naïves of the multi group (N_M: V_S, p-value = 0.009).

With vaccinated horses in both transmission chains shedding a significantly lower amount than naives of the multi group, the vaccine clearly helps decrease shedding, regardless of its composition. However, seeing that neither vaccinated group shed considerably differently to the naïves in the single transmission chain is likely a statistical artefact caused by the great deal of variation in the N_S group. A 'difference-of-means' test like those used above does not consider variation in its estimation. To account for this, general additive models further explore these relationships below.

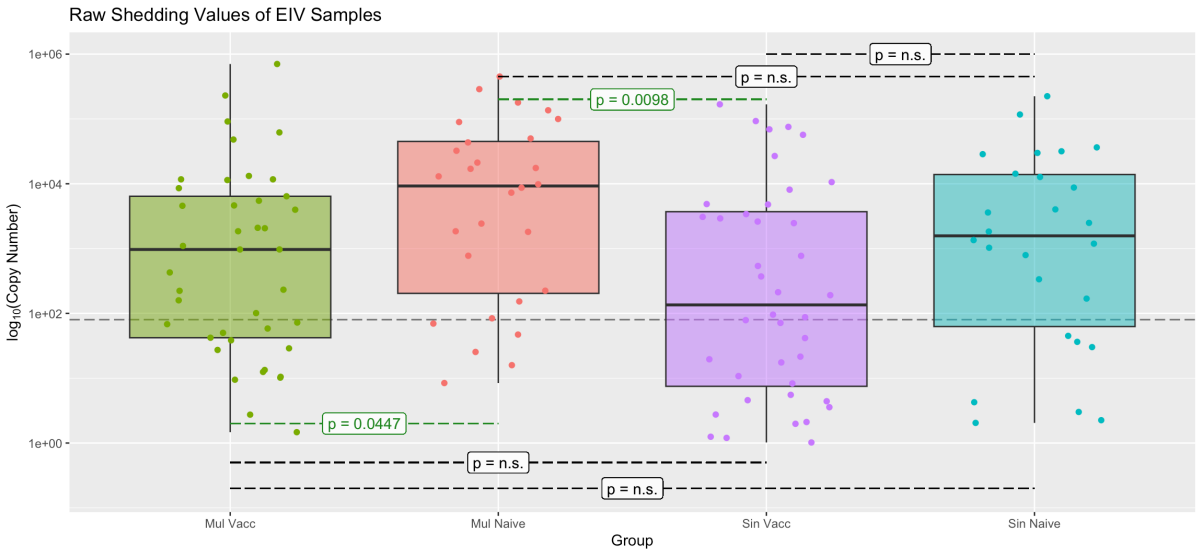


Figure 3.8: Copy numbers of all samples, coloured according to epidemiological groups. Dashed lines connect boxplots showing the results of Wilcoxon rank sign tests, and coloured green if statistically significant.

From the consensus sequence analyses, detailed in Chapter 4, we do detect two distinctly different virus populations by the end of each transmission chain. After

transmission through vaccinated hosts in the single group, a non-synonymous mutation (NP g1445a) becomes fixed in the viruses. The final virus population at the end of the multi group, however, has two fixed mutations, both in segment 2 (the synonymous PB1 t1500c and the non-synonymous PB1 a1853c). These two different viruses may have differing fitnesses, potentially explaining why the viral loads in N_S hosts is more similar to vaccinates than that of N_M hosts. That such a drastic shift in fitness could be mediated by a single non-synonymous mutation away from the consensus in each group (making a distance of two non-synonymous mutations in total between the N_M and N_S viruses [see haplotypes F and J in Chapter 4 Section 3.1.1]) is unexpected. Changes in viral shedding could be caused by viral mutations and/or changes in the host response.

Table 3.2: Model coefficients and resulting viral load estimates for each epidemiological group.

Group	Coefficient	Total log_{10} copies
V_M	12.6	12.6
N_M	-1.5	11.1
V_S	-2.1	10.5
N_S	-3.6	9.0

3.2.1.3 Regression Models

Models of viral shedding were constructed which included the host's transmission group (*Group*), whether the host was vaccinated or not (*Status*) and the day on which the sample was taken, measured from the day that the host was first exposed to infected individuals (Day Post-contact, DPC) as explanatory variables. The first step was to determine whether to use the total amount of virus shed as the response variable or to treat the viral loads of each day as independent response variables. These models were compared in a pairwise manner and with two main ranking processes: average posterior predictive distribution (PPD) and LOOIC (leave-one-out information criterion) (Vehtari et al., 2022). Ultimately, the total amount of virus shed by each host was selected as the most informative and statistically well-supported response (ΔLOOIC = 494.9, supplementary 2.1b).

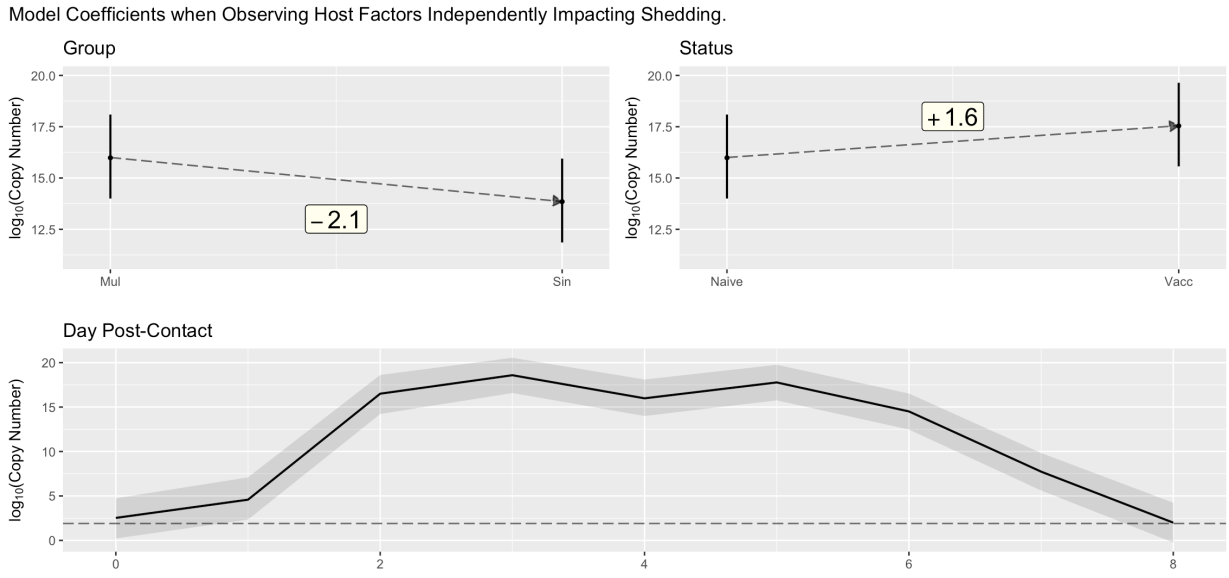


Figure 3.9: Model outputs of the observed effects of the day post-contact, epidemiological group and finally the joined effects of both effect variables.

Observing the modelled viral load over the course of infection, as seen with the real data, viral shedding peaks on the second day and begins to decline from the sixth

day. We can also see that shedding is consistently highest in the V_M group (Table 3). This does not match what is seen in the real data, where N_M hosts shed at a higher level, as expected from a biological perspective. This is likely due to over-dispersion in the model caused by having more data points in the V_M group than N_M, and a broader range of copies in the V_M group. Between the epidemiological groups, hosts in the single transmission chain shed less overall than those in the multi chain, with the total virus shed in the multi group being $2.1e^{12}$ compared to $1.6e^{10}$ across the single chain. Further, the viral load of hosts is heavily influenced by their previous history of EIV exposure; between both transmission chains, naïve hosts shed a total of $6.3e^{10}$ whereas vaccinates shed $2.0e^{12}$ in total.

The higher quantity of viruses shed by hosts of the multi transmission chain suggests a greater degree of uncontrolled infection than those in the single chain. As explored previously, the previous exposure to the specific inoculum seen in the vaccinates of the single chain reduces the amount of virus shed overall.

3.2.2 Transmission Events

From the date of first vaccination (April 1st 2008) to the beginning of each experimental transmission chain (multi: Jul 22nd 2009, single: Oct 14th 2009), all vaccinated individuals had many months for the initial vaccine-responses to acquiesce. However, this extended period may introduce waning immunity of the hosts to our considerations. Sera were collected from the vaccinated hosts for haemagglutinin inhibition tests in order to measure the strength of immune responses by circulating antibodies. The transmission experiment began once passively circulating anti-influenza antibodies had fallen to levels indicating a return to immune senescence, i.e. a value low enough to make them susceptible to infection. Based on both equine influenza vaccine efficacy tests (Wood et al. 1983) and previous EIV transmission experiments (Murcia et al. 2013), this was determined by a single radial haemolysis (SRH) value less than $60mm^2$. Hosts only entered the transmission experiment once it was believed that viral exposure would trigger a secondary, memory immune response. Thus, antibodies raised in response to vaccine antigens would have subsided and any new humoral response would be driven entirely by the hosts' adaptive immune memory (Appendices Supplementary Figure 3.1). Hence the delay between the two experiments; as vaccinated hosts in the Single group mounted a longer-lasting response than those in the Multi group, additional weeks were needed before starting transmission.

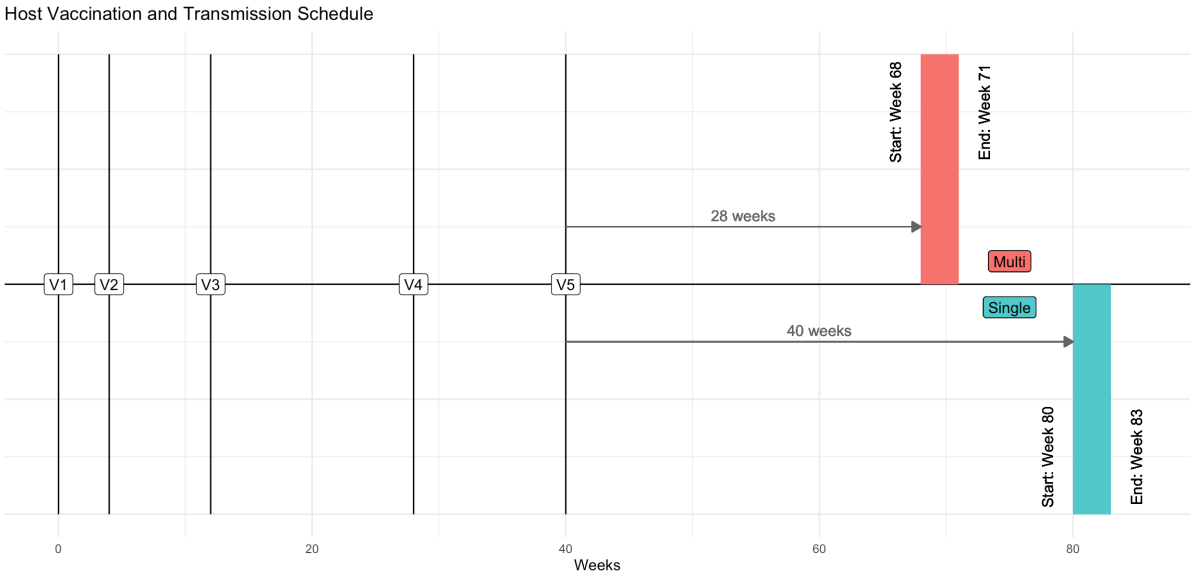


Figure 3.10: Diagram of the vaccine regimen, and the time each experimental transmission chain began. Dates of vaccine administration (V1-5) reference the schedule laid out in Table 1.

Therefore, vaccinated hosts were still susceptible to infection; each host has at least 3 samples surpassing the threshold of 1000 copies from nasal swab qPCR assays.

All transmission events were successful across the experiment; there was always at least one donor and one recipient host wherein sufficient viral particles were transmitted. We are, therefore, unable to gauge the level of the infective dose from this experiment and therefore we cannot model whether or not transmission may die out. However, under the assumption that passage through vaccinated hosts places continually tighter bottlenecks on viral population size (as measured by copy numbers), EIV transmission will eventually halt as the infected, vaccinated donor sheds too little virus to establish infection in a subsequent vaccinated recipient. From this, I hypothesise that a transmission chain with many vaccinate-vaccinate transmission events will be prematurely shortened compared to a chain with more heterogeneity, i.e. fewer transmission events exclusively between vaccinated hosts.

Viral transmission is wrought with stochastic bottlenecks, which may limit the ability of a founder population to establish infection in a recipient host, and this may lead to epidemic burnout. To examine the implications of this on a homogeneous population of horses, the shedding data were used to simulate conditions of an outbreak in which all hosts are vaccinated. These models are extrapolating the trends seen in the actual experiment and are supported by the transmission study of Murcia et al. (2013) in which EIV transmission halts after a host sheds insufficient virus to cause a subsequent infection. In that work, horse 'V4' shed an average of 64 copies daily (±101) and a total of 1221 over the course of the experiment, and was unable to infect horse 'V5'. From this we can safely assume a minimum of 1,000 virions are needed to establish infection (explored further in Methods 2.2.3).

By quantifying the total and daily average amount of virus shed by each host, we can estimate the size of the outbreak as a whole. Summing the amount shed by all hosts in each transmission chain (Table 2) we see $10e^{6.41}$ and $10e^{6.03}$ copies in the multi and single chains respectively. Of course, shedding is not distributed equally, and the size of the founding population can determine the speed and overall viral load of the newly infected host. Each host received at least the minimum infectious dose to establish infection, but a host exposed to $10e^{4}$ viruses is much more likely

to be infected **and** to reach a higher viral load than a host exposed to only $10e^2$ viruses as stochasticity in the establishment of an infection means that a larger population is less vulnerable to change.

Using the additive models created above, the significant host factors in determining shedding profiles can be utilised to make predictions on the shedding of subsequent hosts. Primarily, the transmission chain is simulated to have additional vaccinated hosts. As horses in natural settings are recommended to be fully vaccinated, this simulation could help characterise transmission events between vaccinated hosts; although the implementation of these recommendations likely varies between countries.

3.2.2.1 Linear Predictions

Linear predictions of the viral shedding based on the host's position in the transmission chain, and to which transmission chain they belonged, can be used to extrapolate the amount of virus shed in further hypothetical vaccinated hosts rather than naïve ones. Initially, predictions were based on the mean shedding values and the trends observed between these means. However, they proved inadequate due to their simplicity. The results are briefly discussed here, demonstrating the methodology used and the naïve assumptions such linear models make. Indeed, creating and interpreting these models was a step of the analyses but only a preliminary one; it is included here to demonstrate the analytical process and lay the foundation for the development of the more complex models below.

Using the data from vaccinated hosts, a linear model was created and shows that the data have a strong signal of decline. To note, this incremental decrease in shedding is unlikely to continue at a fixed rate. This modelling is meant to highlight the introduction of stochasticity in transmission events, as hosts shed lower quantities of viable virus the probability of infected hosts shedding sufficient to cause secondary infection falls. Though vaccinated hosts in both transmission chains shed slightly different amounts, an estimated $10^{6.16}$ in total in the multi transmission chain and $10^{5.73}$ in the single chain, the slopes at which shedding decreases incrementally with each transmission event both show the same pattern. Each transmission event reduces the amount of virus shed overall by $10^{0.345}$ or $10^{0.195}$ for the multi and single transmission chains respectively. Hence, eventually, each host is shedding so little virus that on no day are copy numbers above 80 copies/µl ($10^{1.09}$), at which point we can assume that onward infection is not viable. The above models may be expanded upon by adding variation in shedding seen across different days, and so we can further estimate transmission between vaccinates.

3.2.2.2 Additive Modelling

Using a Bayesian additive model in order to better represent the margins of error and standard deviations, the shedding of hosts vaccinated with different vaccine types can be assessed (Figure 3.11). Range-bars represent the variation over the eight days of observation and, as before, the models predict that shedding across vaccinates in the single group declines sooner and more rapidly than vaccinated hosts in the multi group. This contrasts the previous model, in which the multi group showed a steeper decline.

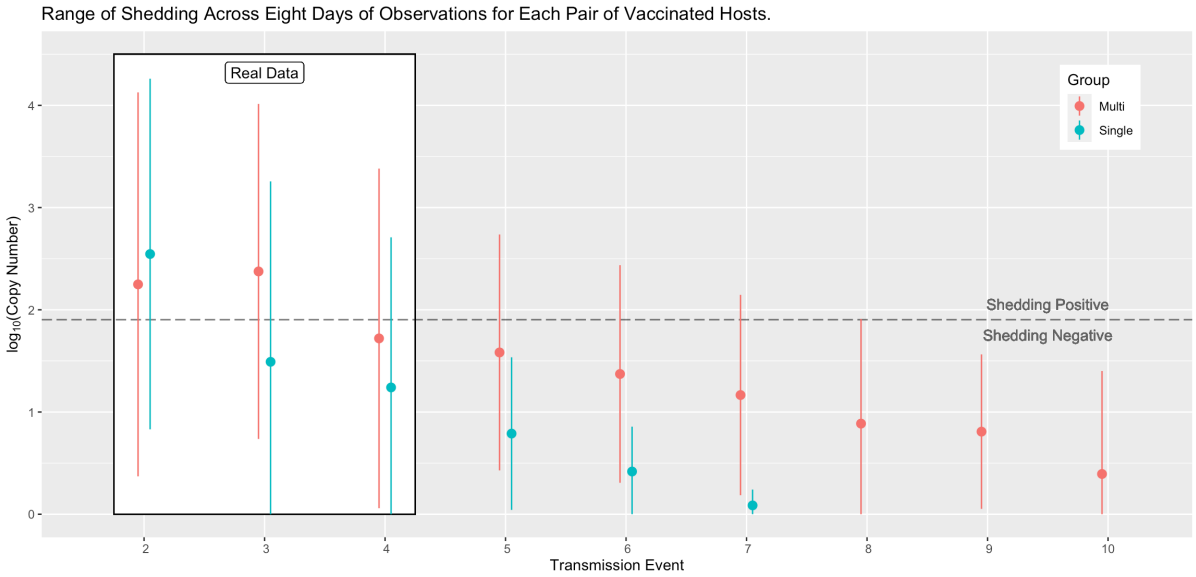


Figure 3.11: Mean shedding, at the level of both hosts in a pair, over the eight days of observation.

Incorporating variability in shedding influenced by the day of shedding into this model thus aims to provide a more comprehensive picture than above, wherein the more simplistic model assumed equal rates of shedding for the entire infectious period. It is unlikely that the 5th hosts in the single chain would shed sufficient virus to subsequently infect a 6th pair of vaccinated individuals. In the multi group, on the other hand, transmission appears to remain viable until the 8th pair. According to the model, EIV will stop local spread after passage through either 5 or 7 pairs of vaccinated hosts, in the single or multi group respectively.

3.3 Discussion

This chapter deals with an initial analysis of viral shedding data from two related transmission experiments, both beginning with the inoculation of immunologically naïve horses with the Newmarket/5/03 EIV strain. Each transmission chain was composed of three pairs of vaccinated hosts followed by two pairs of naïve horses. The vaccine regimen for each transmission chain differed, allowing a total of four distinct experimental groups to be defined: hosts vaccinated with multiple strains of EIV (V_M) and corresponding naive hosts (N_M) and hosts vaccinated with only Newmarket/5/03 EIV antigens (V_S) and the corresponding naïve hosts (N_S). Once they tested positive for EIV infection, nasal swabs were taken from each host daily in order to quantify their viral load with the use of qPCR. From the results presented in the previous section, it can be observed that naïve hosts in the multi group shed substantially more virus than any of the other groups. A high degree of shedding implies a large viral population within the individual over the entire course of EIV infection, the size of which has a major influence on viral evolution. Primarily, a larger population is more resilient to deleterious mutations and also provides a broader pool of potential sites of mutation and/or selection. This, however, is coupled with the disadvantage of dilution of beneficial mutations, as the population may be so large that selective pressures act weakly. Finally, the more virus shed into the environment, the more likely onward transmission is to occur, either directly or through fomites.

A secondary cut-off point was considered, originating from an external EIV transmission study (Stack et al., 2013) which measured a break in transmission once

hosts were shedding beyond this 34.81 Ct value, around 15 copies/µl. However, this study focused exclusively on transmissions between vaccinated horses and so is less informative for transmission chains of heterogeneous hosts. Hughes et al. (2012) defined a horse to be positively shedding once viral load numbered above 150 copies/µl (due to the limits of false positive detection in qPCR), though this study focused on samples collected during a national outbreak of EI and so has a higher threshold than our experimental data require.

The transmission group to which a host belongs moderately influences viral shedding, with $10e^{2.1}$ lower shedding being observed in the single group compared to the multi group (n_{eff} = 74%). The immune status of the host also was found to influence viral load quite dramatically, with vaccinated horses demonstrating $10e^{1.6}$ lower shedding than their unvaccinated counterparts (n_{eff} = 72%). As expected, over the course of infection, the amount of virus shed depended heavily on the number of days since the host was first exposed to the pathogen. Generally, the shedding curves had a bell shape, accelerating from days zero to two, peaking on day two or three and then decreasing over days six to eight. This short period in which an infected host is actively shedding virus and constant close-contact with susceptible hosts during this period is crucial to viral spread during outbreaks. However, it must be noted that horses in the vaccinated class of this experiment had recently undergone an extensive vaccine regimen of five exposures over a period of 40 weeks, which was much more intense than that normally administered in the field, where boosting is recommended every six months. In natural, field settings waning immunity and even exposure to other, non-IAV pathogens would be expected to decrease the strength of the immune memory response. The stark difference in shedding between classes seen here would thus be expected to be less dramatic and more closely resemble one another.

Decreased shedding in vaccinated individuals makes biological sense; a pre-armed immune system mounts a stronger and faster response to infection. Thus, the infecting viral population is suppressed and so the overall viral load is much reduced. This suggests that in natural infections, vaccinated individuals may shed a lower amount of virus throughout the course of infection. Temporal patterns of shedding did not differ substantially between transmission chains; almost all hosts appeared to spend a day or two with low-detectable, non-shedding loads before broaching the defined threshold. Then, once viral loads averaged >1000 copies/µl, hosts were considered as being infective to other horses, with virus being shed for up to three days before dropping back to undetectable levels. The duration and intensity of shedding can influence the likelihood of a host infecting another susceptible individual, a large factor in predicting epidemic properties. As shown, the type of vaccine a host received affected the overall viral load present across the transmission chain, even if the effect on individual hosts itself was minimal. In both transmission chains, the average and total viral load decreased with each pair of vaccinates, i.e. pairs two to four.

The continuous decrease in shedding observed may not necessarily be indicative of epidemic burnout and hosts may eventually shed enough to seed infection in other susceptible individuals. But the important point is that the need to shed for a greater amount of time to reach this threshold necessitates that hosts remain in close contact for longer durations than normal. It is this requirement that increases the chances of burnout; the conditions required for ordinary transmission become more and more demanding, to a point that viral transmission becomes improbable.

Additionally, quantification of viral load in each host comes exclusively from daily nasal swabs. This is clearly missing out on an important, well-known feature of IAV transmission; that of mechanical transmission via fomites. Virus shed into the environment may well be able to supplement, if not entirely substitute, the initial infective dose required to establish infection in a new host. Our results still indicate that such low levels of shedding as seen in vaccinated hosts predict that further down the transmission chain infected individuals would be unable to initiate infection alone. However, with added viral challenge from surrounding fomites, infection may still be established in new hosts.

These models imply that, were the host population *all* vaccinated, there would eventually be a point at which hosts shed too little to establish subsequent infections. This effect is of course two-fold, mounting a strong immune response to infection limits the population size of the virus in the putative donor host and the presence of circulating monoclonal antibodies to the pathogen means that a higher infectious dose is needed in order to initiate infection in a recipient host.

As experimental control, it must be noted that the transmission room was disinfected thoroughly before each new pair of hosts were introduced. In natural disease transmission, influenza virus shed into the environment may remain infective for hours, days or possibly longer (Bean et al., 1982; Thompson & Bennett, 2017) and can contribute to the establishment of infections in susceptible hosts. Thus, our experiment excluded any opportunity for indirect/mechanical transmission of the virus, which has potential implications on our ability to accurately model epidemic burn-out. We have assumed that droplet infection from an infected host is the only way to seed new infection and we have based our models on this minimum infectious dose. Virus shed into the environment may supplement that secreted directly by an infected host (Greatorex et al., 2011), meaning that low level shedding from the nasal cavity would not necessarily interrupt transmission (Wißmann et al., 2021).

Realistically, despite the best efforts of the horse-owning community, 100% EIV vaccine coverage in the field is unlikely to occur and, as seen in the 2019 European outbreak, broad coverage of the horse population doesn't necessarily lead to epidemic burnout. In fact, this outbreak frequently saw the symptomatic infection of fully vaccinated horses, with these individuals contributing to the production and spread of virus capable of infecting other hosts. Furthermore, we can assume that asymptomatic infection of vaccinated horses also occurred as we saw evidence of positive shedding in each host of the vaccinated class, adding yet another pool of actively infective hosts. To relate back to the notion of SEIR epidemiological models, each of these classes of horses likely have their own transmission (ß), eclipse (σ) and recovery (γ) rates. Thus, epidemic maintenance should be considered a complex system where likelihoods of individuals being infected depend heavily on both donor and recipient host factors together with host-pathogen interactions within hosts on either side of a transmission event. However, one factor unable to be measured in this transmission experiment is the presence of super-spreader individuals. As with examples seen in human disease outbreaks, the causes of such super-spreader phenomena are multi-factorial; ranging from individual behavioural or genetic differences to population movements. To date, no specific examination of super-spreaders in EIV epidemic dynamics has been carried out. Lessons could be learnt however from the testing of super-spreaders in foot and mouth disease virus (FMDV) outbreaks in similar, livestock populations (Hidano & Gates, 2019).

Following the 2018-19 outbreak, the Horserace Betting Levy Board recommend booster vaccinations every 6 months rather than annually (*International Codes of*

Practice, 2023); though this is also to account for antigenic drift of circulating EIV. Further work is needed to examine the actual minimum infectious dose and the transmission dynamics that shape continued host-host infection. This should lead to a better representation of the p_{crit}, i.e. the proportion of susceptibles that would need to be vaccinated in order to prevent onward transmission. As guidelines differ depending on the life of horses, i.e. sporting or non-sporting, a reliable figure for p_{crit} is not known and instead owners are encouraged to vaccinate all eligible horses.

3.3.1 Outcomes

To summarise, hosts that received a vaccine shed less than those that did not, with vaccinated individuals exposed to autologous challenge showing the lowest level of shedding. Manufacturers have a choice whether to produce monovalent or multivalent vaccines. The decision to include multiple antigens is one of breadth of coverage, attempting to provide broad immunity to a handful of circulating IAV strains rather than specifically target a single strain and potentially leave lower protection to non-targeted strains. Unsurprisingly, the best recourse to prevent equine influenza in an individual and to help protect others is to ensure that horses have an up-to-date vaccination record. Despite our findings, using a monovalent vaccine in the real-world would only be recommended if a single, well-characterised EIV strain was circulating; as this is rarely the case in a globally distributed virus, the breadth of protection offered by multivalent vaccines outweighs the slight reduction in performance against a specific strain (Blanco-lobo et al., 2019; Daly et al., 2004).

Our findings confirm and broaden understanding of viral load as a key feature of disease processes; following work from Wood (1993); Whitlock et al. (2018) and Smith (2004), with tightly-controlled experimental conditions. The novelty we provide lies in the differing responses of naïve hosts in each transmission chain; N_M and N_S. Even though these two classes should theoretically behave identically, N_M hosts shed considerably more than either vaccinated group whereas N_S hosts are not significantly different to either vaccinated class. Hence, we conclude that the exposure history of hosts can impact the infection dynamics of EIV in hosts further down a transmission chain and that the responses of immunologically naïve hosts may be affected by the immune status of the donor host that infected them.

4 Analysing Consensus Sequences from Influenza Transmission Experiments

As sequencing technology and software have developed, pathogen sequencing has become a mainstay of disease surveillance, treatment and management. Influenza A virus infections are usually acute meaning that viral population may be present in a host for a short period of time. In this time, viruses diversify due to the introduction of random genomic mutations. Rapid reproduction cycles enable viral populations to respond rapidly to host environments by adapting to selective pressures. However, only a subpopulation of viruses leaves the host to establish new infections meaning that some of the diversity generated during an infection stays confined to that host. To explore how viruses evolve both within hosts and between hosts, two transmission experiments were carried out. Over the course of the transmission experiments, influenza A viruses collected from the nasal swabs of infected horses were sequenced to build consensus genomes. 21 unique point mutations appeared in the 53 samples, distributed evenly across the entire IAV genome. Much of the observed diversification was generated in horses that had previously received influenza A vaccines, viruses from unvaccinated horses mostly remained genetically identical to each other.

4.1 Introduction

In the preceding chapter I endeavoured to relate the level of individual viral shedding to host factors such as vaccination status, day post-infection and transmission group. Analyses showed a larger population of viruses in unvaccinated hosts, indicative of larger viral populations in those hosts without vaccine-mediated immunity. Presently, I examine changes in the Equine Influenza Virus (EIV) genome that appear in individual hosts and throughout the experimental transmission chains. Observing the intra-host diversity of viruses relies on the collection and analysis of viral genomes from individuals sampled at multiple timepoints.

Genomic sequence data may be analysed to investigate and understand viruses. Viruses were the first genomes to ever be sequenced (bacteriophage MS2 (Fiers et al., 1976)) and also the first DNA genome to be sequenced (bacteriophage ΦX174 (Sanger et al., 1977)). Since the advent of next-generation sequencing (NGS) technologies, sequencing of viral genomes has become commonplace in many settings and has been applied to clinical (Houldcroft et al., 2017), diagnostic and surveillance fields to name a few (O'Carroll & Rein, 2016). With better understanding of viral genomes and the proteins they encode, regions associated with specific phenotypic changes (e.g. drug-resistance or emergence in a novel host) can be tracked and observed.

The value of genomic sequencing in viral outbreaks, and the investment of money and labour into fulfilling sequence surveillance, further proves the importance of these data (Gardy & Loman, 2018; Nicholls et al., 2021). During the 2013-16 Ebola virus (EBOV) outbreak in West Africa, health and research projects collaboratively sampled over 1600 EBOV genomes (Dudas et al., 2017), representative of over 5% of recorded cases. This was the first viral outbreak in humans to focus on sequencing of pathogen genomes and it provided an unprecedented insight into viral phylodynamics, i.e. the joint analysis of epidemic and evolutionary dynamics. Since then, technology and processing pipelines have advanced rapidly. By 2020 extensive viral sampling, sequencing and analyses pipelines had been created by academic, clinical and governmental bodies following

SARS-CoV-2 emergence. By April 2021, the Covid-19 Genomics UK Consortium (COG-UK) were sequencing over 10% of all reported Covid-19 cases each week (Marjanovic et al., 2022).

A common aim of sequencing pathogen genomes, especially those of bacteria but nowadays also those of viruses, is to track genes involved in the development of drug-resistance (Schürch & van Schaik, 2017). Anti-viral therapeutics often target specific viral proteins and reduced efficacy, or even complete resistance to these compounds, can result from changes to viral protein structures (Das et al., 2010; Mather et al., 2012). This is also the case with viruses that encounter host adaptive immune cells and molecules; recognition by T- and B-cell receptors (TCR, BCR) can eventually lead to clearance of the virus from the host. Hence, following the trajectory of mutations that arise within pharmacologically-targeted molecules can be an early warning sign of the emergence of drug-resistant strains (Park et al., 2009; Vahey & Fletcher, 2019). Sequencing is also used to assist in the design of vaccines matching circulating Influenza A Virus (IAV) strains, attempting to pre-empt any changes to antigenic proteins that would allow escape from pre-existing host adaptive immunity (Henry et al., 2018; Mumford, 2007; Schotsaert & García-Sastre, 2014).

A further use of viral sequence data is in the reconstruction of transmission trees, based on connecting sequences with epidemiological data to estimate chains of transmission (Campbell et al., 2018; Hall et al., 2015; Ypma et al., 2012). In much a similar method to the coalescent theory used to estimate phylogenies (Kingman, 1982), consensus sequences can be sampled from hosts to reconstruct transmission trees based on the genetic distance between two sampled viral populations (De Maio et al., 2016, 2018).

For many decades, emergence of zoonotic viruses into human populations has been understood as a potential public health catastrophe (K. E. Jones et al., 2008; Parrish et al., 2008; M. E. J. J. Woolhouse et al., 2005), with SARS-CoV-2 surprising many who had been anticipating the threat of influenza A from birds (Flanagan et al., 2012; Gibb, 2020; Morse et al., 2012). An application of viral genomic research is surveillance and the prediction of 'host jumps', adapt of the virus to a novel host. First, however, viral genetic determinants of host specificity must be detected and annotated in order to identify the gene or genes that permit cross-species transmission. For HIV-1, adaptations to escape restriction factors such as tetherin (Neil et al., 2008), TRIM5α (Stremlau et al., 2004) or SAMHD1 (Hrecka et al., 2011; Laguette et al., 2011) enabled Simian Immunodeficiency Viruses to develop into human-adapted pathogens capable of anthropogenic transmission. Likewise, the emergence of SARS-CoV-2 appears to be mediated by mutations and insertions in the furin-recognition motif of the Spike protein which binds to the host cell ACE2 viral-receptor (Andersen et al., 2020; Becker et al., 2020; Zhang et al., 2020; Zhang & Holmes, 2020). Thus, knowledge of the viral proteins involved in host-determination and adaptation can guide surveillance and prediction of putative cross-species jumps.

In many viruses, proteins involved in binding and entering host cells are often the first determinant of host permissibility (F. Chen & Cui, 2017; Mackenzie et al., 2007) and for influenza A viruses, host-range is commonly attributed to the surface glycoprotein haemagglutinin. This is not, however, the only element defining

susceptible hosts. Experimental mutagenesis tests have found key roles for mammalian pathogenicity-related mutations (MPMs) in the polymerase (especially PB2) of avian IAV (C.-Y. Lee et al., 2020; J. Li et al., 2009; W. Li et al., 2017; Min et al., 2013). Knowing which parts of the replicative machinery limit the host range of influenza viruses allows guided tracking of mutations with potential phenotypic associations.

IAV continues to be the focus of governmental pandemic preparedness plans and spillover of H5N1, H7N7 and H7N9 viruses from avian hosts are highlighted as one of the largest threats to public health in the UK. Even policy written a year into the Covid-19 pandemic maintained a focus on influenza A viruses, directly stating "pandemic influenza is one of the most severe natural challenges likely to affect the UK" (Health and Social Care, 2020). Globally, the WHO has allocated almost US$240 million to its Pandemic Influenza Preparedness framework (Pietrasik, 2023), while their 148 National Influenza Centres (NIC) keep a vigilant watch over IAV ecology and evolution (WHO, 2023).

Finally, closest to this study, work by Murcia *et al.* explored mutations arising in natural transmission of EIV through naïve (2010) and vaccinated (2013) horses. However, these studies both relied exclusively on sequencing the short HA1 gene on the fourth genomic segment (~980bp) which although a highly variable region, it is not wholly representative of the full 13kb EIV genome. Further, a key difference between the previous two studies and the experiment presented here is that in these prior studies both chains were homogeneous in terms of host type, i.e. transmission either occurred between vaccinated or unvaccinated horses. By comparison, the experiment discussed here featured natural transmission through vaccinated hosts and subsequently through naïve hosts. This was done in the hope of more clearly documenting evolutionary changes in the viral population directly associated with host immunity.

With assembled consensus sequences showing the most prevalent viral genomes present in each sample, the nucleotide sequences may also be translated into protein sequences. These data then grant a further dimension of information from which we can infer the putative impact of non-synonymous mutations. Protein structures of influenza A viruses are well-described, especially the surface glycoproteins (Lopes et al., 2017). Thanks to extensive work on understanding influenza biology, many of the protein structures for commonly studied IAV have been resolved and annotated in great detail (Wiley & Skehel, 1987; N. C. Wu & Wilson, 2020).

By collating consensus sequences of viruses sampled from sequentially-infected horses, changes to the viral genome can be observed. While mutations in viral nucleotide sequences are generated randomly, I choose to investigate mutations that were then fixed or removed from populations at non-random rates. Tracking the trajectory of mutations within an infected host and throughout a transmission chain, I sought to clarify where in the genome mutations appeared and why they were enriched or purged from the viral population. Conclusions drawn here are expected to be applicable to influenza A viruses beyond H3N8 EIV and should aid our understanding of evolutionary dynamics of viral pathogens which cause acute, density-dependent infections.

At time of writing, there are 422 complete genomic segment sequences of equine influenza hosted on the NCBI Influenza Virus Resource from a total 136,712 IAV samples. However, only 192 sampled individuals have sequence data available $\left(\frac{193}{422}\right)$ for all eight genomic segments, the vast majority of reported sequences are of segment four, and often exclusively the short HA1-coding region. This short sequence is commonly used for rapid identification of EIV, thus is overrepresented in databases. Sequence data can tell us a great deal about the viral population and outbreak dynamics. From the analyses presented in this chapter, I aim to highlight mutations that appear during natural transmission of EIV, draw inferences about why they are fixed or removed from the population and what impact they may have on viral phenotypes.

4.2 Results

4.2.1 Multiple Mutations Appear in the EIV Genome Over the Course of Infection

To understand the evolution of viruses during transmission chains, I analysed whole-genome sequences of EIV collected from the nose of horses infected in the transmission studies outlined in Chapter 2. 53 Whole-Genome Sequences (WGS) were generated in the course of the experiment using the Illumina platform and these were assembled against the challenge strain. I then identified mutations in reference to the overall consensus of all the 53 individual consensus sequences.

Table 4.1: Mutations detected at the consensus level across all 53 genomes. Rows are coloured depending on whether the mutation is synonymous (green) or nonsynonymous (blue).

Segment	Nucleotide	Residue	Frequency	Global Freq (n=384)
01PB2	g979a	Gly327Arg	1	1
	c1497t	Asp499	1	0
	c1779t	Ser593	1	6
	g2191a	Val731Ile	1	16
02PB1	a881g	Gln294Arg	1	0
	t1500c	Gly500	18	2
	a1853g	Glu618Gly	16	0
03PA	c201t	Asp67	6	57
	c825t	Pro275	1	0
	g1180a	Asp394Asn	1	3
	t1221c	Ile407	1	4
	a1650g	Leu500	1	0
04HA	g431a	Gly144Asp	1	1
	a1401c	Arg467Ser	3	0
05NP	g1445a	Ser482Asn	13	6
06NA	c690t	Thr230	1	3
	a1024g	Lys342Glu	1	0
	t1385c	Ile462Thr	1	2
07MP	a418g	Thr140Ala	1	0
08NS	t84c	Gly28	1	0
	t87c	Asp29	1	0

As observed previously, vaccinated hosts were capable of getting infected and shedding enough virus to infect subsequent hosts. Having examined the viral population size throughout the transmission chains, I next sought to understand putative effects of mutations in coding regions of the EIV genome upon viral transmissibility. This chapter focuses on the number, types and effects of mutations arising in the vaccinated or unvaccinated hosts through the EIV transmission chains.

Horses were sampled for eight days beginning from the day of contact with an infected individual. This resulted in a total of 80 sampling events in each transmission group. Many samples, however, had insufficient material for RT-PCR

and, therefore, could not be sequenced. Ultimately, 29 and 24 sequences were collected in the multi-strain and single-strain transmission groups respectively. At least one sequence was obtained from each horse, with the exception of vaccinate 4B in the single-strain group from whom no sequences were recovered. A disproportionate number of sequences came from naïve individuals: 16 of the 29 multi-strain sequences and 13 of the 24 single-strain sequences, were derived from naive hosts, despite only four of the ten hosts in each group being naïve.

The challenge virus with which the transmission experiment began, i.e. the inoculum, was the initial reference genome and is designated as sequence A. Mutations were then defined when the nucleotide of a sample with >100 coverage disagrees with sequence A in the consensus sequence. Among the 53 sequences collected during the transmission experiments, 21 mutations were found at the consensus level (Table 1): twelve in the single group, seven in the multi group and two in both transmission chains.

Hypothesis testing, using the non-parametric Kruskal-Wallis & Wilcoxon Signed Rank tests strongly indicated that genetic material was only recoverable in high enough quantities for viral sequencing on days of high viral load (Kruskal-Wallis $chi^2 = 110.62$, df = 1, p-value < $2.2e^{-16}$). Understandably, a host needs to shed a large quantity of virus in order to provide sufficient material to be sequenced. Most notably, the presence and quantity of mutations appeared to be influenced by the host's immune status. The total number of nucleotide mutations was significantly associated with unvaccinated hosts (Kruskal-Wallis $chi^2 = 21.604$, df = 1, p-value = $3.351e^{-06}$) but the transmission group had no effect on the frequency of mutations (Kruskal-Wallis $chi^2 = 1.5823$, df = 1, p-value = 0.2084). Whether this trend towards greater numbers of mutations is due to the immune status of these hosts or simply because mutations are just statistically more likely to appear as time passes is not clarified by these experiments. Though when contextualised with other transmission studies of EIV, populations with homogeneous immune exposure statuses (wholly naïve as in Murcia (2010) or all with previous exposure histories (Murcia 2013)) do not show a significant difference. However, as shown previously, vaccination status has a large impact on the viral population size and so the viral load confounds the variation in the number of mutations detected.

The viral mutational load indicates that a greater number of mutations may actually drive selective processes down (Zhao et al., 2019). Of the 21 mutations observed at the consensus level, 10 were synonymous. Most mutations appeared only once in the study (16 of the 21 are singletons, appearing only once during the study). Additionally, all of the mutations reported in segments 1 (PB2), 6 (NA), 7 (MP) and 8 (NS) are singletons.

Finally, a collection of all 384 reported whole EIV genomes (in personal communication from Laura Mojsiejczuk) was mined for mutations shared with our transmission experiment. Comparing mutations arising in this transmission experiment to those happening at the epidemiological level, I aimed to find mutations shared in both datasets. Though not by itself indicative of changes to viral fitness, the presence of the same mutations at the same nucleotide sites in experimental and wild-type infections implies a certain amount of plasticity at these sites and/or a potential phenotypic effect. 16 of the 21 consensus mutations found in our transmission experiment were detected at least once in the 384 global H3N8 EIV sequences. The number of sequences sharing mutations with our dataset is shown in the final "Global Frequencies" column of Table 1. Knowing that the same mutations that appeared in our transmission experiment have appeared under natural conditions implies, a propensity for variation at these sites without major deleterious consequences.

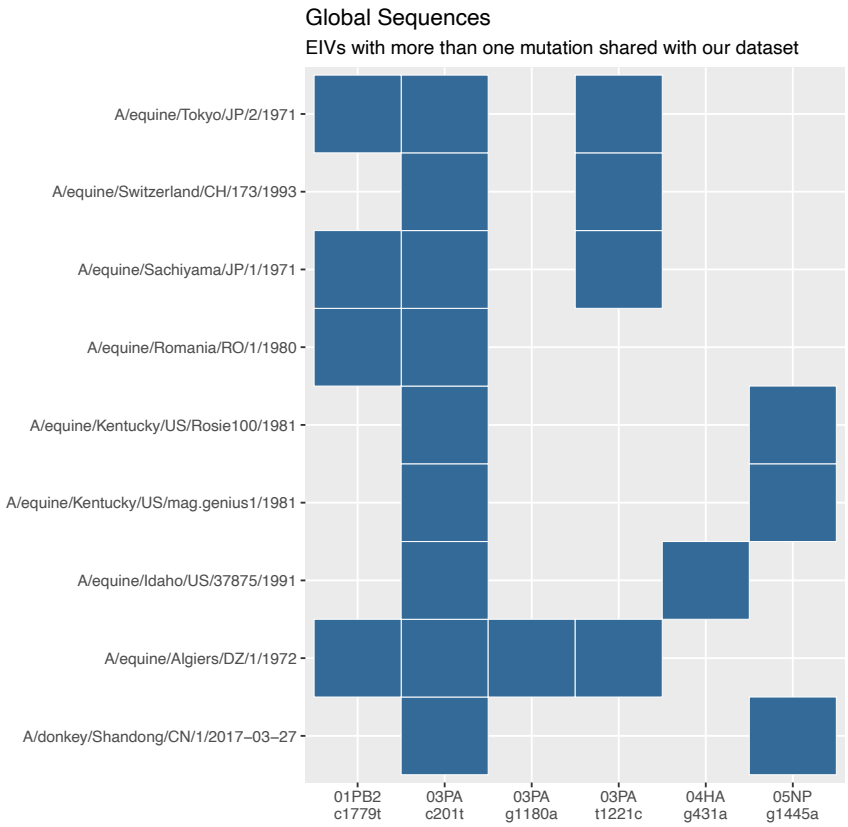


Figure 4.1: Mutations reported in the sequences collected from the transmission experiment which also appear in global EIV sequences. This is then narrowed to the nine global sequences that share two or more mutations observed in the transmission experiment.

To note, of the $\frac{88}{384}$ global EIV sequences that display mutations reported in the experimental transmission chain, $\frac{9}{88}$ sequences have more than one shared mutation with our dataset, as shown in Figure 4.1. EIV sequences have been collected for over 60 years since H3N8 EIV was first detected (1963), hence a lot of variation would be expected in the field. So, finding mutations shared between our dataset and 60 years of global EIV sequences is fairly likely; however, seeing sequences with more than one shared mutation does indicate some level of maintenance in the genome. Furthermore, all of the earlier sequences that shared two or more consensus mutations with our dataset had the synonymous PA-c201t/Asp67 mutation.

Sequences from the four inactivated whole-virus immunogens that comprised each of the vaccines used to inoculate hosts were also compared to approximate the range of adaptive immune specificity raised in response to vaccination. Full genomes of the four inactivated viruses (Miami/63, Newmarket/79, Newmarket/93 and Newmarket/03) shared 93.1% sequence identity with each other. Incorporating the challenge strain, sequence identity remained high between the four vaccine strains and the viral strain that horses were challenged with (lab-grown viruses descended from Newmarket/03) at 92.25% identity. Vaccinated individuals in the Single-strain group had been immunised against the original ancestor of the challenge strain, thus the Newmarket/03 vaccine and challenge strains were very closely related (96.6% identity). Full pairwise comparisons of sequence identity between vaccine strains are displayed in Table 2.

A)

V1	V2	V3	V4	V5
M/63	M/63	NM/79	NM/93	NM/03
NM/03	NM/03	NM/03	NM/03	NM/03

B)	NM/79	NM/93	NM/03	Challenge
Miami/63	96.50%	88.90%	93.60%	90.40%
NM/79		91.20%	96.00%	92.70%
NM/93			92.40%	89.30%
NM/03				96.60%
	μ_1 = 93.10%			μ_2 = 92.25%

Table 4.2: A) Exposure histories of vaccinated horses in the multi and single groups - inactivated virus used in the vaccine regimen are abbreviated and coloured. B) Sequence similarities between the four vaccine strains and the virus that horses were challenged with. Average identity across vaccine immunogens was calculated (μ_1) and then compared to sequence identity of the challenge strain (μ_2). Challenge virus and haplotype A are identical.

4.2.2 Haplotypes

Though many of the mutations we report appear as singletons, some appear together with other mutations and/or in multiple hosts. From these 21 mutations, 13 whole-genome viral haplotypes were identified; these are labelled A-M and are represented graphically in Figure 4.2. Six of these genotypes are more than one step away from the challenge/index virus (**A**), appearing only after other mutations had become fixed prior. There were five fixation events in which a substitution persisted in more than one epidemiologically-connected sample, three of which included a non-synonymous mutation, i.e. PB1-a1853g/Glu618Gly, HA-a1401c/Arg467Ser and NP-g1445a/Ser482Asn. Further, the HA-a1401c/Arg467Ser mutation is a transversion substitution (adenine-cytosine), which is generally considered rarer than transitions.

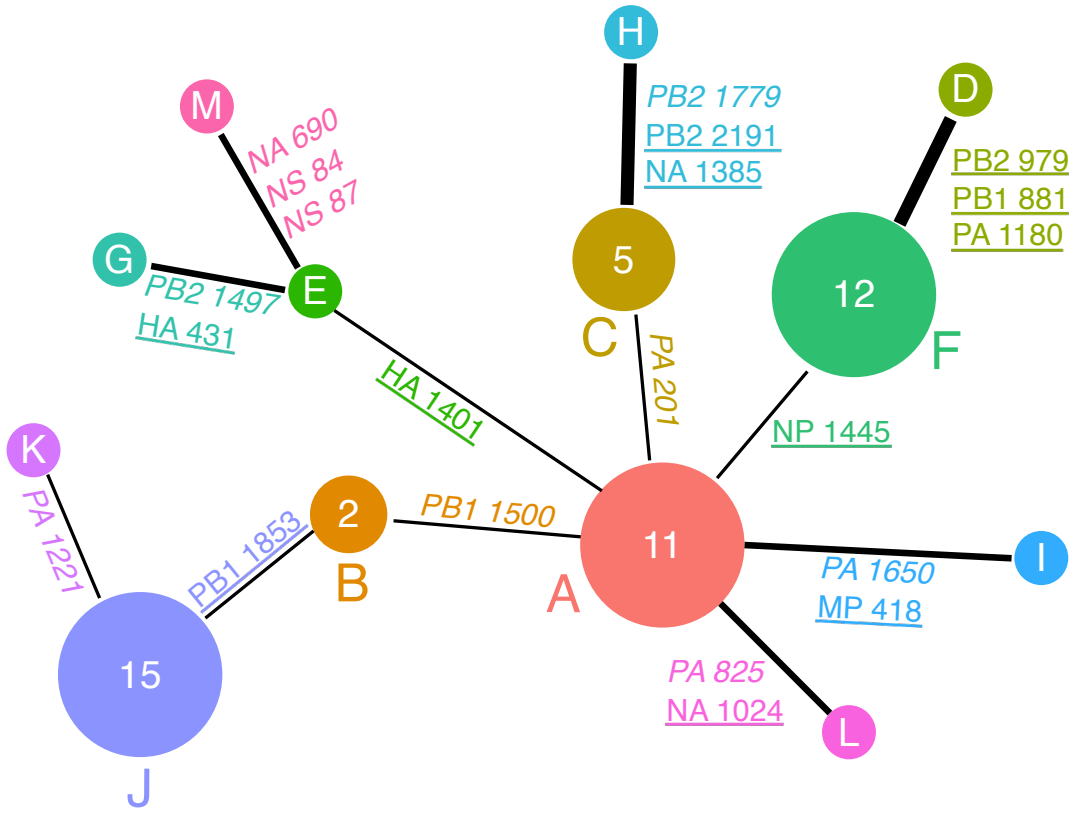
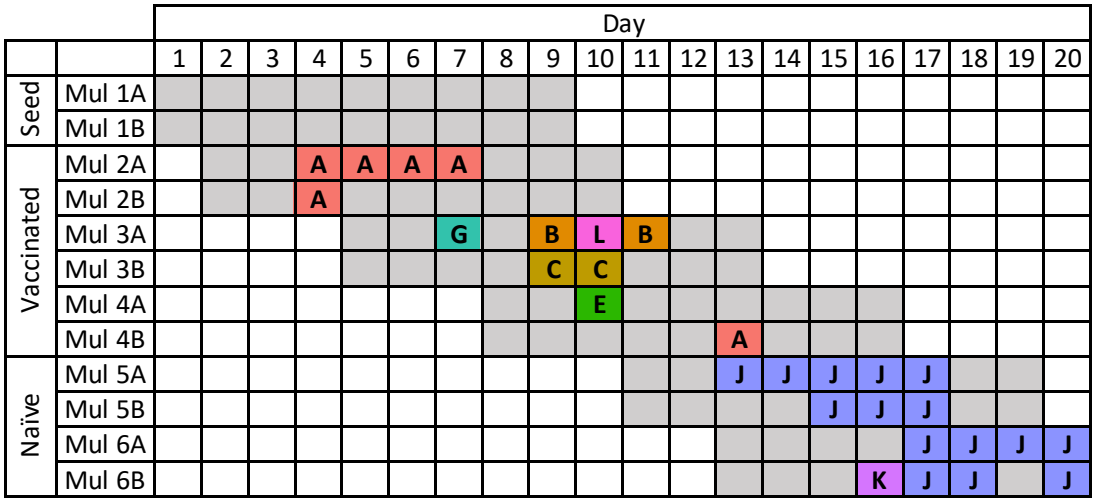


Figure 4.2: Network of all whole-genome haplotypes found among the 53 samples. Mutations are labelled on connecting edges and italicised if synonymous or underlined if non-synonymous. Where a haplotype appears only once, the name of that haplotype appears inside the node; otherwise the name is outside the node and the number inside shows the number of samples sharing that haplotype. Haplotype A is identical to the challenge virus and thus the centre of the network.

Overall, the distribution of genotypes begins very conserved (Figure 4.3) with the majority of genotype A viruses found at the beginning of transmission chains. Since the seeders are experimentally infected with the lab-grown challenge virus (the reference and thus genotype **A**), these individuals mostly remain the **A** genotype and are not shown. These sequences were not included in any of the analyses. Additionally, the few mutations observed in the seeders are transient and thus expected to result from the loss of lab-adaptive mutations upon *in vivo* host infection. Passage through vaccinated hosts generates multiple new genotypes which mostly appear and disappear quickly. Notably though, genotype **C** appears in hosts 3A and 3B from the single group as well as 3B from the multi group. This genotype has a single mutation away from the **A** consensus: PA-c201t\Asp67. The appearance of this mutation in six samples (five sequences of haplotype **C** plus a single **H** sequence which shares this mutation plus three novel mutations) in differing vaccinated horses points to this site deserving a deeper investigation. It is also the mutation that appeared most commonly in the global EIV sequence dataset above.



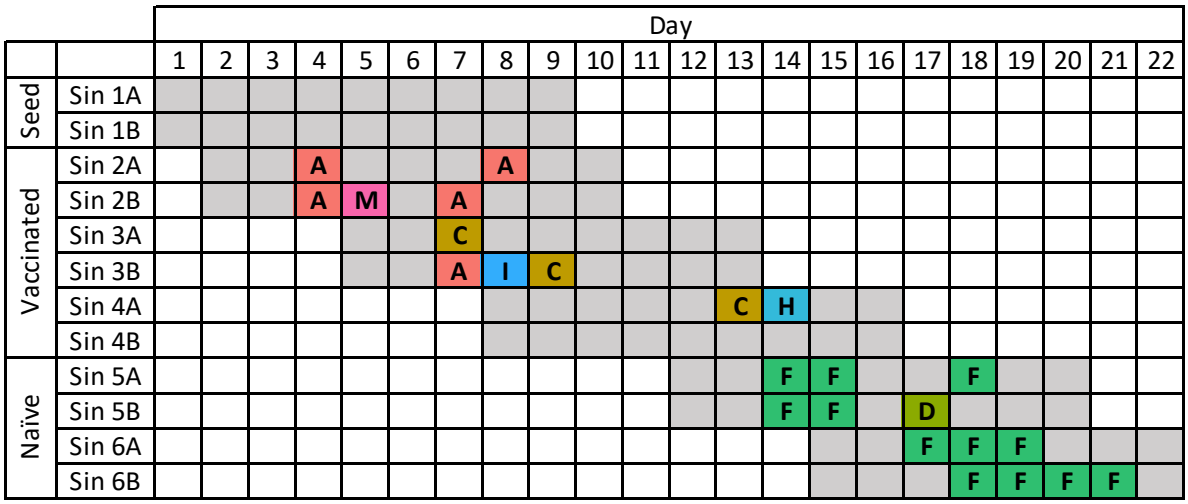


Figure 4.3: Layout of the transmission experiment. Grey boxes show the period of observation and sampling. Days on which a sequence was collected have coloured boxes and are labelled with the corresponding haplotype (A-M).

Tracking consensus heterogeneity helps distinguish how the genome of an infected host can change over the course of infection (as in the change in host Mul_3A from **B** to **L** on the 6th day after contact with an infected individual) and over transmission events. It also shows putative fixation events; almost all the naïve individuals (pairs 5 and 6) of both transmission groups have overwhelmingly the same genotype. Putative effects of these mutations on protein structure and function are explained below. The two exceptions to this are Mul_6B on day 2 (**K**) and Sin_5B on day 5 (**D**). However, even these unique genotypes retain mutations from the prior population.

Initially, I sought to compare the consensus sequence of the transmission experiment with the challenge inoculum and the viral strains used to produce the various vaccines given to horses preceding the experiment. I wanted to establish a solid baseline from which to measure branching diversity, and compare sequences obtained from the experiment with those used to immunise some of the participant hosts. The consensus (**A**) haplotype was identical to the sequence of the challenge virus and shared 96.6% identity with the ancestral wild-type strain of the challenge virus (A/Newmarket/5/03). Hence, we can expect the vaccinates in the single-strain transmission group to have primed immune responses to a virus closely resembling the one they were experimentally infected with. Alternatively, the haplotype **A** virus

shared an average of only 92.25% sequence identity (μ_2 from Table 2) with each of the four vaccine strains used in the multi-strain group. With this in mind, we can expect the adaptive immune memory in single-strain vaccinates to present stronger selective pressures to viruses in these hosts, potentially driving more rapid evolution of these viruses compared to those replicating in vaccinates with broader immune memory.

Counting the number of sequences obtained on each day, and their corresponding genotype, we can see how the proportion of each consensus sequence changes throughout the transmission chain. We see genotypes fixing in the populations over time (Figure 4.3); the heterogeneity generated in the vaccinated hosts is rapidly lost upon transmission to naïve hosts. Some mutations appear in multiple sequences of the same host, different hosts or even different transmission chains. Classing these mutations into haplotypes allows us to see which mutations appear congruently and highlights the homogenisation of genomes upon entering naïve hosts. Though to note, linkage analysis was not a part of this study due to difficulties when working from short-length Illumina reads, hence the presence of physically linked mutations has not been proven and further studies would be required to resolve this issue.

4.2.3 Phylogenetic Analyses

Having detected thirteen distinct haplotypes, the sequence alignments were then assembled into phylogenetic trees, using both maximum-likelihood (ML) and maximum clade credibility (MCC) estimations. Using both methods in conjunction granted a more detailed view of both the relationships between sequences and between each cluster of haplotypes, as well as allowing evolutionary parameters such as substitution rate and branch length to be calculated.

A Maximum Likelihood method aims to find the topology and parameter values of a phylogeny (e.g. branch lengths) that maximise the likelihood of connecting sequence data under a specified evolutionary model. It estimates the probability of observing the data given a particular tree and model of evolution and searches through all possible trees and parameter values to find the combination that maximises this likelihood. However, ML relies on specific assumptions about the evolutionary processes in its estimation. Overall, an ML approach provides the best-fitting tree under the assumed model with estimates of branch lengths and substitution rates.

Conversely, MCC estimation aims to summarise the posterior distribution of trees and parameters to provide a single tree that best represents the evolutionary history. The final MCC tree is a summary tree that incorporates information from all sampled trees, weighted by their posterior probabilities. Like all Bayesian methods, MCC tree estimation depends on prior information about parameters, which can influence the posterior estimates. An MCC tree represents the tree with the highest clade credibility, meaning it is the tree that is most supported by the sampled data, given the chosen model and prior information.

4.2.3.1 Maximum Likelihood Trees

The tips of the ML tree, estimated by IQTree, shown in Figure 4.4 are coloured by genotype and the mutations defining each split are labelled on branches. One genotype (**K**) has two prior fixation events away from the consensus **A**, five others

developed after one fixation event from the root of the tree. Branch lengths of the ML tree average $1.54e^{-5}$ ($\pm$ $2.22e^{-9}$), an unsurprisingly low level of change from sequences collected over the course of 20 days.

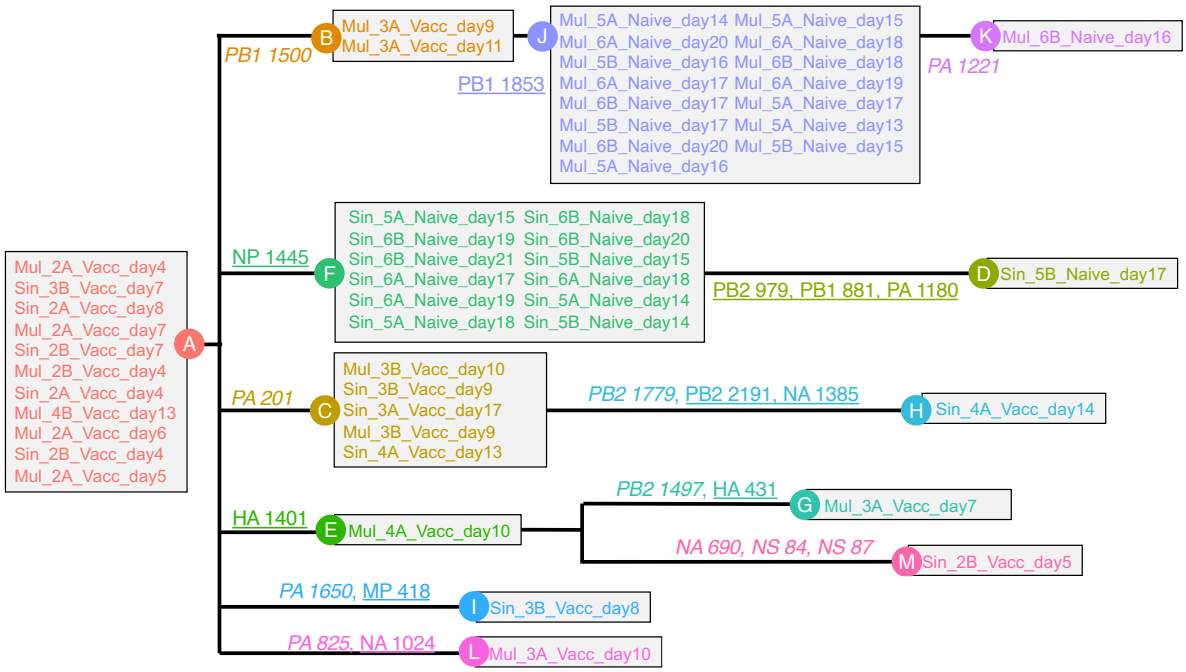


Figure 4.4: A stylised cladogram, based on an ML tree estimated by IQTree, showing each sequence from the experiment grouped into its corresponding haplotype.

Three main clusters of over-represented genotypes can be distinguished (Figure 4.4) which correspond well to the host transmission group and/or vaccination status. 11 samples were genotype **A**, showing no difference from the initial challenge inoculum. The g1445a/Ser482Asn mutation in segment 5 (NP) first appears in hosts 5A and 5B of the 'single' transmission group on day 2, and from this point forwards the mutations became fixed as the **F** genotype. Only one sample after this date diverges from this genotype: while retaining the original NP-g1445a mutation, three novel mutations appear (PB2-g979a/Gly327Arg, PB1-a881g/Gln294Arg and PA-g1180a/Asp394Asn) giving the consensus of 'Single_5B_Naive_day17' the **D** genotype. All four of these mutations are non-synonymous. Similarly, all samples taken from hosts at the end of the alternate 'multi' transmission group had two mutations in segment 2 (PB1); all had the genotype **J** defining a1853g/Glu618Gly mutation. Sample 'Multi_6B_Naive_day16' acquired PA-t1221c/Ile407 mutation in addition to the PB1-a1853g/Glu618Gly mutation (genotype **K**) which was then removed from the population on the following day, reverting to genotype **J**.

Overall, of the 13 genotypes present at the consensus level, most are connected by at least one shared mutation; only two haplotypes (**I** and **L**) appear completely independently, sharing no mutations with any other sequence. These two haplotypes appear on days with samples immediately before and after (**B**-**L**-**B** in host Mul_3A and **A**-**I**-**C** in Sin_3B), representing *de novo* generation and reversion of mutations. The first example, **B**-**L**-**B**, involved two synonymous mutations (PB1-c1500t/Gly500 and PA-c825t/Pro275) and a non-synonymous mutation (NA-a1024g/Lys342Glu) which resulted in the **B**-**L** shift; these were then reversed on returning to the original haplotype (**L**-**B**). Rather than assume that the dominant

virus in this host lost and then recovered the exact same point mutations over three consecutive days, I presume that the appearance of the **L** haplotype was an effect of sampling; perhaps a slightly different sub-population was being shed on this day or this non-dominant variant was amplified more than the **B** virus by chance. Samples from the second host, Sin_3B, are also presumed to be spurious. Moving from the most dominant, and the challenge strain (virus **A**) to a virus with two *de novo* mutations (virus **I**: PA-a1650g/Leu500 and MP-a418g/Thr140Ala) is reasonably likely. However, of all viruses sampled in this pair (Sin_3A and 3B) and the subsequent pair (Sin_4A and 4B) four of the six viruses displayed mutations of the **C** haplotype (three explicitly **C** viruses and one direct descendant of the **C** virus, **H**). We thus have $\frac{1}{6}$ viruses continuing the most dominant virus from the previous strain (**A**), $\frac{4}{6}$ viruses sharing a common ancestor (**C**) and a final sample (virus **I**) sharing no common mutations with the viruses before or after it. Notably, of the 21 consensus mutations, only two (PA-c201t/Asp67 and HA-a1401c/Arg467Ser) appear in both experimental groups, with the remainder being observed only in one transmission chain.

4.2.3.2 Maximum Clade-Credibility Trees

As above, trees were estimated with whole genomes concatenated for each individual sample (Figure 4.5), additionally partitioned depending on the experimental group (multi or single) from which each sample was taken. Metadata (i.e. group and vaccination status) were incorporated in the tree estimation process to delineate clades in order to only group viruses with epidemiological connections, they were not used to partition sequences in any way. Trialling trees without partitioning exposure history (i.e. trees are only partitioned by experimental groups) led to decreased confidence in estimations. The final estimated tree shows very high confidence, with the exception of the top clade which is estimated with 93.91% confidence. This, however, is due to the inclusion of two sequences from a host earlier in the chain (Mul_3A) which are thus allocated because of their genesis of the synonymous PB1-t1500c/Gly500 mutation which is then found in all sequences in the naïve multi group.

Using a monophyletic tree in this way ignores the possibility of reassortment between virions; though with the small sample size and serial sampling over eight days. Bypassing reassortment estimation in this way is justified by the work of Rabadan (2008) and Lauring (2020) who showed that diversifying reassortment is rare, due in part to the similarity between viruses co-infecting a single host cell.

4.2.3.3 Analysis of Evolutionary Rates

Partitioning the sequences by transmission chain allowed for separate evolutionary rates to be estimated for each group. The rates differed between transmission groups; averaged across each genomic segment, we observed in Tracer a mean substitution rate of $8.77e^{-4}$ in the multi group and the single group rate of $1.57e^{-3}$ substitutions/genome/year. These differences were not significant though, with considerable overlap of 95% highest posterior densities (HPD) values. Nor were there appreciable differences between the evolutionary rates of genomic segments.

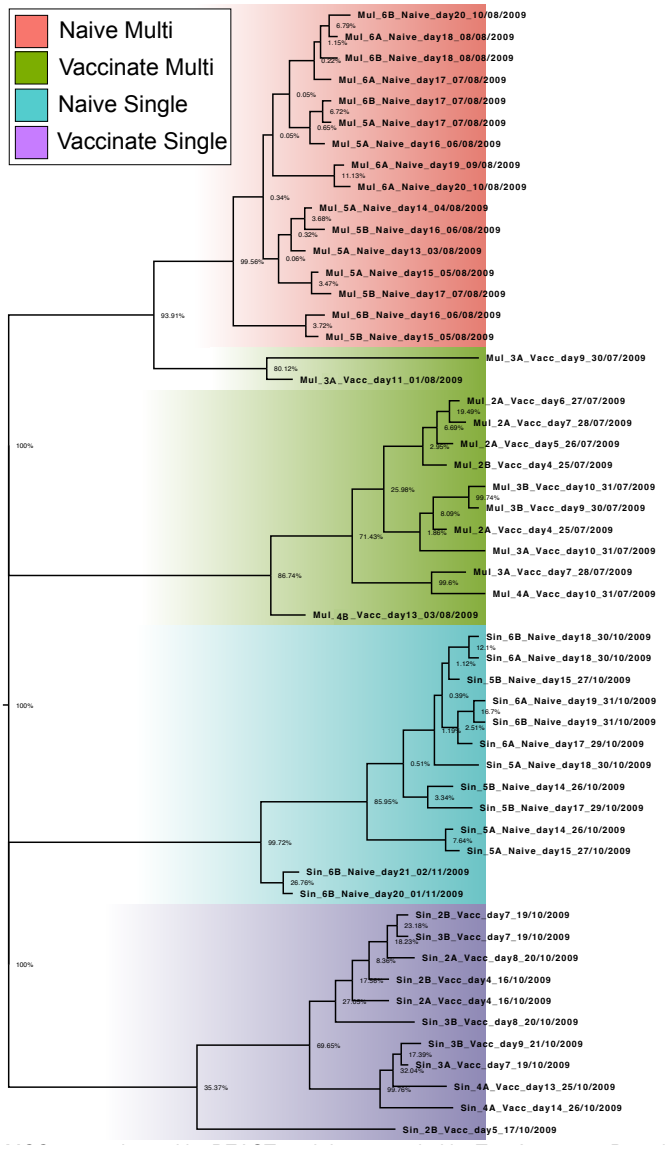


Figure 4.5: MCC tree estimated by BEAST, and downsampled by TreeAnnotator. Branches are coloured according to the transmission chain and vaccine status of the corresponding host. Nodes are annotated with their mean$_{PPD}$ to represent confidence of each predicted split.

4.2.4 Selection Analysis

I next assessed whether selection of mutations may have influenced the evolution of consensus sequences collected throughout the experiment. Each genomic segment across the entire 53 sequence dataset was assessed separately by each of the eight tests offered by the HyPhy bioinformatic suite (Pond et al., 2005). Sequences were not stratified because non-synonymous mutations were generally too rare to appear multiple times across multiple sub-groups. To note, segments 7 and 8 were excluded as they do not have enough diversity to measure any kind of evolution at the consensus level.

Only one segment showed evidence of negative selection by FEL: segment 2 (PB1) at position t1500c/Gly500 (Kruskal-Wallis chi^2=18.18, p-value = 0.0079). SLAC also detected negative selection at PB1-Gly500, as well as positive selection at NP-g1445a/Ser482. The four other tests (MEME, FUBAR, BUSTED and aBS-REL) found no significant evidence of selection or directional evolution. The sparse evidence for selection across viral genomes is perhaps due to the short time-frame of the experiment, limiting the period in which selection could act effectively at the viral population. This was also attributed to 'noisiness' within the relatively small dataset, which further clouded patterns of evolution.

The sites at which evidence of selection did appear (PB1-Gly500 [haplotype **B**] and NP-Ser482 [haplotype **F**]) are also some of the most abundant mutations in the dataset, hence which may confound estimations due to their frequency. PB1-Gly500 appears in only $\frac{3}{384}$ global EIV sequences, while the NP-Ser482 mutation seen in the final virus of the single group **F** isn't observed in any other sequences.

4.2.5 Sequence Diversity

An important population measure is genetic diversity. Patterns of genetic diversity can be informative of a population's evolutionary past; for example, low genetic diversity may be evidence for a recent population bottleneck. Furthermore, the variation of diversity within the genome can be informative of different evolutionary effects, such as the strength of selection in different parts of the genome.

Here, a few commonly used metrics of genetic diversity are selected to show the diversity in each of the four experimental groups (V_M, N_M, V_S and N_S) as well as that of all hosts in each transmission group (Multi or Single) and finally that of the entire experiment together. Analysing the sequences of each group separately, before collating these into an analysis of each transmission chain as a whole and then the entire experiment, we will see how estimates of population diversity depend heavily on the background against which they are being compared. Multiple different diversity metrics were used in the analyses in order to get around the inherent biases present in some algorithms, such as the classic problem of overinflation of non-rare species in Simpson's Index or the grouping-blindness apparent when using Shannon's Entropy.

4.2.5.1 Shannon Diversity

Shannon Entropy is calculated at divergent sites in each genomic segment, then aggregated into transmission groups. Entropy was calculated in three rounds of analysis: first a single dataset of all 53 samples of that segment; next, two datasets split only by multi (29 sequences) or single (24 sequences) transmission chains and finally full stratification into each epidemiological group (16 N_M, 13 N_S, 13 V_M and 11 V_S samples of each genomic segment). This grants a broad view of the patterns of diversity across the experiment, as Shannon's Entropy is calculated based on the differences to the rest of the population under comparison (Table 4.3). An example of this is the lack of diversity in PB1 of the N_M dataset; examining the

sequences, all samples contained both t1500c/Gly500 and a1853g/Glu618Gly mutations but because this is seen in every sequence of the N_M group, the resulting entropy score was zero.

Table 4.3: Shannon's Entropy of each genomic segment, for each group & vaccination status then averaged across those four groups.

Segment	Multi Group			Single Group			Average
	Vacc	Naïve	Total	Vacc	Naïve	Total	
1 PB2	0.091	0	0.050	0.102	0.091	0.058	**0.031**
2 PB1	0.143	0	0.226	0	0.091	0.058	**0.150**
3 PA	0.117	0.078	0.061	0.160	0.091	0.089	**0.049**
4 HA	0.117	0	0.067	0.102	0	0.058	**0.052**
5 NP	0	0	0	0	0	0.230	**0.186**
6 NA	0.091	0	0.050	0.102	0	0.058	**0.031**
7 MP	0	0	0	0.102	0	0.058	**0.031**
8 NS	0	0	0	0.102	0	0.058	**0.031**
Average	**0.070**	**0.010**	**0.057**	**0.084**	**0.034**	**0.083**	

Diversity appeared highest in the vaccinated hosts with values of 0.07 and 0.084 in the multi and single groups respectively. Sequences were much more homogeneous in N_M, while the diversity of N_S sat between the two. We observed a high level of homogeneity across the N_M dataset indicating that sequences remained highly conserved; a single synonymous mutation, PA-t1221c, was the only point of divergence [haplotype **K**]. Though Segment 3 was diverse in all examined groups, at the scale of the entire dataset, Segment 5 showed the most variation. Likewise, N_S sequences were mostly homogeneous, with only one of the 13 genomes being unique, as a result of three non-synonymous mutations [haplotype **D**] compared to the most prevalent haplotype in this group [haplotype **F**]. The average entropy of sequences in both vaccinated groups was considerably higher than in either naïve dataset. Finally, when each transmission group was analysed as a whole, the Multi-strain group was found to be least diverse (H = 5.7% diversity compared to H = 8.4% in the Single-strain group).

4.2.5.2 Tajima's D

Tajima's D test examines whether each sequence is undergoing neutral evolution (null hypothesis) or not. It utilises the total number of polymorphic sites in the sampled genome and the average number of mutations between pairs across the dataset (Figure 4.6).

Upon examination against host factors (vaccination status and transmission group), I saw that the distribution of diversity across genomic segments and host factors (measured by Tajima's D) was non-random (Figure 4.6). Initially, Wilcoxon Rank tests showed a strongly contrasting diversity values between transmission groups (hosts in the single group showed more diversity than those in the multi, Kruskal-Wallis chi^2 = 8.4235, df = 3, p-value = 0.03802), whereas the diversity between vaccinates and naïves was only marginally significant (Kruskal-Wallis chi^2= 7.9079, df = 3, p-value = 0.04795). Vaccinates in the multi group showed lower diversity than N_M (-0.21 lower D value), whereas in the single group, vaccinates have slightly higher diversity than naïves (+0.1065).

The population size, as measured by copy numbers, did not impact values of Tajima's D (p = 0.28). Key to note however, is that diversity differed across genomic segments. To investigate further, host factors and viral genomic segments were used to create an additive linear model highlighting differentiation of Tajima's D diversity across the dataset. Diversity in segments encoding the polymerase complex (1-3) is

generally estimated as much lower than that of the shorter genomic segments. To note however, the difference between the diversity of segments 1 and 2 is not significant.

4.2.5.3 Nucleotide Diversity (π)

Nucleotide diversity is a distance measure, indicating the number of sites that differs between sequences averaged across a dataset. Each base position in the dataset is then compared to ultimately calculate the net number of nucleotide differences between populations; practically, π diversity gives the average number of differences between two randomly selected sequences.

Across the EIV genome, π diversity is higher in hosts of the Single transmission group (multi = 1.61, single = 1.81). However, breaking down these values to their component epidemiological groups shows that both vaccinated and naïve hosts in the multi group had very similar levels of π diversity (V_M = 1.46, N_M = 1.61) whereas both the vaccinates and naïves in the single-strain transmission group differed greatly: V_S = 2.15 and N_S = 0.46.

4.2.5.4 Consensus Genome Diversity

These three measures are presented together in Figure 4.6. Nucleotide diversity (π) and Shannon Entropy both measure diversity by the differences in nucleotides; the first averages the number of different nucleotides between any two given sequences while the latter uses the mean entropy of nucleotides across the sites in any given sequence. Thus, the greater the value of π and Shannon Entropy, the more differences are present between two sequences. Tajima's D, however, is a neutrality test, which compares the diversity seen within a sample to that which would be expected during neutral evolution (Korneliussen et al. 2013). Ultimately, a D score of zero indicates that the variation seen in a population matches what would be expected. When D is greater than 0, genomes contain lower levels of mutations than would be expected, implying that the population is undergoing balancing selection rather than purifying (negative) selection.

On the basis of these tests diversity was found to be highest in viruses isolated from vaccinated hosts in the Single transmission group. Further, the neutrality test (Tajima's D) showed that this population of viruses had evolved in a non-neutral manner. This implies that strong selective pressures were placed upon these viruses by the host's strain-specific vaccine-mediated immune responses which led to greater genomic diversification than in viruses collected from other hosts (N_S, V_M and N_M).

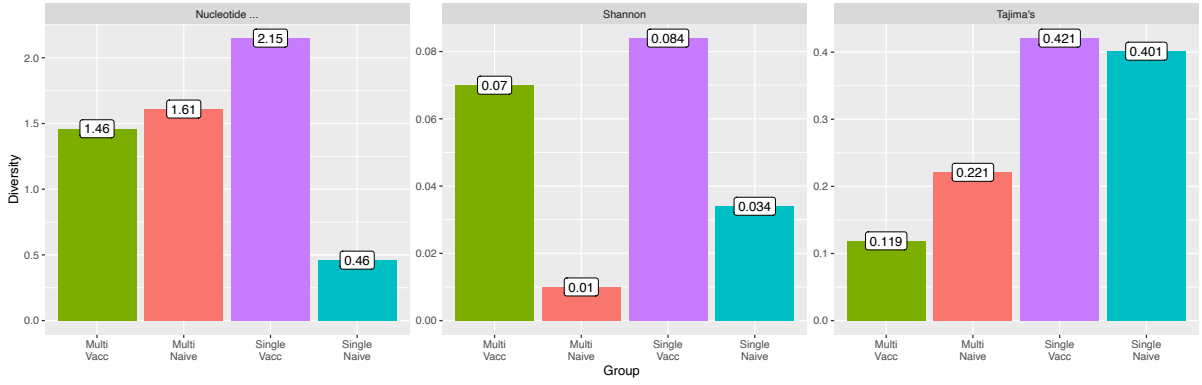


Figure 4.6: Three measures of genetic diversity, applied to sequences from each of the four tested groups.

4.2.6 Structure and Function of Mutations

So far, we've seen where mutations are located, how frequently they appear (within the experiment and at the epidemiological scale) and whether genomes are more diverse in certain groups. But do any of the mutations have phenotypic effects associated with them? Are different host environments driving the evolution of different phenotypes?

Initially, information on the proteins encoded by each segment was mined from PDB's UniProt after alignment with the nearest homolog in the database. I also estimate the first structural models of the whole H3N8 proteome using AlphaFold. Having reliable 3D models of each protein allows for not only analysis of proteins in their natural tertiary/quaternary structures but also simply provides better visual aids for studying. Knowing how amino acid residue variation at particular sites affects protein structure, permits an insight into the role of mutations. Further, structural models are now used widely in the manufacture of products for disease-control: vaccine immunogens can be designed based on the expected interactions of viral and host proteins while specific anti-viral protein inhibitors, such as oseltamivir, can be developed based on our structural understanding of pathogenic proteins.

With these structural models, I then proceed to investigate, *in silico*, the impact of non-synonymous mutations on protein form. This also enables investigation of similarities between EIV HA and other HA proteins. Finally, I estimate the antigenicity of the two EIV surface proteins, haemagglutinin and neuraminidase, to map epitopes onto protein structures as well as assess the impact of amino acid polymorphisms at antigenically-available sites.

4.2.7 Protein Analysis

The ProtParam tools on the Expasy Proteomics Server (Duvaud et al., 2021; Gasteiger et al., 2005) allow for estimation of a range of protein chemical properties such as weight, charge and hydrophobicity (Table 4A). These physiochemical properties can then be used to further predict structural properties. Implementation of the flexibility method of Vihinen et al. (1994) uses a sliding window to approximate the likely structures created by short stretches of residues. On comparison with prototypical human IAV protein sequences (Igarashi 2010), the properties estimated here for H3N8 EIV appeared remarkably similar (Table 4B). Haemagglutinins from pandemic, seasonal and emergent (A/California/04/2009 [CA2009], A/Brisbane/59/2007 [BR2007] and A/South Carolina/1/1918 [SC1918] respectively) H1N1 strains had similar molecular weight (63.24kDa, ±0.26), hydrophobicity (gravy = -0.34, ±0.01) and aromaticity (0.1, ±0) to those estimated from our H3N8 EIV sequences (genotype A). As a reference, the difference in weights between pre-pandemic (BR2007) and contemporary (CA2009) human haemagglutinin to original 1918 H1N1 haemagglutinin is roughly the same as the differences between these human and equine haemagglutinins (~0.37kDa difference both between contemporary vs 1918 HA and contemporary vs equine HA).

While not a perfect comparison, this quick sanity-check shows that the predicted properties of EIV proteins mirror those of experimentally-determined human IAV haemagglutinins. This demonstrates the general homogeneity of

influenza A protein properties, and the applicability of using human IAV structures for homology modelling of EIV proteins.

A)	Segment	Protein	Weight (kDa)	gravy*	Aromaticity	Instability	Isoelectric Point
EIV H3N8 (haplotype A)	01PB2	PB2	86.02	-0.31	0.0660	47.74	9.50
	02PB1	PB1	86.53	-0.51	0.0890	38.95	9.38
	03PA	PA	82.75	-0.49	0.0980	50.09	5.47
	04HA	HA	63.61	-0.34	0.0970	32.78	8.19
	05NP	NP	56.16	-0.56	0.0720	40.84	9.33
	06NA	NA	52.11	-0.25	0.0960	36.90	8.48
	07MP	M1	27.86	-0.23	0.0520	38.16	9.30
		M2	11.22	-0.26	0.0930	58.53	4.99
	08NS	NS1	24.86	-0.32	0.0680	49.25	6.45
		NEP	14.42	-0.46	0.0740	65.67	5.23

B)	Segment	Protein	Weight (kDa)	gravy*	Aromaticity	Instability	Isoelectric Point
SC1918 (H1N1)	04HA	HA	62.87	-0.33	0.0989	37.42	6.05
BR2007 (H1N1)	04HA	HA	63.20	-0.35	0.0991	34.41	6.74
CA2009 (H1N1)	04HA	HA	63.28	-0.36	0.0989	32.29	6.93
EIV H3N8	04HA	HA	63.61	-0.34	0.097	32.78	8.19
Average			63.24	-0.34	0.10	34.22	6.98
StDev			± 0.26	± 0.01	± 0	± 2.01	± 0.77

Table 4.4: A) Properties of EIV H3N8 proteins, as predicted by ProtParam. B) repeats this analysis with the haemagglutinin of three human influenza viruses. *Grand Average of hYdrophobicity

4.2.7.1 Protein Localisation

Knowing whether a protein is situated within, outside or traversing the viral membrane can indicate function and provide insight into whether it may be a potential target for disease control methods. For example, most of the neuraminidase protein is presented on the surface of the virion, which facilitates its function of cleaving sialic acid during viral exit from a host cell. This activity also marks neuraminidase as a potential target for pathogen control and, indeed,

oseltamivir is an inhibitor of neuraminidase that is able to work in tissues because of the exposed nature of its target.

There is a large bias in research toward contemporary human-infecting IAV such as H1N1/2009$_{pdm}$, H3N2 or high-pathogenicity avian influenza (HPAI) H5N1. So, while the location of proteins throughout the virion is known, knowledge of specific features of the EIV proteome are often reliant on extrapolations from viruses naturalised to other hosts. Conversely, this means that EIV can be used as a model for general IAV biology; knowing how the properties of H3N8 EIV proteins compare to influenza viruses naturalised to other hosts means that findings here can be assumed to mirror those in other IAV.

The proteins of influenza A viruses are well-documented and the predicted localisation of EIV proteins generally matches what is known for other IAV. However, knowing the precise location of residues of EIV proteins helps identify which mutations appear intra- or extra-virion and can provide an estimate of function. It also bridges gaps in knowledge, so we don't have to rely on assumptions from other IAVs. Annotation may be transferred from well-characterised proteins in order to highlight active sites, RNA binding sites and monomer polymerisation domains. This assists in gauging the functional impact of nucleotide substitutions in viral coding sequence. Proteins used for mapping and annotation are shown in supplementary Table S4.1.

4.2.7.2 Surface Accessibility

I next estimated the degree to which protein residues are exposed on the viral surface (Figure 4.7) using the Emini Surface Predictability algorithm. Unsurprisingly, many of the antigenic sites on the haemagglutinin head are expected to be exposed to their surroundings until the virion is endocytosed by a host cell. This is also the formation of haemagglutinin when among host tissues when it is at risk of encountering extracellular immune molecules. During this extracellular period, epitopes on the extra-virion proteins of EIV can be targeted by antibodies and complement. Hence, identifying where mutations are located on the structure of extra virion proteins can lead further studies into whether conformational changes to epitopes by non-synonymous mutations affect the binding capability of immune molecules and thus the immune response of a host. This is, therefore, a key molecule of interest for the present study.

The five labelled antigenic sites are estimated from studies of A/England/878/1969 (H3N2) and A/Hong Kong/1/1968 (H3N2) which circulated in the human population in the latter half of the 20th century (Skehel et al., 1984; Wiley et al., 1981; Wilson et al., 1981).

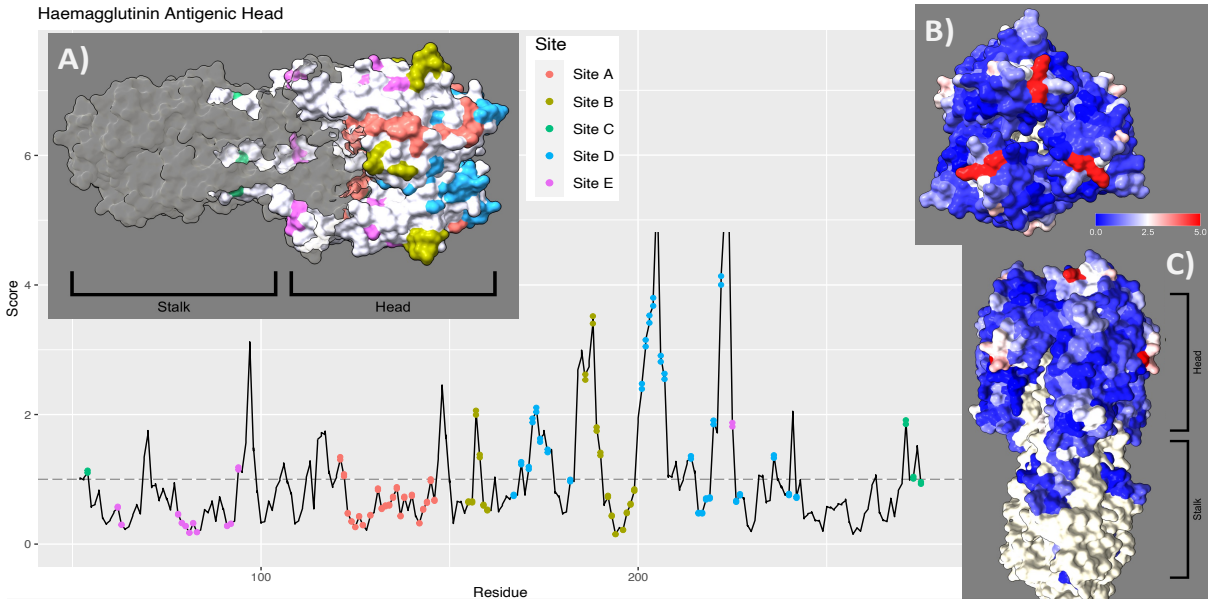


Figure 4.7: Haemagglutinin head protein expected to be accessible outside the virion, exposed to extracellular environments. Antigenic sites A-E are shown coloured. Inset 7A) shows the 3D structure of the haemagglutinin trimer with antigenic sites coloured corresponding to the points on the line graph. Insets B) and C) show the estimated surface accessibility of the protein scaled from low (blue) to high (red).

Homologous residues at these sites were annotated correspondingly in A/Newmarket/5/03 (H3N8) for alignment with the 53 samples obtained during the transmission experiment. The sequences A/England/1969, A/Hong Kong/1968 and A/equine/Newmarket/5/03 viruses shared 82.4% sequence identity in segment 4 and 86.2% across the whole genome. During the transmission experiment, two non-synonymous mutations appeared in the haemagglutinin protein. Site 144 sits directly within antigenic site A (residues 142-146), a region of the protein expected to be targeted directly by antiviral antibodies. Thus, the substitution of Glycine at this site with a larger, more acidic and less hydrophobic Asparagine residue may have phenotypic and/or immune-evasion effects on the haemagglutinin protein. Potential physio-chemical, immune and spatial impacts of residue substitutions are explored further below. Figure 4.8 shows the placement of mutations on the homotrimeric haemagglutinin structure.

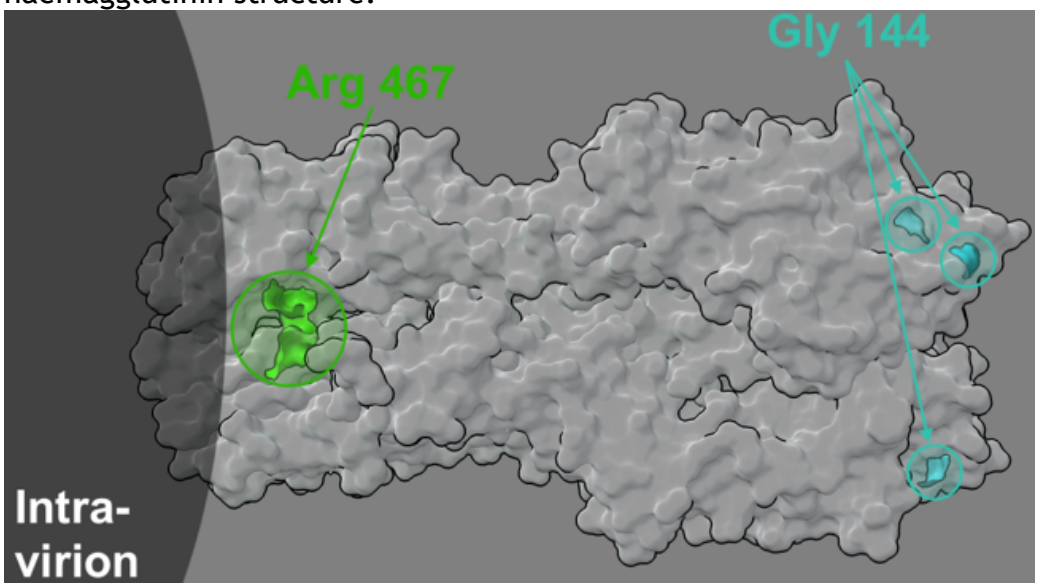


Figure 4.8: Both sites of non-synonymous mutations seen in haemagglutinin.

4.2.7.3 Kolaskar & Tongaonkar Antigenicity

Can we measure the antigenicity of proteins from their sequences alone? Knowing the baseline antigenicity of haplotype A viruses, we can measure how non-synonymous mutations away from this consensus may affect protein antigenicity; further, whether antigenicity increases or decreases in response to vaccine-mediated immunity. Antigenicity was calculated using the method of Kolaskar and Tongaonkar (1990), implemented by a tool hosted on http://tools.iedb.org/bcell/.

Only four total non-synonymous mutations are observed throughout the transmission experiments between both antigenically available proteins: HA and NA. By predicting the antigenicity of the consensus form of the protein (belonging to haplotype **A**) and comparing the scores with the predicted antigenicity of the mutant proteins, non-synonymous mutations with the potential to alter antigenicity (via conformational changes resulting from differing properties of residue side chains) can be detected. It should be noted, the scores are relative and there is no binary threshold defining whether a region is or isn't antigenic.

The two observed non-synonymous mutations in haemagglutinin (Figure 4.9) show different estimated phenotypes: g431a/Gly144Asp does not affect the predicted antigenicity of the protein while the a1401c/Arg467Ser mutant increases antigenicity. This Arg467Ser mutation is seen in three samples, two vaccinates from the multi group and one vaccinate from the single group (viruses **E**, **G** and **M**). Gly144Asp appears only once throughout the dataset, as well as once among the global EIV sequences (A/equine/Idaho/US/37875/1991, NCBI:txid415988) over a decade before the challenge strain for our experiment was isolated. It must also be noted that viruses with the mutation Arg467Ser often had some of the highest viral loads, though due to the rarity of these samples, and a lack of global EIV sequences sharing this mutation, it is impossible to find any statistical significance regarding this association.

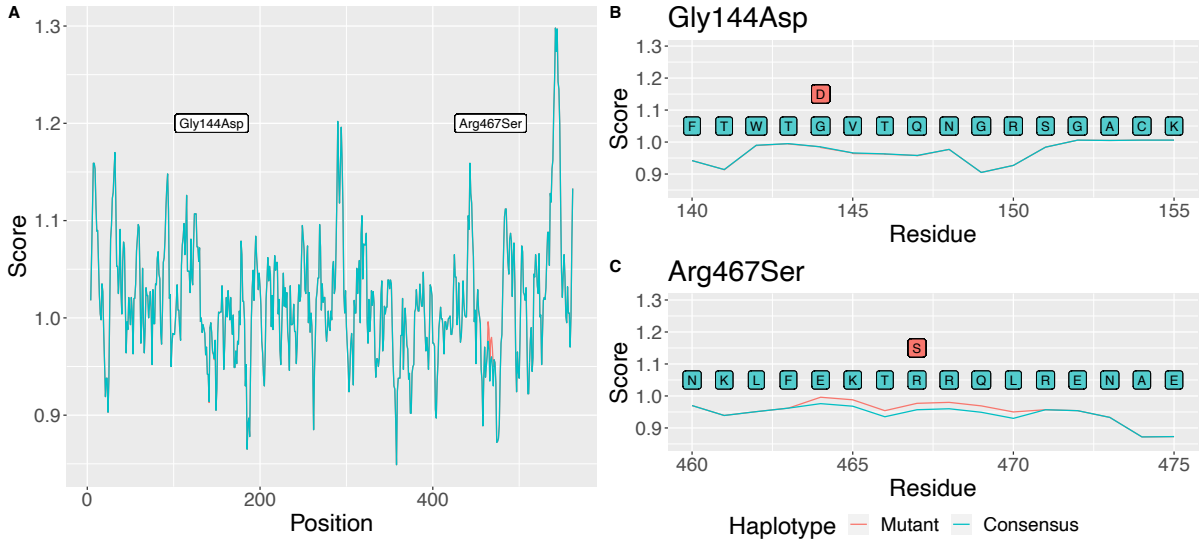


Figure 4.9: A) Predicted antigenicity of all residues in H3N8 EIV haemagglutinin, with a focus on the two mutations detected in our transmission experiment; B) Gly144Asp and c) Arg467Ser.

Conversely, both of the non-synonymous mutations detected in neuraminidase are estimated to decrease antigenicity of the protein (Figure 4.10). Again, however, these mutations only appear once each within our dataset; t1385c/Ile462Thr appears twice in publicly available EIV sequences collected from

China (A/equine/Heilongjiang/CN/1/2010-04-23, NCBI:txid1125808, next in 2013 at A/equine/Xuzhou/CN/1/2013-08-27, NCBI:txid1417962) whereas a1024g/Lys342Glu is not reported in other EIV sequences.

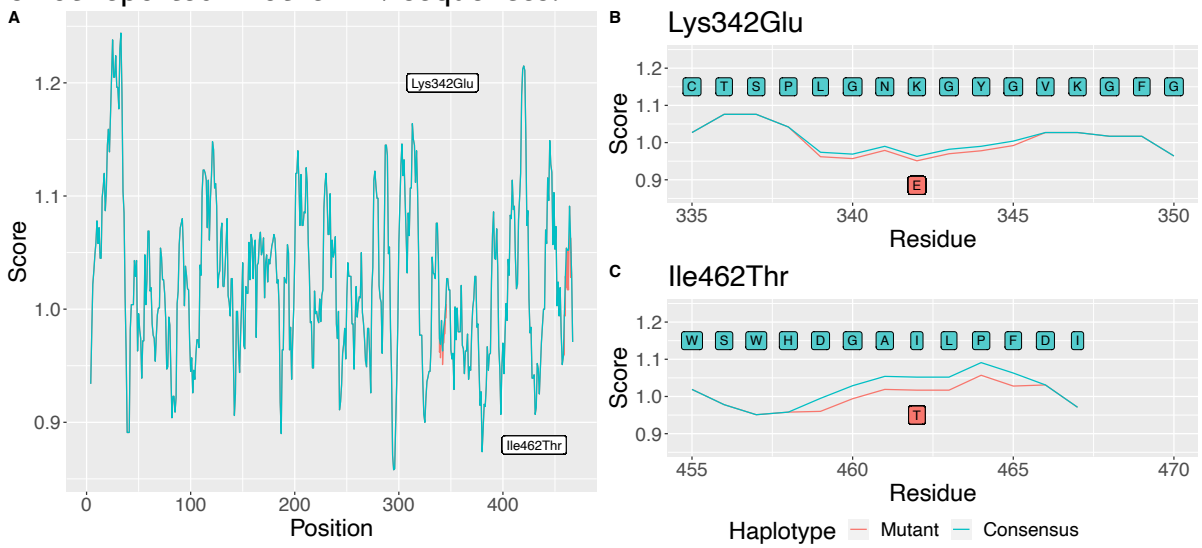


Figure 4.10: A) Predicted antigenicity of all residues in H3N8 EIV neuraminidase, with a focus on the two mutations detected in our transmission experiment, B) Lys342Glu and C) Ile462Thr.

This time, sequences sharing this mutation are dated some years after the original isolation of the challenge strain, A/equine/Newmarket/UK/5/2003, used in this experiment. Unfortunately, due to the sparsity of samples we cannot verify whether either of these mutations impacted the viral loads of hosts.

4.2.7.4 Epitopes

I then examined the propensity of amino acid substitutions to affect potential epitope sites via putative protein conformational changes, the results of which are presented in Figure 4.11. At the consensus level, we detect two non-synonymous mutations in haemagglutinin (Gly144Asp & Arg467Ser) and two in neuraminidase (Lys342Glu & Ile462Thr).

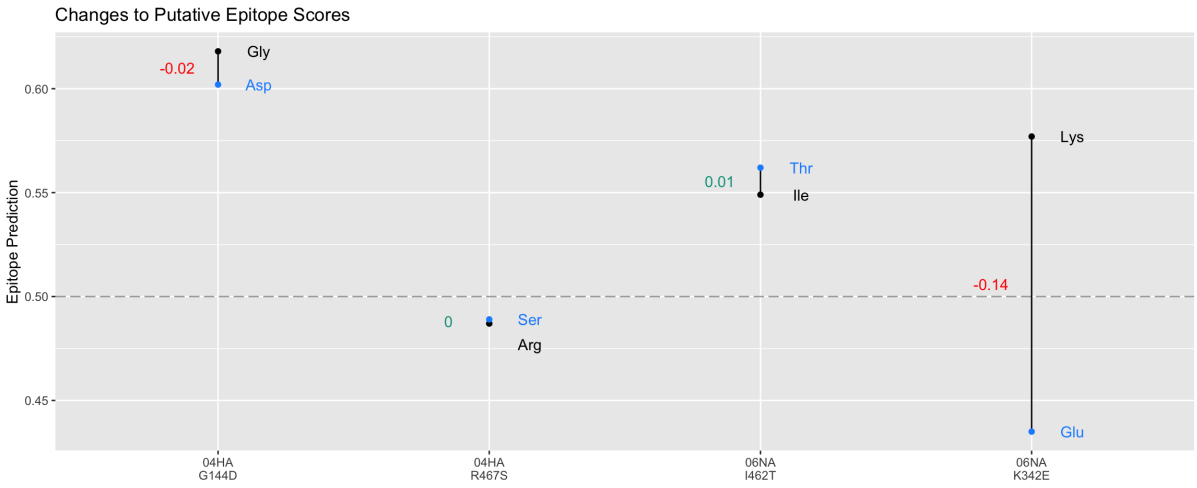


Figure 4.11: Epitope scores, as predicted by BepiPred, of both the consensus (black) and mutant (blue) residues observed throughout the transmission experiment. The changes in values is given by each range and is coloured red for a decreased likelihood of epitope availability or green for increased chances of epitope availability. A value above 0.5 is likely to display some epitope activity, below 0.5 and the location is unlikely to be epitopic; the dashed line demarcates this boundary.

Neither HA mutation substantially changes the estimated likelihood of epitope presentation, according to BepiPred's Linear Epitope Prediction tool (Jespersen et al., 2017). The Lys342Glu NA mutation, however, is estimated to drop the probability of epitopic availability from 'strong likelihood' (0.577) to 'unlikely' (0.435). This matches the above Kolaskar & Tongaonkar test, where this mutation is expected to lower the antigenicity of neuraminidase compared to the consensus seen in other EIV samples.

The estimated changes to viral epitopes are seen clearly in the Lys342Glu mutation of neuraminidase. Strongly basic lysine is substituted by an acidic glutamic acid residue. This change in protein charge is expected to alter protein structure enough to mask a previous epitope site. This mutation, however, only appears once throughout the transmission experiment, in genotype **L** which is quickly lost from the viral population.

4.2.8 Structural Modelling

As the HA trimer is the only resolved structure for equine IAV, modelling the nine other major proteins of EIV allowed us to observe differences between IAVs of other species and see any large changes caused by the mutations detected throughout our experimental transmission chain. I aimed to create usable models for EIV structural biology, vaccine design and analysis of host-pathogen interactions. A substantial amount of data exists on IAV biology and protein annotations, however the vast majority of this is focused on human influenza viruses. I, thus, utilised proteins of non-equine IAV as a template from which to estimate structures and functions of EIV H3N8 proteins. Even if protein structural models are not perfect, they may still have value in estimating the impact of changes caused by non-synonymous mutations.

4.2.8.1 Validating Predictions

Using the results of a Local Distance Difference Test (LDDT) to assess model confidence, the estimated structure of each protein was mostly well-predicted. Averaging the confidence of each predicted amino acid location as a proxy for total model confidence, seven of the proteins were modelled well, with >75% confidence (Figure 4.12). As a reference point, the authors of the tool classify this LDDT into very high model confidence (>90%), confident (70-90%) and low confidence (50-70%) (Varadi et al., 2022).

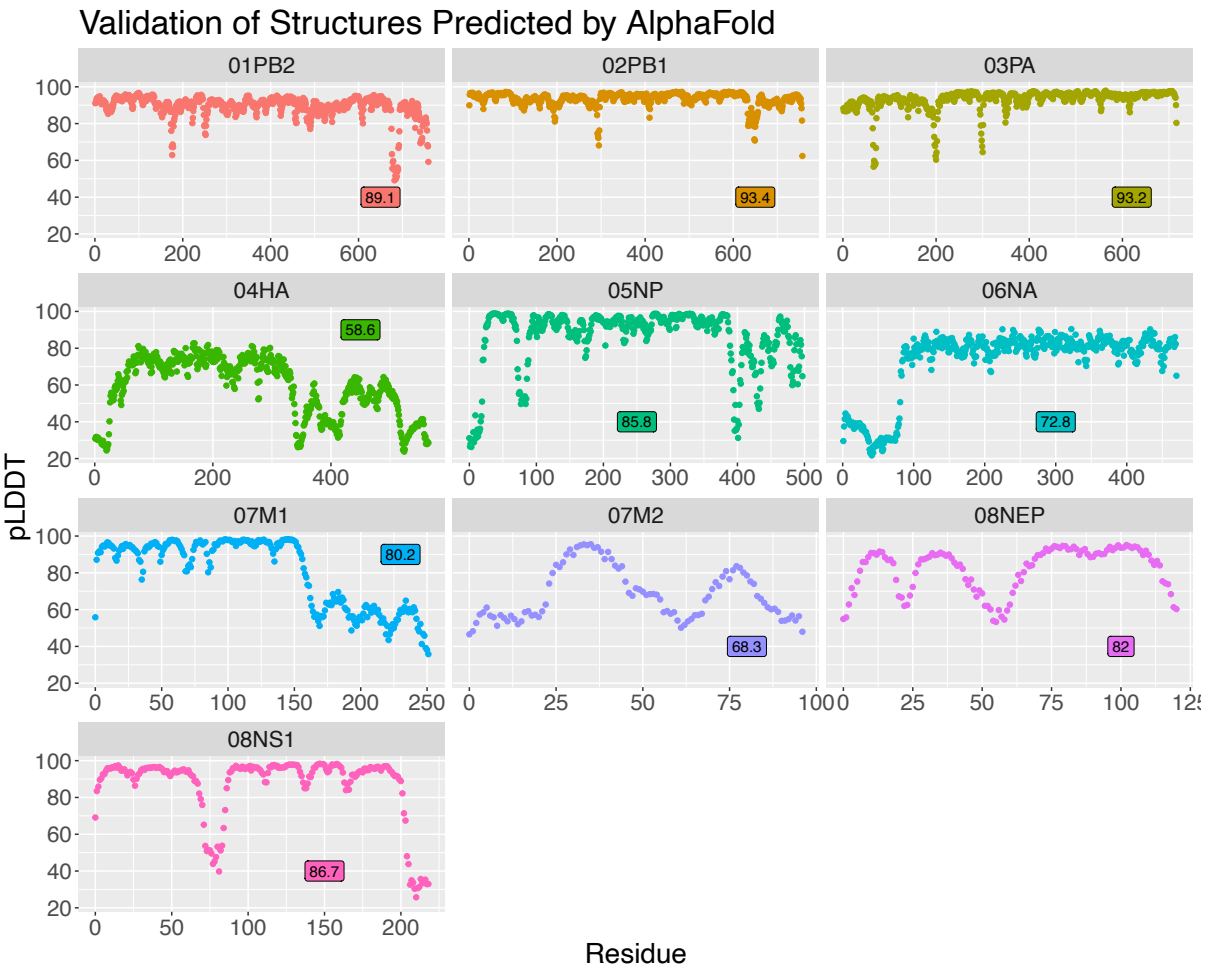


Figure 4.12: The local distance difference of each residue for each modelled EIV protein, calculated with AlphaPickle (Arnold 2021). LDDT values averaged over the whole protein are labelled on each plot. Most proteins are modelled with high (+80%) confidence. Notably, the three transmembrane proteins are estimated with lower confidence.

Haemagglutinin, neuraminidase and matrix 2 were predicted with lower confidence (58.6%, 72.8% and 68.3% respectively), although these are still considered moderately reliable. Notably, these proteins are homotrimers (HA) or homotetramers (NA and M2), which are known to be more challenging for AlphaFold to predict. They are also transmembrane proteins, which are expected to be more difficult to model. Overall, predictions of the influenza proteome generally provided high-confidence 3D structural models, thus enabling *in silico* experiments which could be used to investigate the impacts of amino acid mutation on protein morphology and function.

4.2.8.2 Comparisons with other Models

In silico experiments can often occur in a vacuum and so before commencing testing on estimated protein structures, I sought to compare the models with resolved IAV protein structures. If structural models are close to published resolved crystal structures then we can expect the results of *in silico* experiments to closely resemble what happens in actual proteins. The homology modelling approach may be used in the absence of known EIV protein structures and can provide useful inferences on how non-synonymous mutations may impact proteins. As reference, the matchmaking procedure was initially tested with two closely related

H1N1/2009$_{pdm}$ haemagglutinins, A/Darwin/2001/2009 (PDB:3M6S) and A/California/04/2009 (PDB:3LZG) which gave the following results: sequence alignment score = 1600.7; the RMSD (root mean^2 distance) between 318 pruned atom pairs is 0.571Å and across all 322 pairs is 0.805Å.

In order to validate the use of the predicted structural models in further analyses, I then examined the similarity of these homologous haemagglutinin molecules with the model I developed. As shown in Figure 4.13, the sequence similarity of proteins was compared to the overall difference in protein structures. This Ångstrom difference was calculated by overlaying the AlphaFold structure onto the sample and measuring the resulting mismatch.

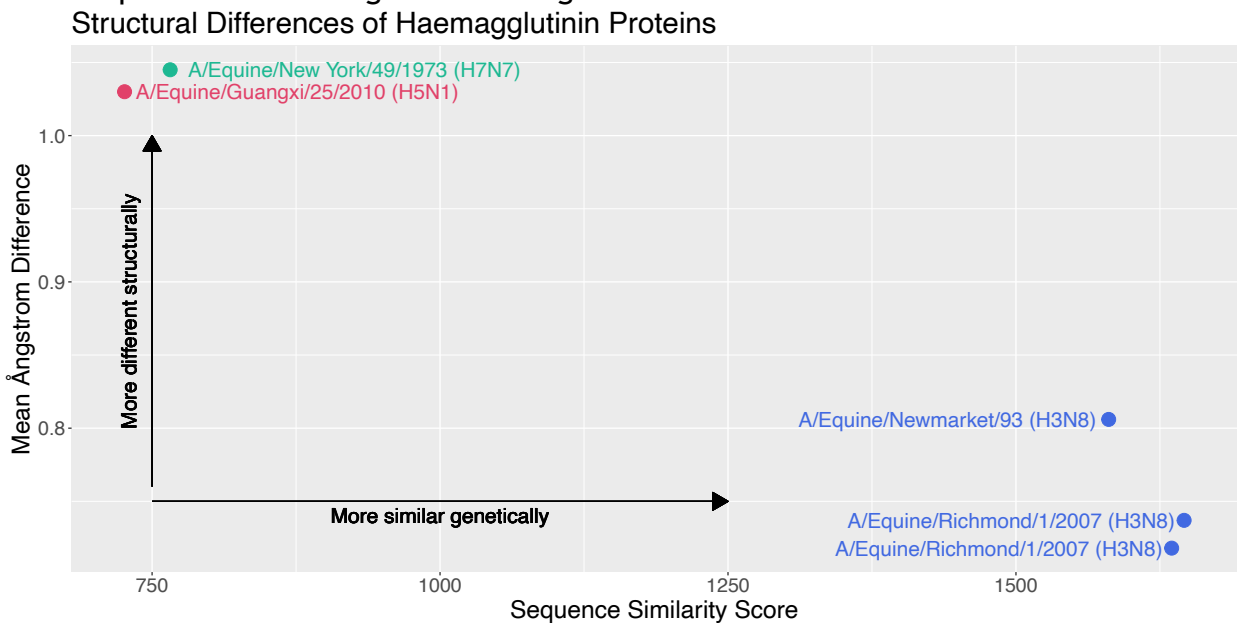


Figure 4.13: Differences between resolved EIV haemagglutinin structures and an AlphaFold prediction. Each published protein is marked with their similarity to (Score) and the distance between (Å), the HA structure as predicted by AlphaFold.

Of the five EIV haemagglutinin structures in PDB, three are H3N8, thus matching the subtype of the inoculum used in our transmission experiment. These three haemagglutinin proteins (from A/Equine/Newmarket/2/93 (PDB:4UNW), A/Equine/Richmond/1/2007 (4UO3) and a mutant form of A/Equine/Richmond/1/2007 (4UO0)) only differ from the AlphaFold model by an average of 0.753Å; considering that the two 2009 H1N1$_{pdm}$ haemagglutinins differed by only 0.571Å, our Å average from viruses sampled in 1993 and 2007 match nicely to the estimated model. The two remaining haemagglutinin structures are from an H7N7 virus, A/Equine/New York/49/1973 (PDB:6N5A) and H5N1 A/Equine/Guangxi/25/2010 (PDB:7WL5). Unsurprisingly, these proteins are very genetically and structurally different from H3N8 viruses.

Figure 4.14 shows how all the resolved IAV structures in PDB compare to the EIV prediction. All have been matched (in ChimeraX) to the AlphaFold prediction of haemagglutinin to quantify how closely related protein sequences are, and how spatial structures of studied proteins compare to those modelled with proteins from our transmission study.

Figure 4.14: All of the available influenza A protein structures were assessed against our structural models. Comparing the alignment score and spatial differences to resolved structures of other IAV proteins, we see high sequence and structural homology with human IAV samples and a distancing from proteins of swine and bat influenza.

From comparing each modelled protein to resolved influenza structures from PDB, I have shown how EIV proteins cluster together with proteins of IAV naturalised to other hosts. The colours also differentiate IAV subtypes. For example, the haemagglutinin model has high sequence identity with other H3N8 equine HA, as well as closely matching the structure of previously resolved equine HA. Conversely, swine H1N2 haemagglutinin proteins have low sequence homology and a substantial difference in structural similarity. To note, however, human-sampled IAV (+) are over-represented in PDB and thus can occlude similarities or differences between viruses of equines and other hosts.

4.2.9 Structural Analysis

Physio-chemical properties of amino acid side-chains play a large role in determining the location of residues among the protein's tertiary/quaternary structure. Amino acid side-chains interact with each other and thus orient the surrounding environment. By *in silico* testing the spatial impacts of non-synonymous mutations, we explore potential changes to protein structures over the course of transmission.

Amino acid substitutions can result in conformational changes to proteins, with differing physiochemical properties of residue side-chains altering the interactions between residues. In an effort to measure these changes, *in silico* experimental substitutions were carried out on the structural models created above. These substitutions then had associated changes to the bonds linking the α-Carbon to both amino (N-C_α) and carboxyl groups (C_α-C); as amino acids display chirality, the angles of these rotations in the amino-C_α and C_α-carboxyl bonds can be measured as Φ and Ψ angles respectively. These concepts are clarified and illustrated in Figure 2.4 of the Methods chapter. Of the 11 non-synonymous mutations detected across the experiment, six are predicted to substantially alter the structure of their homologous crystal structure, as determined by the Φ and Ψ shifts that occur.

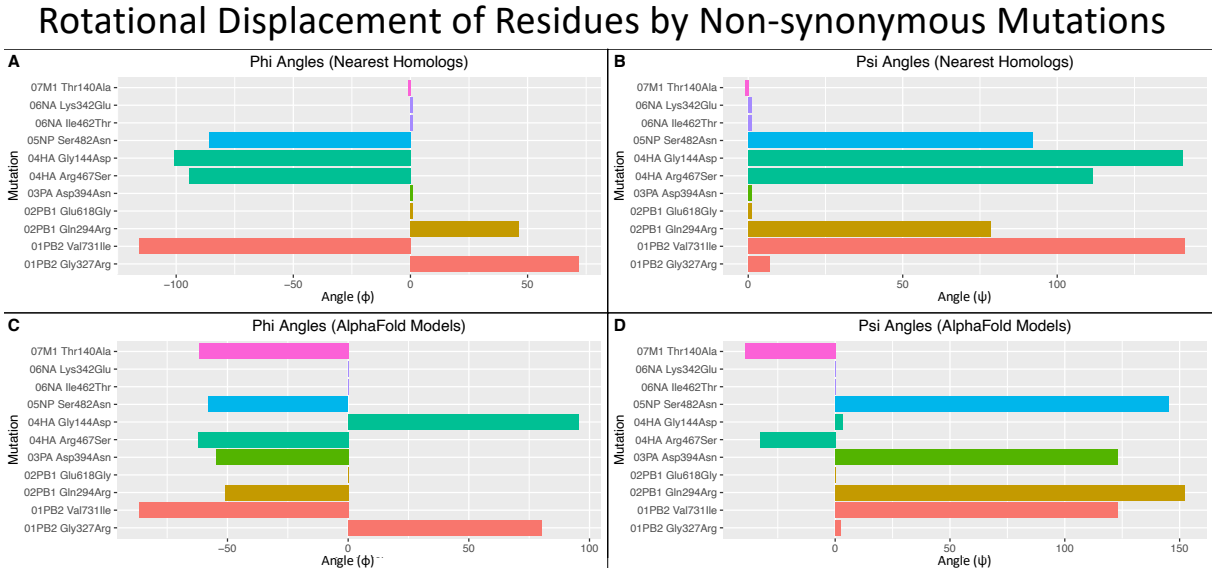


Figure 4.15: Twist angles of non-synonymous mutations from *in silico* experiments in EIV structures. A and B show the phi and psi angles estimated on homologous crystal structures. C and D model the same mutations on structures estimated by AlphaFold. These angles represent structural changes and are a proxy for the impact of mutations on protein tertiary structures.

Not all of the non-synonymous mutations are equal. As shown, some of the physical changes associated with mutations are predicted to be negligible. While this does not necessarily indicate a change in protein function, it indicates that some of the mutations likely alter the structure of viral proteins. These, however, are all modelled virtually, and rotamer activity may not necessarily translate to functional changes. A number of these non-synonymous mutations appear in important active sites of their proteins.

In addition to the predicted spatial changes associated with mutations, point accepted mutation (PAM) matrices describe the likelihoods of amino acid substitutions arising in the context of the chemical properties of amino acid residues. Each cell of a PAM matrix is then the probability of a residue being substituted with the corresponding amino acid after nucleotide mutation(s) (Dayhoff & Foundation, 1979; Pevsner, 2009). As an example, Leucine and Isoleucine have residues with very similar physio-chemical properties and only a single nucleotide substitution (CTT to ATT) can cause this change. In contrast, Leucine and Tyrosine have disparate codon sequences and require more than one nucleotide replacement in the codon meaning that this substitution is less likely than Leu-Ile. From this, we can infer that certain non-synonymous mutations may have greater fitness costs than others.

4.2.10 Physico-Chemical Impacts of Non-synonymous Mutations

The following section summarises what we've learned about non-synonymous mutations in EIV proteins so far and also incorporates physio-chemical properties of residues to potentially explain why we do or do not expect structural changes. Tables 5-12 summarise the structural metrics associated with non-synonymous mutations observed in each genomic segment throughout the transmission experiment. This begins with physical properties (side-chain charge, hydropathy, molecular weight, isoelectric point [the pH at which the amino acid is not electrically-charged] and class of amino acid side chain) of the consensus and mutant residue, then the

PAM250 value for the substitution. These are then followed by two paired columns of predicted Φ and Ψ angles associated with the substitution modelled on both a homolog and the structure predicted above. Homologous proteins were found by searching the Protein Data Bank (PDB) for resolved structures and extracting those with the highest sequence similarity score to sequences sampled from the transmission experiment; the corresponding four-character codes of each homolog are contained in the header of such columns. When the protein is polymeric, the chain used in the prediction follows the four-character ID code.

4.2.10.1 Segment 1: PB2

Site g2191a/Val731 sits in the Karyopherin α binding site; the change to an isoleucine is predicted to cause a substantial effect on the local structure, though chemical properties of the residue itself do not differ dramatically from the original valine. On the other hand, the mutation at position 327 replaces Gly with the hydrophobic, heavy side chain of Arg [g979a/Gly327Arg]. For this reason, the

Polymerase Basic Protein 2							Homolog 6QNW – C		AlphaFold	
Mutation	Charge	Hydropathy	Weight (Da)	Isoelectric Point	Type	PAM 250	φ	ψ	φ	ψ
Gly327Arg g979a	Neutral	-0.4	75.07	6.06	Aliphatic	-3	72.0	7.0	80.0	2.3
	Positive	-4.5	174.20	10.76	Basic					
Val731Ile g2191a	Neutral	4.2	117.15	5.96	Aliphatic	4	-115.8	141.4	-86.6	123.0
	Neutral	4.5	131.18	6.02	Aliphatic					

likelihood of this mutation occurring, according to PAM250 matrix, is $10e^{-3}$. Val731Ile appears but once in this experiment (Table 1), and yet is reported 16 times in the 384 global full-genome EIV sequences; its putative impact on protein structure and function could warrant further investigation.

4.2.10.2 Segment 2: PB1

First, the replacement of Gln with Arg at position 294 [a881g/Gln294Arg] in the PB1 protein is predicted to moderately alter the structure (φ = 46.2, ψ = 78.5) and drastically shift the charge (isoelectic point = +7.54) of the immediate surroundings. The second mutation in PB1, at position 618 [a1853g/Glu618Gly] has little physiochemical impact on the protein, other than a change to local hydrophobicity. While it appears in 16 of the 53 collected samples, it is not present in any of the full-genome EIV sequences available on global databases.

Polymerase Basic Protein 1							Homolog 6QNW – B		AlphaFold	
Mutation	Charge	Hydropathy	Weight (Da)	Isoelectric Point	Type	PAM 250	φ	ψ	φ	ψ
Gln294Arg a881g	Neutral	-3.5	146.15	3.22	Amidic	1	46.2	78.5	-51.0	152.4
	Positive	-4.5	174.20	10.76	Basic					
Glu618Gly a1853g	Negative	-3.5	147.13	5.65	Acidic	0	1.0	1.0	0.0	0.0
	Neutral	-0.4	75.07	5.97	Aliphatic					

4.2.10.3 Segment 3: PA

Only one non-synonymous mutation was detected in the PA subunit of the polymerase complex: g1180a/Asp394Asn. This residue is found in the PB1 binding-site, implying the potential for functional, phenotypic changes.

Polymerase Acidic Protein							Homolog 6QNW – A		AlphaFold	
Mutation	Charge	Hydropathy	Weight (Da)	Isoelectric Point	Type	PAM250	φ	ψ	φ	ψ
Asp394Asn g1180a	Negative	-3.5	133.10	2.77	Acidic	2	1.0	1.0	-54.8	122.9
	Neutral	-3.5	132.12	5.41	Amidic					

4.2.10.4 Segment 4: HA

Three viral haplotypes detected in the experiment show evidence of haemagglutinin mutations: a1401c/Arg467Ser is found in all three haplotypes (E, G and M), while haplotype G has an additional g431a/Gly144Asp mutation. The Gly144Asp singleton is predicted to largely impact the rotation of atoms in the surrounding area, as well as a shift in hydrophobicity and charge. This residue is within the well-described antigenic site A of the HA1 head domain (Kawaoka et al., 1989; Webster & Laver, 1980). Thus, any conformational change to this sensitive region could potentially alter the binding of the EIV haemagglutinin to host cell receptors and/or host epitope-binding immune molecules; both possibilities would have a dramatic effect on viral fitness.

Haemagglutinin							Homolog 4UNW		AlphaFold	
Mutation	Charge	Hydropathy	Weight (Da)	Isoelectric Point	Type	PAM 250	φ	ψ	φ	ψ
Gly144Asp g431a	Neutral	-0.4	75.07	5.97	Aliphatic	1	-101.0	140.6	95.7	3.3
	Negative	-3.5	133.10	2.77	Acidic					
Arg467Ser a1401c	Positive	-4.5	174.20	10.76	Basic	0	-94.6	111.6	62.2	32.7
	Neutral	-0.8	105.10	5.68	Hydroxylic					

The more abundant mutation in our dataset, Arg467Ser, has similarly large effects on the rotational orientation of the residue. This residue is located on the stalk domain of haemagglutinin, though before the transmembrane portion of the stalk and so is not expected to alter function or efficiency of the viral protein.

4.2.10.5 Segment 5: NP

The g1445a/Ser482Asn mutation found in the viral nucleoprotein is the second most abundant mutation observed in our dataset. It is predicted to have large effects on the spatial conformation of the protein (φ = -86, ψ = 92) but otherwise the wild-type (Ser) and mutant (Asn) residues have similar physiochemical properties.

Nucleoprotein							Homolog 2IQH		AlphaFold	
Mutation	Charge	Hydropathy	Weight (Da)	Isoelectric Point	Type	PAM250	φ	ψ	φ	ψ
	Neutral	-0.8	105.09	5.68	Hydroxylic					

Ser482Asn g1445a	Neutral	-3.5	132.12	5.41	Amidic	1	-86.0	92.0	-57.8	145.1

4.2.10.6 Segment 6: NA

The two non-synonymous mutations reported in segment six, a1024g/Lys342Glu and t1385c/Ile462Thr, neither change the structure of the protein (referring to their φ and ψ angles) or are less likely to be substituted for this residue than any other. They are, thus, not expected to affect protein function. The likelihood that Lysine is replaced by Glutamine, or that Isoleucine is replaced by Threonine, due to similarities in residue properties, rather than being substituted due to chance, is zero (as represented by their PAM250 score of 0).

Neuraminidase							Homolog 5HUK – A		AlphaFold	
Mutation	Charge	Hydropathy	Weight (Da)	Isoelectric Point	Type	PAM 250	φ	ψ	φ	ψ
Lys342Glu a1024g	Positive	-3.9	146.19	9.74	Basic	0	1.0	1.0	0.0	0.0
	Negative	-3.5	147.13	5.65	Acidic					
Ile462Thr t1385c	Neutral	4.5	131.18	6.02	Aliphatic	0	1.0	1.0	0.0	0.0
	Neutral	-0.7	119.12	5.60	Hydroxylic					

4.2.10.7 Segment 7: MP

The single mutation seen in segment seven is minor; Threonine and Alanine have relatively similar physio-chemical properties and so are not expected to substantially alter the protein structure or function. This mutation is transient, vanishing from the population by the end of the experiment.

Matrix Protein 1							Homolog 1EA3		AlphaFold	
Mutation	Charge	Hydropathy	Weight (Da)	Isoelectric Point	Type	PAM 250	φ	ψ	φ	ψ
Thr140Ala a418g	Neutral	-0.7	119.12	5.6	Hydroxylic	1	-1.0	-1.0	-61.8	-39.2
	Neutral	1.8	89.09	6.0	Aliphatic					

4.2.10.8 Segment 8: NS

No non-synonymous mutations were detected in segment 8.

The eleven non-synonymous mutations observed are expected to have a range of effects on protein structure and function. Some are minimal; for example, NA-Ile462Thr involves substitution with a residue with similar chemical properties and from *in silico* testing is not expected to alter protein structure, although this may be confounded by its location at the end of the protein chain Other mutations, however, do show a proclivity for structural and functional changes; in haemagglutinin the Gly144Asp mutation causes a substitution with a residue almost twice as large (75kDa to 133kDa) and with a much more acidic isoelectric point (5.97 to 2.77) leading to a large shift in protein structure around this site.

4.3 Discussion

From the 53 sequences collected over the course of this transmission study, I sought to understand possible changes to viral genomes while being transmitted through vaccinated and unvaccinated hosts. After assembling the sequences and reporting mutations within the alignment, I addressed the problem from a range of angles. First, I used a phylogenetic approach whereby the sampled sequences were analysed to determine whether host immunity or transmission chain could explain genetic changes.

The putative impacts of mutations on EIV protein structure and function were analysed next; tools qualifying protein properties, surface availability and propensity for interaction with host immune molecules and cells were used to characterise each of the ten major EIV proteins residue-by-residue. With these features described, *in silico* experiments were used to explore putative effects that non-synonymous mutations may have on EIV replicative fitness. Would certain amino acid substitutions alter how exposed a region of the protein was, and would that impact the binding of antibodies?

An important aim I had for this work was to create high quality models of the EIV proteome, which had previously consisted of only a single crystal-resolved haemagglutinin structure. Establishing Equine Influenza Virus H3N8 as a suitable model virus was necessary were these findings to be applicable to influenza epidemiology as a whole. By estimating the structures of EIV proteins, these models could then be compared to resolved IAV proteins to determine their similarity and hence whether these results could potentially apply to other IAVs. This meant estimating and validating structures from assembled genomic sequences, before undertaking a comparative analysis with other IAV proteins and finally performing *in silico* experiments simulating the effects of non-synonymous mutations on protein structure. Ultimately, most of the major proteins of EIV were estimated with great confidence (>70% pLDDT) and were found to closely resemble crystal structures of other IAV proteins.

4.3.1 Sequence Analysis

4.3.1.1 Mutations detected

Collating consensus sequences from the nasal swabs of horses infected with EIV, I observed 21 mutations across the 13kb viral genome. The majority of these mutations (n = 16) were singletons, though two appeared independently in both transmission chains: PA-c201t and HA-a1401c. Both populations homogenised upon infecting immunologically naïve hosts, resulting in two distinct end-point virus populations: **F** became fixed in the single group and **J** in the multi group. Viruses in V_S hosts were attempting to circumvent strain-specific adaptive immune recognition whereas in the V_M hosts, the breadth of immunologic memory appeared to create an environment with weaker selective pressures for infecting viruses. Greater viral population, fewer fixed mutations and a slightly slower branch rate in viruses from the V_S hosts indicate lower selective pressures in these hosts.

The appearance of the non-synonymous HA-a1401c/Arg467Ser mutation in both transmission chains is surprising. For a mutation to appear *de novo* at the consensus level in three individuals (Mul_3A, Mul_4A and Sin_2B) suggests either low levels of circulation in the original inocula or a selective pressure for its creation in the viral population. It is notable that this mutation is a transversion, a comparatively unlikely form of point-mutation.

4.3.2 Phylogenetic Analyses

Analysing the phylogenetic trees of viruses from each transmission group, we estimate a higher mutation rate in viruses of the single group than of the multi group ($1.57e^{-3}$ and $8.77e^{-4}$ substitutions per annum respectively), though these differences are not statistically significant. These values align with previously published substitution rates of influenza A viruses (Lloyd et al., 2011; Murcia et al., 2013). Though the rates are similar, viruses in the single group appear to change faster than those in the multi group; one possible reason for this is the greater selective pressure placed upon viruses that are infecting a host with pre-existing immunity. As the shedding analysis showed, viral populations of hosts in the single group were smaller than those in the multi chain. Hence, we assume that the single group viruses encountered more barriers to replication (namely a primed adaptive immune response). A greater mutation rate may be taken as evidence of the virus attempting to adapt to this challenge. In the multi group, it may be hypothesised that viruses did not have to contend with the same degree of immune recognition.

Viruses appear to face different challenges across each transmission chain. In vaccinated hosts of the single chain, the immune system is primed to specifically combat the challenge strain. Thus, we could expect strong, specific adaptive immune responses when these hosts are naturally infected by the challenge strain. In the shedding analyses, vaccinated hosts did have lower viral loads than naïve hosts. Vaccinates in the multi chain have a broad history of IAV exposure. Original Antigenic Sin theory hypothesises that the multiple exposures will confound the specificity of antibody binding, decreasing the protective response. This certainly matches the patterns seen in the viral load and evolutionary rate of these viruses, suggesting an increased viral population size and lesser selective pressure compared to the single transmission group. As vaccinated hosts in the single group (V_S) had already been exposed to antigens from the challenge strain (Newmarket/5/03) five times over the previous year, their adaptive immune systems were primed to respond rapidly when they were naturally infected by a virus derived from lab-grown Newmarket/5/03 EIV. It is supposed then that viruses in V_S hosts experience different selective pressures to those in V_M hosts; the rapid activation of adaptive immunity causing viruses in V_S hosts to diversify to a greater degree, in order to escape elimination.

4.3.3 Selection Analyses

Multiple algorithms were used to examine the viral evolution for evidence of purifying/negative or enriching/positive selection. As many of these sequences show little diversity overall, the analyses were not well powered to detect selection. Sites

that appeared to be under selection were the negatively selected synonymous mutation PB1-Gly500 as well as the positively selected HA-Gly144Asp and NP-Ser482Asn. Individual sites throughout the protein were found to differ in their levels of selection, indicating these sites may play an important role in protein function and, therefore, warrant functional studies.

4.3.4 Consensus Diversity

The population diversity, as measured by Shannon Entropy, was greater in the vaccinated hosts than in naïves in each transmission chain. Perhaps unexpectedly, N_M hosts had the lowest overall diversity; I attribute this to the vastly different selective pressures between the two groups. The strong adaptive immune response in vaccinated hosts already primed to EIV infection subdues viral replication and thus, in theory, the viruses endure greater mutational plasticity to attempt to adapt to the challenge. Contrary to Shannon Entropy, other diversity metrics do not necessarily estimate that the highest diversity is found in N_M hosts. Tajima's D, as estimated by PoPoolation, observes diversity twice as high in hosts of the single group to those in the multi group.

While univalent vaccines are obviously better at granting protective immunity against specific strains of EIV (as seen previously by the viral load), they clearly are unable to provide fully neutralising immunity. They in fact appear to be driving the diversification of viruses; applying strong selective pressures. Similar levels of diversity in both vaccinated and unvaccinated hosts of the multi-strain group suggests that selective pressures for the virus were not associated with vaccination status in this group. Alternatively, viral genomes collected from horses vaccinated with a univalent vaccine (V_S) have four-fold higher diversity than unvaccinated hosts in the same transmission group. This indicates a much greater challenge for viruses replicating in V_S hosts than in unvaccinated hosts; viruses appear to be diversifying greatly in order to attempt to overcome host adaptive immune selection. Such a dramatic response is not however seen in the multivalent vaccine group (V_M), possibly due to less concerted immune activation.

4.3.5 Protein Analyses

Coding sequences from the consensus genomes were translated *in silico* to protein sequences and additional tests of protein properties were carried out. Many tools exist hosted on web servers to estimate the properties of proteins from their primary structure alone. These bioinformatic tools were first used to simply assess basic properties such as weight, charge and hydrophobicity of each of the ten main EIV proteins. Though easily obtained, these features were not published for H3N8 viruses, and therefore I saw a gap in the knowledge base. Beyond this, however, measuring the properties of H3N8 EIV allowed for comparison with other, more popularly studied IAV. As this study is in large part meant to be applicable to the dynamics of all mammalian influenza viruses, knowing how similarly H3N8 EIV proteins resembled the proteome of other IAV subtypes granted us some validity in applying our conclusions to epidemiological and evolutionary dynamics of influenza in general.

4.3.6 Antigenicity

The reported non-synonymous mutations in the surface proteins HA and NA are expected to have little effect on protein antigenicity. Despite substantially different physio-chemical properties and a large change in predicted twist angles, the two mutations in HA (Gly144Asp and Arg467Ser) have an equal likelihood of arising (according to a PAM250 matrix). Furthermore, selection is only detected at the first mutation (Gly144Asp), which is estimated to have no impact on the antigenicity of the protein as a whole. Both NA mutations resulted in a decrease in predicted antigenicity, despite little to no impact on the protein structure. This is mirrored in calculating the probability of each site to be an epitope; only NA Lys342Glu had a substantially different epitope score (a decrease in probability of 14%) and was estimated to no longer be an epitope. The impact of this Lys342Glu substitution in neuraminidase implies a shift towards immune evasion, evidenced by lowered availability of the site to immune cells plus decreased antigenicity.

4.3.7 Structural Modelling

One novel finding presented here is the structural modelling of the whole EIV proteome. Only a single EIV protein structure, haemagglutinin, has thus far been resolved by crystallography. The remaining major proteins of equine H3N8 have thus relied on homologous protein structures for any structural analyses. An aim of this study was to obtain reliable and accurate predictions of protein structures using *in silico* modelling. Actual crystallographic resolution of proteins is an expensive and time-consuming labour requiring highly-skilled technicians; modelling *in silico* can grant us a reasonably trustworthy structure with limited time and expense. With dependable protein structures, analyses such as targeted drug/antibody manufacture or binding-affinity can help elucidate protein activity and inform on treatment of viral infections.

Though carried out *in silico*, with all the caveats accompanying such modelling, I present characterisation of the ten main proteins of the EIV proteome with corresponding properties and localisations. Understanding the placement and properties of these viral proteins enables comparative approaches between EIV and other influenza A viruses, and further broadens the use of EIV to model other IAV systems.

The equine influenza virus proteome has not, as of yet, been fully resolved and much of what is known is inferred from studies of other IAV; the mapping of antigenic sites on EIV H3N8 haemagglutinin, for example, is based on H3 haemagglutinin subtypes from human infections. Hence, generating 3D structural models from the high-quality genomic sequences obtained over the transmission experiment was an important contribution I sought to add to the knowledge base of EIV biology. I obtained mixed results in terms of the reliability of structural predictions; transmembrane homopolymers (HA, NA and M2) were particularly difficult for AlphaFold to reliably model, despite finding highly homologous sequences for each of these proteins. It is recognised that transmembrane proteins are difficult to model in this way due to the complexities of protein-lipid interactions.

4.3.8 Structural Analyses

Using Ramachandran (twist) angles of amino acid chains, I examined the structural impacts of non-synonymous mutations. I decided to use two versions of the seven proteins in which non-synonymous mutations were reported, first fully-resolved influenza homologs and secondly the *in silico* structural EIV models. The positioning effect of an amino acid substitution can indicate whether the mutation impacts the protein tertiary or quaternary structure. The first observation was the difference in twist angles (ψ and φ) when estimated on homologous IAV proteins compared to those estimated on the *in silico* EIV structural models. These disparities indicated either a poorly modelled structure, which contradicted the results seen with the LDDT plots (Figure 4.12), or that the homologs are sufficiently different from the EIV proteome to cause errors in structural analysis.

4.3.9 Physio-chemical Differences in Non-synonymous Mutations

Having explored evolutionary and viral load trends of samples, I then referenced potential functional changes to the observed non-synonymous mutations of EIV proteins. The eleven non-synonymous mutations detected throughout our transmission experiment range from large impacts on local protein structures to substitutions by amino acids with very similar residue properties. Further laboratory work on these candidate mutations is needed in order to investigate possible phenotypic effects on protein function and fitness.

5 Influenza Virus Evolution at the Sub-consensus Level

As in the chapter above, studies often simplify genomic sequences to create a consensus genome, the most abundant viral genome is assumed to represent all of the viruses within a sample. Realistically however, the error-prone replication of influenza virus genomes leads to heterogeneous populations within infected hosts. In an attempt to capture some of this heterogeneity, virus genomes present at frequencies below the dominant genome were examined from horses in two transmission experiments. Sub-consensus genomes collected from hosts on multiple days and from hosts connected in a transmission chain allowed for the study of intra-host and inter-host diversity of influenza viruses. These data also enabled the estimation of the size of the transmission bottleneck; how many viruses needed were needed to establish an infection that reflected the observed genetic diversity. Contrary to the patterns seen in the consensus data, vaccinated hosts had much lower diversity of sub-consensus genomes suggesting that the application of strong selective pressures such as host adaptive immune response created environments unfavourable to broad diversification.

5.1 Introduction

5.1.1 Reporting sub-consensus viral genomes

With the development of deep-sequencing tools and metapopulation genetics, we are now able to see the genetic diversity of both intra- and inter-host pathogen populations (Gallagher et al., 2018; Gelbart et al., 2020; Lauring, 2020; Nelson & Hughes, 2015). Next-generation sequencing (NGS) technologies and the genome assembly bioinformatic processes are now sensitive enough to account for experimental error when assembling read libraries. This enables sub-consensus mutations to be recognised with some confidence that the variation is not generated by erroneous sampling (McCrone et al., 2020).

The mutant spectra present in an infected individual can have a range of clinical and public health repercussions (Domingo & Perales, 2019). One example of this is in chronic viral infections like HIV, where the emergence of drug-resistant strains can quickly dominate the overall viral population. Thus, combination treatments are required to combat the dominance of drug-resistant viruses. Rather than preventing such resistant strains from emerging, a combination of therapeutics with differing mechanisms of action can ensure that when sub-consensus variants that are resistant to one drug appear, there are multiple other anti-retrovirals restricting their proliferation. This highly-active anti-retroviral therapy (HAART) usually consists of at least one reverse transcriptase inhibitor in combination with an inhibitor of viral protease and/or integrase proteins (Waters et al., 2016).

The diversity generated during infection of hosts provides the “raw material” for global genetic drift (Rodríguez-Nevado et al., 2018; Simmonds et al., 2019). Understanding the causes and consequences of viral mutant spectra is obviously important for virologists, public health workers and clinicians (Houlihan et al., 2018; Kwong et al., 2015); but how do we actually detect and observe them? By definition, genomes with sub-consensus mutations are a small minority of the overall population and so conventional genome amplification and sequencing techniques do not reliably amplify all of the genomes present equally and may display biases in the sequences they enrich. Many bioinformatic procedures were designed specifically to exclude

spurious outliers, so how then do we obtain this information from a viral sample such as a clinical specimen? Balancing the need for preserving data present at very low proportions while also ensuring that sample-preparation and technical errors are minimised is thus a key function of bioinformatic processing pipelines. Many pipelines incorporate sequence metadata (base quality [Phred score], read quality [Q score] and, if possible, strandedness [0-100%]) into their assignment of 'variant' or 'error'.

Further, the epidemiological mechanisms in which these variants are maintained, transmitted and/or lost from the viral population can illuminate specific evolutionary bottlenecks that viruses experience as they infect subsequent hosts (Sobel Leonard et al., 2017; Stack et al., 2013). Referencing the term used in macro-ecology, a genetic bottleneck describes an event that severely reduces the amount of genetic diversity within a population (Ørsted et al., 2019; Rees et al., 2009). In obligate parasitic pathogens, such events occur during transmission, as well as when migrating between host body compartments, wherein a subsection of the population in one host leaves to establish infection in a secondary host. These bottlenecks can be heavily influenced by the transmission route. A pathogen spread through close-contact over a long period of time (Frothingham, 1999), such as *Mycobacterium leprae* causing Hansen's Disease (neé leprosy) is afforded ample opportunities for multiple cumulative transmission events that have a higher chance of sharing the within-host diversity of the bacterium leading to conserved bacterial populations (Weng et al., 2011). In contrast, fomite transmission of an acute respiratory virus such as IAV is limited to a snapshot of the viral population present in the upper respiratory tract at a time when a suitable surface is seeded (Bean et al., 1982; Thompson & Bennett, 2017; Wißmann et al., 2021). Selective pressures enacted upon viruses undergoing transmission bottlenecks shape the overall epidemic viral population; they may also influence which mutations can become fixed in the broader viral population (Johnson & Ghedin, 2020; Sigal et al., 2018). Some viruses, notably HIV, even display distinct transmission phenotypes which have only been observed through low-frequency variant (LFV) analyses (Kariuki et al., 2017).

Once a sizeable proportion of the total viral population has been sampled and sequenced, bioinformatics tools must then distinguish mutant variants from the consensus genome while excluding variation caused by experimental/sequencing error. Due to the sheer quantity of viral genomes in most viral samples, most minority variants fall below a set threshold and thus are excluded from analysis. This threshold varies depending on the efficacy of the genome amplification technique and the specificity of the sequencing procedure but, as a standard, most labs place a cut-off value at genomes that constitute less than 1% of the total viral population meta-genome after amplification (Domingo et al., 2017; Lauring, 2020).

Deep sequencing technologies can generate deep-sequence data and illustrate the genetic composition of samples at very high coverage. Deep-sequencing approaches enable the detection of rare variants in samples, but can also have the undesired effect of generating and amplifying sequencing errors and artefacts. Distinguishing real variants from such noise is not straightforward. Errors in sequencing can arise at many steps, commonly during reverse transcription, PCR amplification and the sequencing process itself; most deep sequencing pipelines can now reliably detect variant genomes present at or above only 1% proportion of a sample (Xue et al., 2018). Following these variants throughout the course of disease

in an infected individual can help infer transmission events (De Maio et al., 2016, 2018) and even assist in estimating transmission trees (Campbell et al., 2019). As genetic sequencing technologies improve, the ability to explore the dynamics of viral diversity within hosts is expanding, held back only by the limitations of informatically distinguishing technical errors from true, naturally occurring mutations.

5.1.2 Influenza Within-host Variation

Influenza A Viruses have high mutation rates ($10e^{-4}$ to $10e^{-5}$ substitutions per nucleotide per replication (Lauring, 2020; McCrone et al., 2020)), caused by their large population sizes, rapid replication and low-fidelity polymerase (Dunham et al., 2009; Grear et al., 2018; Smith et al., 2009). Over an acute IAV infection, viral load averages between 10^5 and 10^7 copies/µl (Hughes et al., 2012; Neira et al., 2016; Ward et al., 2004), so we can expect to observe significant variation during infection. The proclivity of these viruses to mutate means that the viral population within an individual is often genetically distinct enough to form a viral cloud, or mutant swarm (Ørsted et al., 2019) in which subsections of the population may exhibit stark differences to the parsimonious consensus of all genomes in the population.

However, the actual impact of genomic variability of RNA viruses in terms of influencing pathogenic outcomes is poorly understood (Jombart et al., 2014). Mutations in all influenza proteins occur at an observable rate within a single host (Chen & Cui, 2017; Illingworth & Mustonen, 2012; Kenah et al., 2016). But how relevant is this diversity on an epidemiological scale given that the vast majority of mutations observed at the consensus level are transient? Previous transmission studies have implicated hosts with chronic influenza as being disproportionate sources of IAV evolution on a global scale (Houlihan et al., 2018). Hypothesising that lessened selective pressures enacted by weaker immune responses in addition to the longer period of disease in chronically infected hosts simply allows for more mutations to both appear *de novo* and to survive. Indeed, Lumby et al. (2020) reported that long, non-acute IAV infections allow for greater periods of time for selective forces to act upon viral populations. However, as chronic infections form a minority of overall influenza infections in any host population, tracking mutations through immunocompetent hosts to observe whether sub-consensus variation is stochastic or driven directionally may provide insight into the way within-host variation can shape influenza virus population structures.

Viruses exist in an ecological community; virions will be infecting host cells amidst a plethora of other competing, complementary and/or antagonistic viruses and bacteria present in the host mucous membranes. Additionally, once infection with a particular virus is established, the diverse range of progeny from that virus will also be interacting with each other, often competing but in some instances playing a complementary role (Leeks et al., 2018). The diverse mutant spectra generated in an infected host are then subject to selective pressures within the host (Bessière & Volmer, 2021).

Key to understanding and measuring this range of sub-consensus viral genomes are the methods which may be used for quantifying genetic diversity. Myriad methods of describing diversity exist, scaling in complexity and abstraction, many

of which are leveraged from studies of macro-organism ecology (Reeve et al., 2014). Commonly used methods, however, range from counts of non-reference nucleotides (such as Mutational Frequency or Simpson's Index) to more complex methods accounting for unequally polymorphic sites (per-site Shannon Entropy) or for sampling bias (nucleotide π distance) (Fuhrmann et al., 2021).

Previous studies have shown that despite the relatively high *de novo* mutation rate of most IAV, those mutations arising later in infection have a much lower chance of surviving to be transmitted to a secondary host (Sigal et al., 2018). Hence, though traditionally associated with rapid mutation, within-host diversity of influenza A viruses is generally low (Xue & Bloom, 2019).

An important caveat of these and other studies is that measurements of viral population diversity are reliant on a subsample of genomic information taken at a specific point in time (Bessonov et al., 2020; Didelot et al., 2017; Houldcroft et al., 2017). Genetic diversity over the course of infection in a single host (from exposure to colonisation, infection, possible transmission events and finally clearance/death) fluctuates in response to host environments and both intra- and inter-species competition (Pauly et al., 2017; Poon et al., 2016).

5.1.3 Transmission Bottlenecks of Naturally Transmitted EIV

Transmission events of pathogens are both a necessity for continued infection, and thus survival, and a huge disruption to population dynamics. A subset of viral particles leaves the current host, seeds an infection in a new host and becomes a founder population in this secondary host. This founder population is composed of a collection of viral genomes that may be wholly unrepresentative of the population size, diversity and even phenotype of viruses in the donor host. In addition to impacting population dynamics of viruses, transmission bottlenecks can also illustrate putative links in epidemiological networks. Influenza A viruses usually experience relatively tight bottlenecks (Sobel Leonard et al., 2017; McCrone & Lauring, 2018), measured at between 1 and 5 viral particles in ferrets, mice and guinea pigs (Bergstrom et al., 1999; Varble et al., 2014). Some estimates, like results obtained by Sobel Leonard et al. (2017), have shown that bottlenecks between human transmission pairs range from 100 to 200 IAV particles. However, these large values have since been re-examined and have been shown to be erroneous due to contamination within read-pairs (Sobel Leonard et al., 2019). The actual values are much lower, below ten particles and are comparable with the studies presented above.

Further, the epidemiological mechanisms by which these variants are maintained, transmitted and/or lost from the viral population can tell us a great deal about the specific evolutionary bottlenecks viruses experience as they infect subsequent hosts (Sobel Leonard et al., 2017; Stack et al., 2013). Selective pressures enacted upon viruses undergoing transmission bottlenecks shape the overall epidemic viral population and determine which mutations become fixed in the broader viral population (Johnson & Ghedin, 2020; Sigal et al., 2018). Some viruses, notably HIV, even display distinct transmission phenotypes which have only been observed through low-frequency analyses (Kariuki et al., 2017).

Most notably, due to its importance, HIV-1 infection is an exemplar culmination of all the dynamics discussed thus far (Mak et al., 2020). Lots of variant genomes appear below consensus level within infected hosts (Frost et al., 2018; Theys et al.,

2018), requiring multiple chemotherapeutics with overlapping mechanisms of action (highly-active anti-retroviral therapy [HAART]) to combat the emergence of drug-resistant strains (Power et al., 2016). Additionally, studies of the transmission bottleneck show distinct phenotypes, from those surrounding the transmission event (associated with host colonisation) to those present in established infections (Kariuki et al., 2017; Zwart & Elena, 2015). Finally, detailed HIV-1 transmission networks have been made at varying scales, utilising epidemiological and genetic data, to reconstruct transmission trees. These have been used for research, public health and even legal purposes, showing the interplay of within- and between-host evolution in shaping HIV-1 population dynamics. Lessons learned from other viral systems can provide insight into dynamics of influenza A infections despite differences in pathology, epidemiology and biology between the viruses (Giardina et al., 2017; Yu et al., 2018).

5.1.4 Aims

With the diverse composition of variants in EIV populations during natural transmission chains, I aimed to understand the role of transmission bottlenecks in shaping the evolution of influenza viruses. Furthermore, as hosts had heterogeneous immune experiences (naïve or vaccinated with either exclusively immunogens matching the challenge strain or alternatively, a range of four EIV) differing adaptive immune responses may also be reflected in viral populations. Thus, I shall see the fate of genetic diversity within hosts, and use this diversity to quantify transmission bottlenecks, in order to understand the limitations and influences that transmission bottlenecks place on viral evolution.

Myriad ways exist to extract these variants from the oft-times deep read libraries generated by deep sequencing. First, I explore a range of publicly available tools designed to extract such variants from large viral genomic assemblages. In comparing these tools, I process a range of datasets using bioinformatic tools with default settings. The use and analysis of such deep-sequencing data can vary dramatically depending on the experimental procedure, sequencing technology and bioinformatic pipelines used, so multiple datasets were selected in an effort to account for this variation. To establish this, five control datasets were chosen in order to test scenarios with simple, regular mutations up to complex sequence libraries collected from clinical samples. Both publicly available sequence data and read libraries simulated *in silico* with ART-Illumina datasets were used and thus both represents real and synthetic sequences of Influenza A viruses.

Their segmented genome and overlapping reading frames provide sizeable processing challenges to tools, many of which may be capable only of analysing more basic, linear genomes, or indeed are not designed for viral genomes at all.

In the following comparative analysis of published variant call tools, I aim to select one or more variant call tools (VCT) suitable for examining EIV sequences obtained from two transmission experiments. Experimentally testing the advantages and disadvantages of an array of tools, I then establish a bioinformatic pipeline with which to process H3N8 EIV sequences and document the emergence, elimination and fluctuation of low-frequency variants along natural transmission chains.

In order to examine these evolutionary and transmission dynamics, deep-sequencing data from the above transmission experiment collected by daily nasal swabs allowed the tracking of the trajectories of viral variants within and between hosts. From these highly detailed descriptions of viral populations, sub-consensus variation was examined, with additional detailed investigations at sites of known consensus polymorphisms. The depth of genomic detail also enabled estimation of transmission bottleneck sizes. By utilising the beta-binomial modelling of Sobel Leonard et al. (2017), differences between two populations of viral genomes can be compared; the proportion of shared variants and the frequency at which they appear can estimate how many genomes passed from one population to the other. As I have at least two samples from most individuals, this enabled quantification of within-host transmission events, where the "donor" and "recipient" hosts were samples from the same individual on day_x and day_{x+1}, and inter-host transmission events.

5.2 Methods

5.2.1 Comparative Analyses of Variant Call Tools

5.2.1.1 Control Datasets

To thoroughly evaluate variant calling tools across a breadth of increasingly complex datasets, a dataset of simulated influenza reads was created and combined with four previously published influenza deep sequence datasets:

Simulated Dataset

ART sequence simulator by Huang et al. (2012) is able to produce deep sequence read data synthesised from sample reference genome and defined Illumina machine error profile data. In doing so, ART is able to produce single-end or paired-end reads with the error rates seen in sequencing technologies. *In silico* generated sequences with mutations spiked at known, regular intervals and proportions were created under two Illumina sequencing procedures, a single-end (Illumina Genome Analyser II – library "GA2") and a paired-end (Illumina NextSeq – library "NS50"). Both simulated libraries were based on A/Equine/Newmarket/2003 (H3N8) to get datasets of non-human IAV reads. Mutant reads contained nucleotide substitutions every 50, 100 and 200 bases and comprised 5%, 10% and 25% of the total reads in the sample, respectively. A full diagram of where mutations are spiked into the genome is shown in Supplemental Figure 5.1.

Sample Dataset 1 (McCrone16)

Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling: Bioproject PRJNA317621

The samples published in this dataset are all lab-grown IAVs adapted from A/WSN/1933 (H1N1). They have a range of viral population sizes, as determined by copy number. Twelve samples published from *in vitro* mutagenesis experiments were selected, with known variant frequencies and positions. Proportions of variant bases ranged from 0.2% to 5% at designated variant sites. Each library was pooled for the removal of adapters by gel isolation with the GeneJet gel extraction kit prior to sequencing by an Illumina HiSeq 2500, with 2×125 paired-end reads. Additionally, twenty nucleotide mutations were experimentally spiked into A/WSN/1933 (H1N1)

viral genomes with a pHW2000 reverse-genetics system, listed below. These mutant genomes were then mixed with wild-type genomes in known concentrations to create samples wherein 2%, 1%, 0.5% or 0.2% of the genomes present carried these 20 mutations. The results of the authors' variant calls (using the tools deepSNV and LoFreq) were then compiled into a CSV file which was then used to create a reference (https://github.com/lauringlab/Benchmarking_paper/blob/master/data/process/2015-6-23/Variants/all.sum.csv).

1. PB2: a1854g, a440t and a1167t;
2. PB1: g599a, g1764t and t1288a;
3. PA: t964g, t237a and a1358t;
4. HA: t1583g, g1006t and g542t;
5. NP: a454c and a1160t;
6. NA: g1168t and c454t;
7. MP: t861g and a541c;
8. NS: g227t and a809g

Sample Dataset 2 (McCrone18)

Stochastic Processes Constrain the Within and Between Host Evolution of Influenza Virus: Bioproject PRJNA412631

Influenza virus samples were collected from the Household Influenza Vaccine Effectiveness (HIVE) study by Ohmit et al. (2015). Households of at least 3 individuals in Michigan, USA were followed prospectively from October to April. Nasal & throat swabs are collected by the individual on appearance of respiratory symptoms for viral identification via RT-PCR. Over five seasons of observation (2010-2015), 77 cases of A/H1N1$_{pdm09}$ and 313 cases of A/H3N2 infection were reported by the authors. Approximately half of the cases $\left(n=\frac{166}{313}\right)$ were identified in the 2014–2015 influenza season, hence this is the subset of samples from which we pulled eight random read libraries.

cDNA of all eight genomic segments was amplified from 5µl of viral RNA with universal influenza A primers. Libraries were then sequenced on Illumina HiSeq 2500 with paired-end reads. Variants were called using a modified DeepSNV protocol laid out in previous studies (McCrone & Lauring, 2016). Eight A/H3N2(2014-2015) samples with a range of genome copies were chosen for the comparative analysis, all originally aligned to the reference strain A/New York/WC-LVD-15-031/2015 (H3N2). Notably, only the results of segment 4 (HA) are published as accompanying data, hence results from this project will only be shown for segment 4. Variants that the authors detected, using DeepSNV, were published with their estimated frequencies alongside the paper – specifically the supplementary file 'fig1-data4-v3.csv' (https://doi.org/10.7554/eLife.35962.011). This record of identified SNV was then used as the control against the outputs of my *in silico* testing of tools.

Sample Dataset 3 (Han21)

Within-Host Evolutionary Dynamics of Seasonal and Pandemic Human Influenza A Viruses in Young Children: Bioproject PRJNA722099

Han et al. (2021) collected samples from children observed in a longitudinal influenza study, comprising 303 sequences (H1N1$_{pdm09}$: 47 nasal, 12 throat; H3N2: 146 nasal, 98 throat) collected from 82 longitudinally-sampled individuals in South-East Asia. H3N2 sequencing libraries were prepared using a Nextera XT DNA Library Preparation kit (Illumina, FC-131-1096) then sequenced using Illumina MiSeq 600-

cycle MiSeq Reagent Kit v3 (Illumina, MS-102-3003). $H1N1_{pdm09}$ samples were sequenced in Roche FLX+ 454.

The variants detected using a custom python code were published as supplementary CSV files within the code base on GitHub (https://github.com/AMC-LAEB/Within_Host_H3vH1/tree/main). The presence and quantity of variants called by the authors were then used as a control dataset in my comparative studies.

Sample Dataset 4 (Poelvoorde22)

A General Approach to Identify Low-Frequency Variants Within Influenza Samples Collected During Routine Surveillance: Bioproject PRJNA692424

The dataset from Poelvoorde et al. (2022) has sequences from surveillance efforts from the 2016-17 influenza season in Belgium. This work looked for sub-consensus variants in influenza samples from regular surveillance of patients. A total of 48 (24 $H1N1_{pdm09}$ and 24 H3N2) influenza viruses from the 2016–2017 Belgian influenza season were collected by the authors and eight were randomly selected from each group for our analyses. All libraries were sequenced on an Illumina MiSeq platform to produce 2×250 nucleotide paired-end reads. The sheet 'InputR' in the 'Input_File.xlsx' found in Supplementary Methods S2 contained the authors' findings when they called variants from their dataset; this then formed the reference dataset for my comparative analyses.

5.2.1.2 Data Set Overview

A brief overview of the 5 datasets used to evaluate variant calling tools is presented below (Table 1). Multiple datasets were chosen to control for the various ways in which deep-sequencing datasets are produced, removing potential bias in instances where results are heavily influenced by the analytic pipeline used. Repeated analyses on these disparate datasets function as technical replicates. Supplementary Table 1 reports the exact library used for each of the 46 samples, alongside associated metadata.

Table 5.1: The datasets used to test and compare variant call tools together with the NCBI taxonomy ID of the reference strain and the average number of reads in the selected samples.

Dataset	Reference	NCBI taxID	Average Reads	Samples
SimData	A/Equine/Newmarket/5/03 (H3N8)	568375	1,758,612	2
McCrone 2016	A/WSN/1933 (H1N1)	382835	5,985,094	12
McCrone 2018	A/New York/2015 (H3N2)	1895544	928,876	8
Han 2021	A/Brisbane/10/2007 (H3N2)	476294	339,744	4
	A/California/04/2009 (H1N1)	641501	2,175,225	4
Poelvoorde 2022	A/Bretagne/7608/2009 (H1N1)	1506405	358,616	8
	A/Victoria/1003/2012 (H3N2)	2044087	275,214	8

5.2.2 Variant Calling Pipelines

First, paired-end reads were trimmed using PRINSEQ (Schmieder & Edwards, 2011). Then reads were assembled and mapped onto the reference genomes with the standard samtools pipeline. Reads were mapped using Bowtie2 v2.3.5.1 with default options in '*local*' mode (Langmead & Salzberg, 2012). Though nine tools were initially considered for comparative analyses, two programs were unable to be tested: V-Phaser2, and VirVarSeq which are briefly discussed below.

- V-Phaser2 (Yang et al., 2013): Unable to utilise due to dependency issues [libbamtools2.5.2].

- VirVarSeq: (Verbist et al., 2014): Intermediary dependant R package [rmgt] is no longer supported and hence could not be tested.

5.2.2.1 DiversiTools (Hughes, 2016)

Many existing tools can determine the frequency of mutations from deep-sequencing data, but most have been developed for diploid genomes. DiversiTools, written in Perl, focuses on determining mutations in haploid genomes. Specifically designed for analysing viral deep sequence data, it simply reports the counts of all bases and indels at all genome positions. It runs on a user provided BAM file and outputs the data in text tab delimited format.
Tool available at: http://josephhughes.github.io/DiversiTools

5.2.2.2 DeepSNV v1.42.1 (Gerstung et al., 2012, 2014)

DeepSNV is a targeted deep-sequencing approach combined with a custom statistical algorithm for detecting and quantifying sub-consensus SNVs in mixed populations. Utilising a probabilistic method, DeepSNV incorporates knowledge about the distribution of variants in terms of a prior probability. The authors present a novel approach for calling mutations from large cohorts of deep-sequenced cancer genes. Their work claims to be capable of detecting variants at proportions as low as 0.0001%.

5.2.2.3 FreeBayes (Garrison & Marth, 2012)

FreeBayes is designed to find SNPs using a Bayesian statistical framework to model multiallelic loci in sequences with non-uniform copy numbers. It uses short-read alignments plus a reference genome to determine the most-likely combination of genotypes for the population at each position in the reference and then reports putative polymorphic positions.
Tool available at: https://github.com/freebayes/freebayes

5.2.2.4 iVAR v1.4.2 (Grubaugh et al., 2019)

iVar is a generic tool that can be used for calling variants, determining consensus sequences and trimming primers off amplicon sequences. iVar is written in the C++ programming language and processes the output of the mpileup function of samtools to subsequently call observed variants from a BAM file. iVar is not a very sophisticated variant caller and relies on user defined thresholds for minimum read depth, minimum base quality and minimum variant frequency. Results are outputted in text tab format.
Tool available at: https://andersen-lab.github.io/ivar/html/manualpage.html

5.2.2.5 LoFreq v2.1.5 (Wilm et al., 2012)

LoFreq is a fast and sensitive variant-caller for inferring SNVs from next-generation sequencing data. Sensitivity is derived from the tool's processing; each variant call is assigned a p-value, allowing for rigorous false-positive controls. LoFreq is generic and fast enough to be applied to high-coverage data and large genomes. LoFreq is written in C and Python, runs on a user provided BAM file and reference file, and outputs results in the VCF format. LoFreq has a number of in-built filters that filter variants for strand bias, depth and snv-quality.
Tool available at: http://csb5.github.io/lofreq/

5.2.2.6 Varscan2 (Koboldt et al., 2012)

VarScan2 was developed as a platform-agnostic caller for the detection of mutations in the genomes of tumour-normal pairs. The algorithm reads data from both samples simultaneously; a heuristic and statistical algorithm detects sequence variants based on user defined thresholds in coverage, read quality and variant frequency. VarScan2 is written in the Java programming language and runs on the output from the samtools mpileup command (which itself runs on a BAM file), and creates a tab delimited text file of variants.
Tool available at: http://varscan.sourceforge.net/

5.2.2.7 vSensus

VSensus is similar to DiversiTools and simply reports the observed counts of all bases and indels at all genome positions and then creates a consensus sequence. It is written in the Java programming language, runs on a user provided BAM file and outputs results in a text tab format. VSensus does not have any filter checks for strand-bias etc, but the user can apply base quality filters.
Tool available at: https://github.com/rjorton/VSensus

Table 5.2: Default parameters of each variant calling tool

Tool	Min. Base Quality	Min. Mapping Quality	Min. Read Coverage	Min. Variant Frequency
DeepSNV	25	0	100	-
DiversiTools	-	-	-	-
FreeBayes	0	1	0	-
iVar	20	-	0	3%
LoFreq	6	0	1	-
VarScan	15	-	2	1%
vSensus	0	0	0	-

5.2.3 Accuracy and Hamming Distance

Simply put, p_i is the proportion of non-consensus reads at site n in the control dataset. The difference in variant reads at the corresponding site in my experimental dataset (q_i) is then calculated to give the Hamming Distance between the two datasets at a single site (L_i). Averaged across all sites in the genome, each sample then has a distance value between the control and one of seven experimental datasets.

$$L = \sum_{i=1}^{n} |p_i - q_i|$$

If the caller did call a variant that also was recorded in the published dataset, then how much did the frequency differ between the original call and each of the caller tools? This measure of difference, the Hamming Distance, can show how closely a reported mutation mirrors a mutation detected by other tools/researchers.

5.3 Results

5.3.1 Comparative Analysis of Variant Call Tools

Testing measured seven variant calling tools (VCT) using three main metrics: 1) computational demand in seconds, as given by either the unix 'time' command or, for DeepSNV, the R library 'microbenchmark' 2) accuracy & precision and 3) resulting viral population characteristics. This three-fold analysis touches on some of the main factors involved in deciding which tool to select for low-frequency variant calls. A processing pipeline is presented in Figure 5.3, representing the source files required and output by each VCT. Ultimately, all of the output files were transformed into csv (comma-separated value) files so results could be compared. All tested VCT require BAM inputs, though some (deepSNV, VarScan and vSensus) need additional prior processing before to calling variants. Additionally, all of the tools except deepSNV could process the multi-segmented influenza A genome under the assumption that each genomic segment was analogous to independent chromosomes. Due to the unique necessity of an averaged reference BAM needed for deepSNV processing, this then required variants to be called separately for each genomic segment.

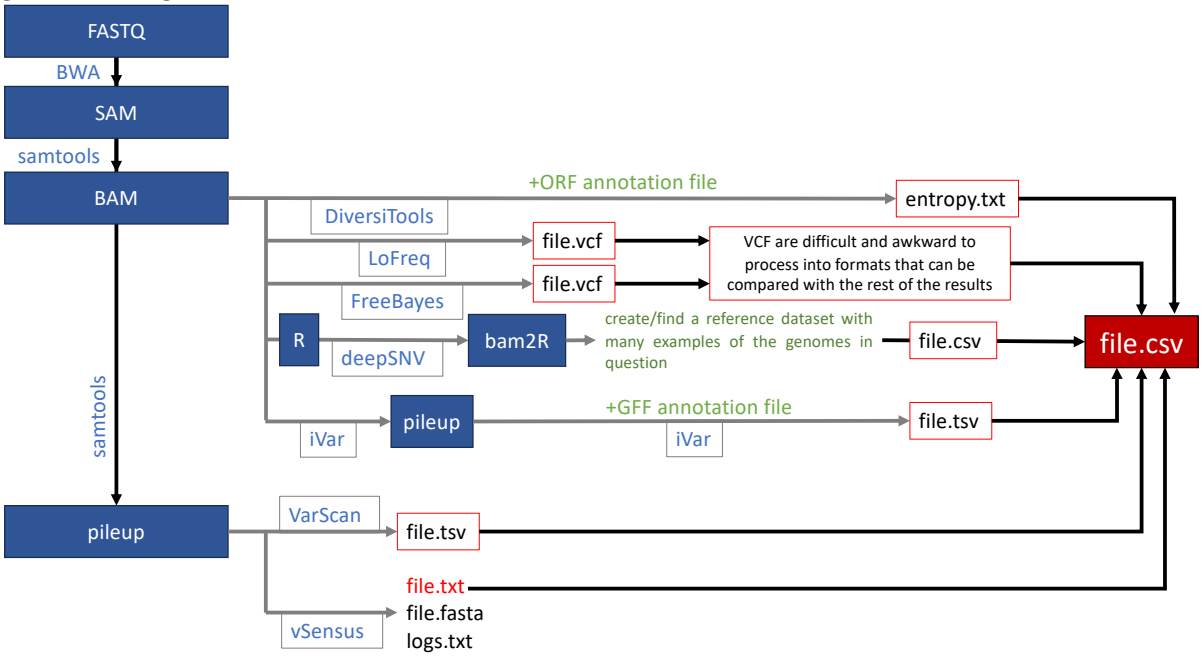


Figure 5.1: Processing pipeline of variant call tools, and the steps involved in obtaining a final output, in a widely-used format that can be compared across tools, i.e. a csv file.

5.3.1.1 Computational Demand

Overall, the tools DeepSNV, VarScan and vSensus perform fastest for the processing of each dataset (Figure 5.4). Diversitools and LoFreq rather are the slowest

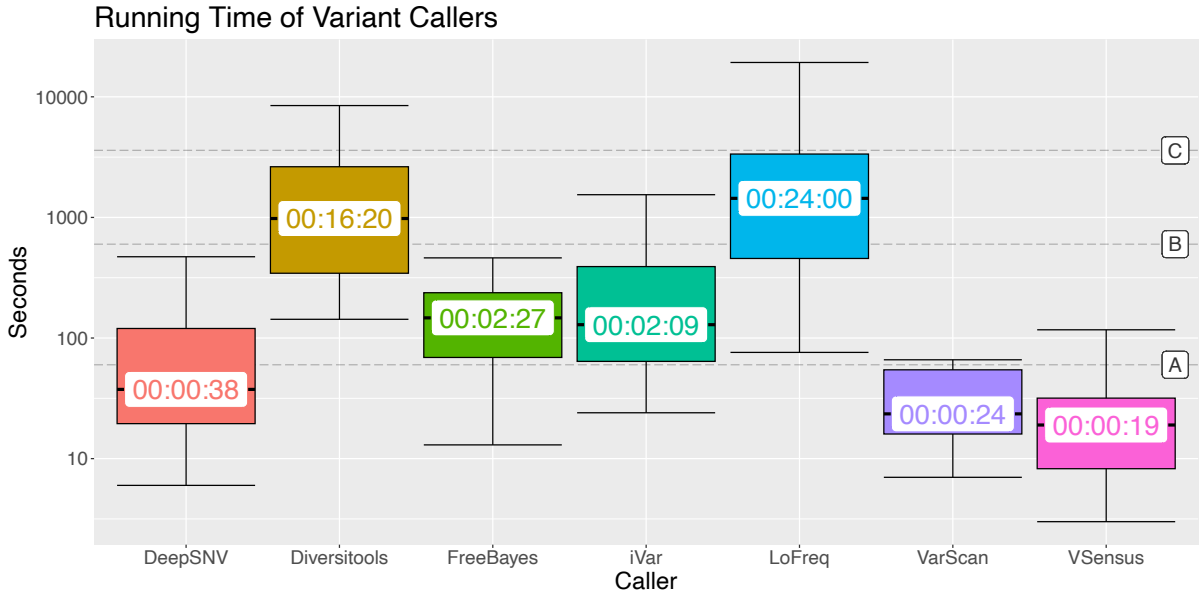


Figure 5.2: Processing time of each variant call tool, averaged for all five datasets. Median times are labelled and dashed lines are added at 1 minute (A), 10 minutes (B) and 1 hour (C).

processes, sometimes reaching a scale of hours per sample. Importantly, this measures only the actual running time of the tool, and does not include data pre−processing steps.

A caveat of these time-trials is that the model genome used, Influenza A Virus, has a relatively small genome compared to some other RNA viruses. Additionally, the eight-segmented genome of IAV often artificially inflates processing time/resources required that may not be seen in analysis of other, non-segmented viral genomes. Not all bioinformatic tools are designed to handle segmented genomes or proteins encoded by frame-shifts and splicing. Finally, even if processing the sample can be relatively fast the preamble and necessary set-up can be significantly disproportionate to the running time reported here. Despite its' fast run-time and comprehensive output, deepSNV requires a great deal of pre-processing before analyses and additionally has numerous other packages on which it depends; as a consequence of its R language and utilisation of graphical outputs.

5.3.1.2 Accuracy

In order to quantify how successfully each tool was able to report LFVs, the location and frequency of variants from tool outputs were compared against the originally published variant calls, or the expected variants in the case of the simulated data set.

Each base position of the 38 selected datasets was declared as either *variant* or *non-variant* in the reference dataset (True/False) and the outputs of my experiments (Positive/Negative); overall giving 38 matrices with eight rows (reference dataset plus outputs of the seven tested tools) and a column for each base position of the IAV genome. These classic confusion matrices were then populated with a binary pass/fail for each cell in which a variant was reported.

A)

Sample 1	Position	1	2	3	4	n...
	Reference	FALSE	FALSE	TRUE	FALSE	TRUE
Variant Call Tool	DeepSNV	-	+	+	-	+
	Diversitools	-	-	+	-	+
	FreeBayes	-	-	+	+	-
	iVar	-	-	+	-	+
	LoFreq	-	-	+	-	+
	VarScan	-	+	-	-	+
	vSensus	-	-	+	-	+

B)

Sample 1	Position	1	2	3	4	n...
Variant Call Tool	DeepSNV	TN	FP	TP	TN	TP
	Diversitools	TN	TN	TP	TN	TP
	FreeBayes	TN	TN	TP	FP	FN
	iVar	TN	TN	TP	TN	TP
	LoFreq	TN	TN	TP	TN	TP
	VarScan	TN	FP	FN	TN	TP
	vSensus	TN	TN	TP	TN	TP

TRUE-Positive	FALSE-Positive
TRUE-Negative	FALSE-Negative

C)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP} \qquad Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

Figure 5.3: In comparing VCT, variants at each position of the sample were marked in a simple presence/absence matrix (A) which could then be compared with the Reference. The Reference dataset was compiled from both variant sites introduced into the dataset at known locations by the original authors, and the results of variant calls within those original publications. B) For any given site, the Reference is used as a gold standard against the output of my testing; these match-values were then used in calculating confusion matrices to measure accuracy, using the calculations presented in C.

Confusion matrices were produced (using the R package caret) by comparing the proportions of variants found by each tool to those published by the original authors as the 'real' observations (Figure 5.5). To note, the published results are taken as correct, the published datasets use one or more variant call procedures, often with tools other than the ones tested here. The emphasis therefore is on how repeatable these variant calls are, rather than absolute verifiability. Each row of these matrices then gave a number of match-values (true-positive, true-negative, false-positive or false-negative) equal to the number of nucleotides in each genome.

To note, each site is limited to a binary value: it matches the reference or it doesn't match the reference. Finally, every genome sample had one set of match-values for each of the seven tools. Our results in Figure 5.6 mark the ability of each tool to find SNV, based on four common measurements used in machine learning.

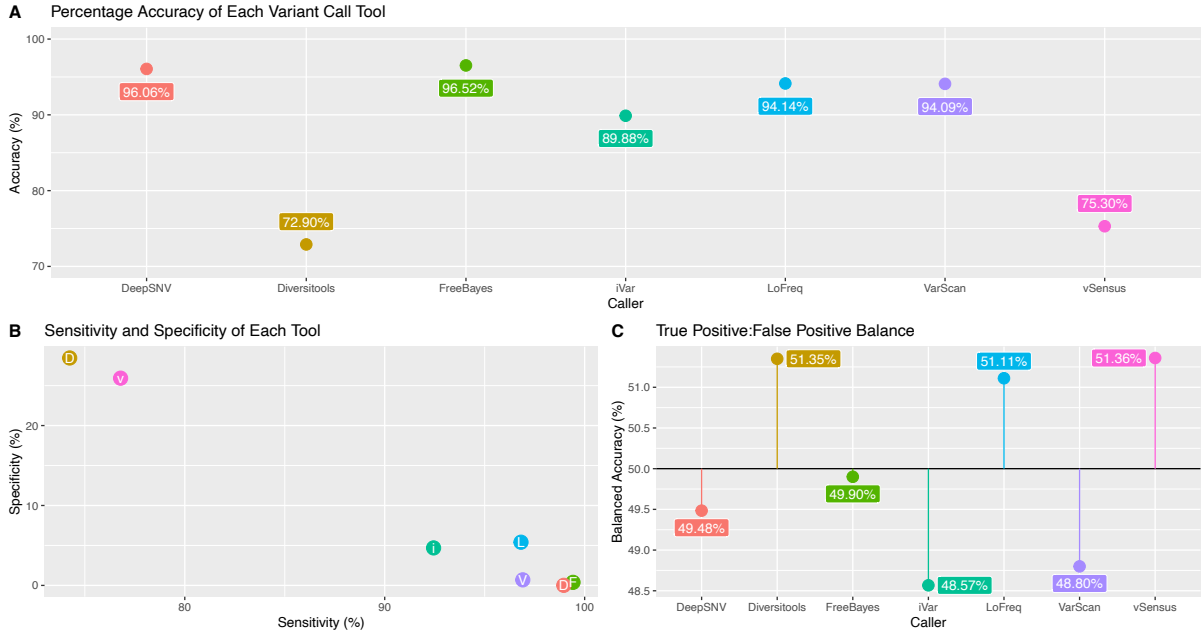


Figure 5.4: Four measures of performance for each tool trialled: A) Accuracy, B) Sensitivity (True-Positive Ratio) & Specificity (True-Negative Ratio) of each tool averaged across all five datasets are represented by dots, coloured and labelled with abbreviated names of each tool. C) The ratio of True-Positives to False-Positives gives the balance likelihood.

To review these metrics, accuracy (Figure 5.6A) measures the proportion of reads at each site that match the reference sequence, sensitivity is the percentage of true-positives over the total number of correct calls $\left(\frac{TP}{TP+FN}\right)$, mirrored by the specificity which shows the proportion of incorrect calls $\frac{TN}{TN+FP}$ (Figure 5.6B). With default settings, both Diversitools and vSensus lag behind with lower accuracy (72.90% and 75.30% respectively) and sensitivity (74.24% and 76.79% respectively) than the other processes; this is likely due to the creation of many false-positives from the lack of filtering in both of these programs. Additionally, the specificity of each tool may appear artificially low because during the creation of BAM files, particularly egregious negatives are purged from the dataset due to low base-quality, prior to processing with a VCT.

Using the confusion matrices provides simple, but helpful, insight. Quantifying the accuracy of each tools' output grants a levelled base-line from which to compare and rank them. Importantly, it must be noted that this quantitative scoring does not give a definitive answer; different tools may suit certain purposes better than others. A clinical variant-call pipeline likely values accuracy above all else, perhaps favouring the use of VarScan. Environmental viral sampling, however, may be able to sacrifice fine accuracy for more faster processing times.

5.3.1.3 Diversity Metrics

Using well-established methods of measuring genomic diversity (Fuhrmann et al., 2021; Gregori et al., 2016), our final comparisons denote the population diversity of samples as determined by the variants each tool reported. These diversity values are then compared against those calculated using the originally published set of low-frequency variants to show differences, if any, in viral population analyses. Observing the population diversity of LFV detected by each tool

in the tested datasets will allow for comparison of the resulting outputs of each tool. Should the output from one VCT show radically different population diversity to that calculated from the outputs of the other tools, it suggests that variants are being missed or created spuriously. Theoretically, all VCT should produce the same output, i.e. the detection of identical locations and frequencies of variant sites throughout the genome, and so this homogeneity is . Thus, calculating the diversity of these datasets enabled identification of spurious, less accurate VCT procedures.

5.3.1.3.1 Richness

First, the simplest diversity metric, richness, was calculated for each genomic segment of the control sequence libraries. The richness of genomic variants represents how many nucleotide positions show evidence of LFV (*V*), defined as a site with <99% read identity (i.e. a variant present at a frequency of at least 1%). As IAV has eight genomic segments (*n=8*), the product of richness values for each segment is given: $\prod_{n=8} \frac{V}{N}$ where *N* is the total number of bases in the genome. Thus, Figure 5.7 shows violin plots for the richness estimates of each control dataset, each comprised of the diversity of each genomic segment calculated from the output of each VCT; SNV richness of eight genomic segments, as calculated from the results of seven different tools meaning that each violin (except project 412631 which looks exclusively at genomic segment 4 (HA)) has 56 unique points.

Due to the specific cut-off points and processes for defining what makes a variant, VCT show markedly different numbers when analysing the same dataset. Values from the SimData sequences are the most polarised; mutations were spiked into the genome at known locations and frequencies. This is of course because of the repeated nature of mutations. To note, segments are not labelled as differences in the diversity of each gene are not the focus of this tool comparison step.

Conservative VCT like FreeBayes, DeepSNV and LoFreq all have a maximum number of variant sites below 20%. The stringent rulings used by these tools to classify *variant* or *non-variant* result in only the most egregious variants to be reported. Alternatively, tools like Diversitools, iVar and vSensus simply report the proportion of mismatched bases at each site, leaving the filtering and classification of *variant/non-variant* to the user in post-processing. Because of the lack of filtering in vSensus and Diversitools, the richness calculated from data produced by these tools is generally very high, making them unsuitable for the in-depth analyses I intended to pursue. I sought to focus on clearly distinguishable dynamics of LFV, and so the background noise generated by these non-specific tools would have hindered further analyses.

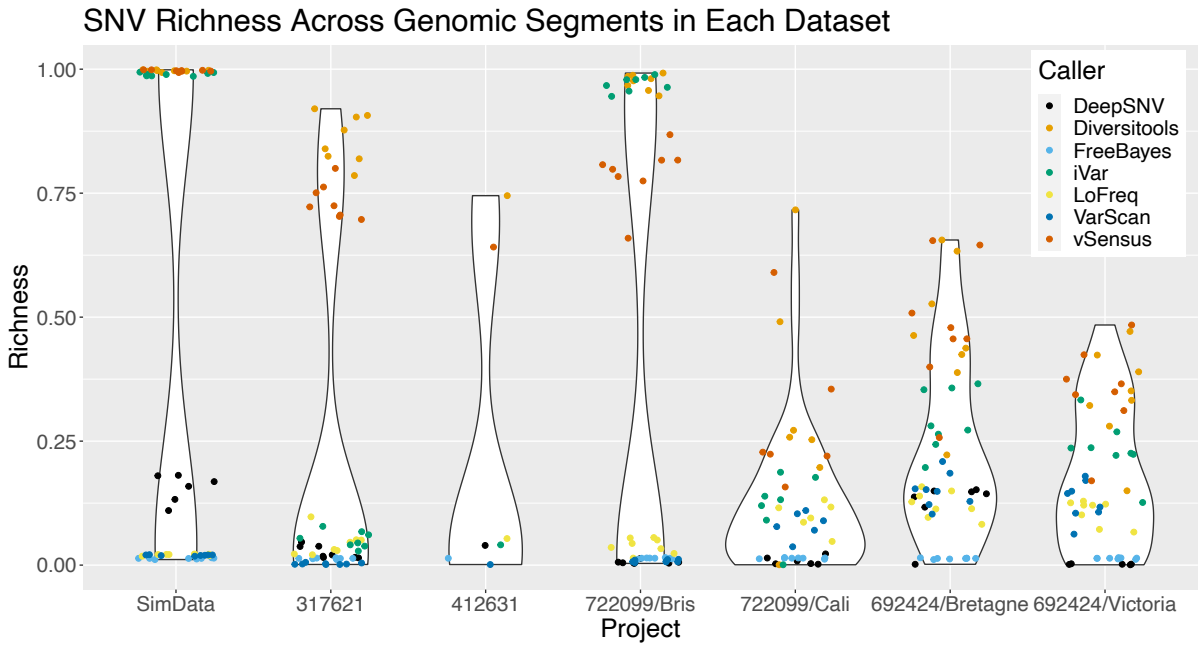


Figure 5.5: Violins show the proportion of nucleotides showing some evidence of variation, each point representing the average richness of a genomic segment using a specified tool (colour) and the dataset sequences from which it originated (x-axis). Each point represents the richness for that genomic segment averaged across that dataset. Datasets created with controlled populations of viruses (SimData, 317621 and 412631) tend towards biphasic violins, with the population richness calculated either very high or very low, depending on the VCT used.

5.3.1.3.2 Shannon Entropy

Sub-consensus population diversity depends largely on the tool used to find variant nucleotides. Figure 5.8 shows entropy of genomic segments calculated from the variant proportions obtained from each VCT. Diversity between segments is small yet significant ($p < 2.2e^{-16}$, Kruskal-Wallis test).

Figure 5.6: Shannon Entropy as calculated from the variant frequencies obtained by each of the seven tools. As above, project 412631 contains only results from segment 4 (HA).

The different proportion and frequency of variants detected by each VCT, however, results in substantially different entropy values based solely on the chosen VCT ($p < 2.2e^{-16}$, Kruskal-Wallis test). Much like the results from the richness of LFV

in the control datasets, tools without internal filtering (Diversitools and iVar) tend give overestimations of population diversity. Here, calculating diversity from the variants called by FreeBayes shows that within each dataset, the eight genomic segments show very similar values of Shannon Entropy.

5.3.1.3.3 Distance Measures

The frequency at which each mutation was originally reported is compared against the frequency detected by our VCT trials. Measuring disparities in the frequencies of variants that were reported in both the original publication and these trials clarifies just how much of an impact the chosen VCT may have on further analyses of viral populations.

Knowing how closely the variant proportion reported by a VCT matches that recorded in the control dataset goes further than the binary true-positive/false-positive accuracy used above in confusion matrices. Figure 5.9 presents each variant found in both my analysis and the published dataset, then compares the frequency that variant is reported at. The addition of a straight line representing an absolute positive correlation helps distinguish when the frequency of mutations are over-estimated (below the line) or under-estimated (above the line). Consistency is key for bioinformatic tools. Using tools at their default settings, great variation is seen between the results obtained through my experiments and those originally published. Diversitools, iVar and VSensus have very weak correlations; these tools report the proportion of variants at every site in the genome, without any real filtering or recognition of error.

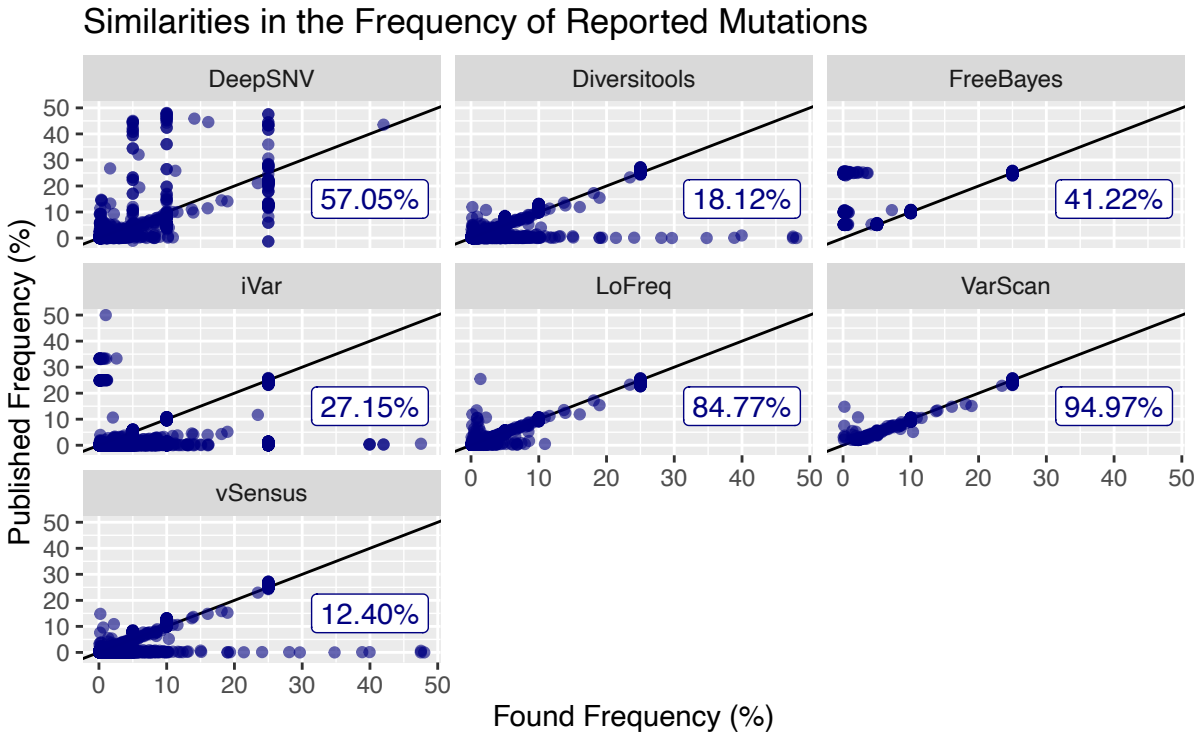


Figure 5.7: Each point shows the frequency of a mutation that was reported in both the original dataset (y-axis) and in my experimental replicates (x-axis). A line is added to show what would be expected if the mutations were found at exactly the same frequency in both datasets, i.e. a perfect correlation. If a point deviates towards the x-axis, it was detected by my protocol at higher frequencies than in the original work. Conversely, a point closer to the y-axis is at a lower frequency in my dataset compared to the original publication. Spearman correlations between the abundance of mutant genomes are annotated on each graph.

VCT that employ more stringent filtering algorithms that can account for sequencing error/bias, such as LoFreq, consequently have much higher congruity in the proportion of genomes displaying variant alleles between the datasets. This time VarScan appears as the frontrunner, with very high correlation between the frequencies of variants in the expected (control) and observed (test) datasets. However, note the lack of points below the dividing line for this plot – there were few, if any, false-negatives reported with this tool. That is to say, all mismatches between the expected and observed datasets were due to variants detected during my testing that were not apparent in the published dataset. This is reflected in the tests above (5.8), the ratio of true-positives to false-positives in VarScan is an outlier. Hence, the next most congruous tool was LoFreq.

5.3.2 Observed Variants

The above conclusions indicated LoFreq as the best tool of choice for analysing sub-consensus patterns and behaviour in the small, high-depth dataset obtained from our EIV transmission experiment. Auxiliary results from FreeBayes were cross-examined against those from LoFreq, to ensure no low-frequency variants found by FreeBayes, the more exact and stricter of the two, were missed by LoFreq. After this cross-referencing procedure, all analyses were carried out using the LoFreq output. Hence, a library of LFV was generated from each of the genomic samples collected by nasal swabs of infected horses, totalling 53 lists of LFV.

5.3.2.1 Tracking the Trajectory of Mutations Before Broaching Consensus Level

21 mutations were recorded at the consensus level throughout the transmission experiment (Chapter 4.3.1); 11 non-synonymous and 10 synonymous. Exploration of this diversity began at these 21 sites in order to observe any potential sub-consensus dynamics that may be underlying emergence at the consensus level. It was also helpful in deducing whether these mutations appeared *de novo* or existed at low levels in the population before being enriched enough to appear in genomes at the consensus, population level. The proportion of reads at each base position showed patterns of diversification, enrichment or removal. As the consensus nucleotide is defined here as >50% reads, nucleotides that appeared in 10-50% of reads are declared low-frequency variants (LFV). Nucleotides present at a proportion of 1-10% reads are rare LFV, 0.1-1% are very rare LFV and anything below 0.1% is too low to reliably distinguish from sequencing error. This frequency threshold was manually filtered from the collated outputs of each tool.

Two consensus-level nonsynonymous mutations in segment 1 (PB2-t979c/Gly327Arg and PB2-g2191a/Val731Ile) are transient and appear only once with very minimal presence outside of these samples. These mutations are seen in two separate hosts of the Single transmission group. The two other mutations seen at the consensus level in this segment have more fluctuation in sub-consensus proportions and are both synonymous. Mutation PB2-c1497t (Asp499) only appears at the consensus level once; however, in the days around this sample, 1-10% of viral genomes also show this mutation (Figure 5.10). Notably, it is also seen across three different vaccinated hosts of the multistrain transmission group: 2A, 2B and 3A. Additionally, a host in the other transmission chain (singlestrain_2B) shows a proportion of 9.28% genomes with this mutation, but only at one timepoint. The other consensus-level mutation in this genome segment, PB2-c1779t (Ser593), also appears in a single sample from a vaccinated host (singlestrain_4A).

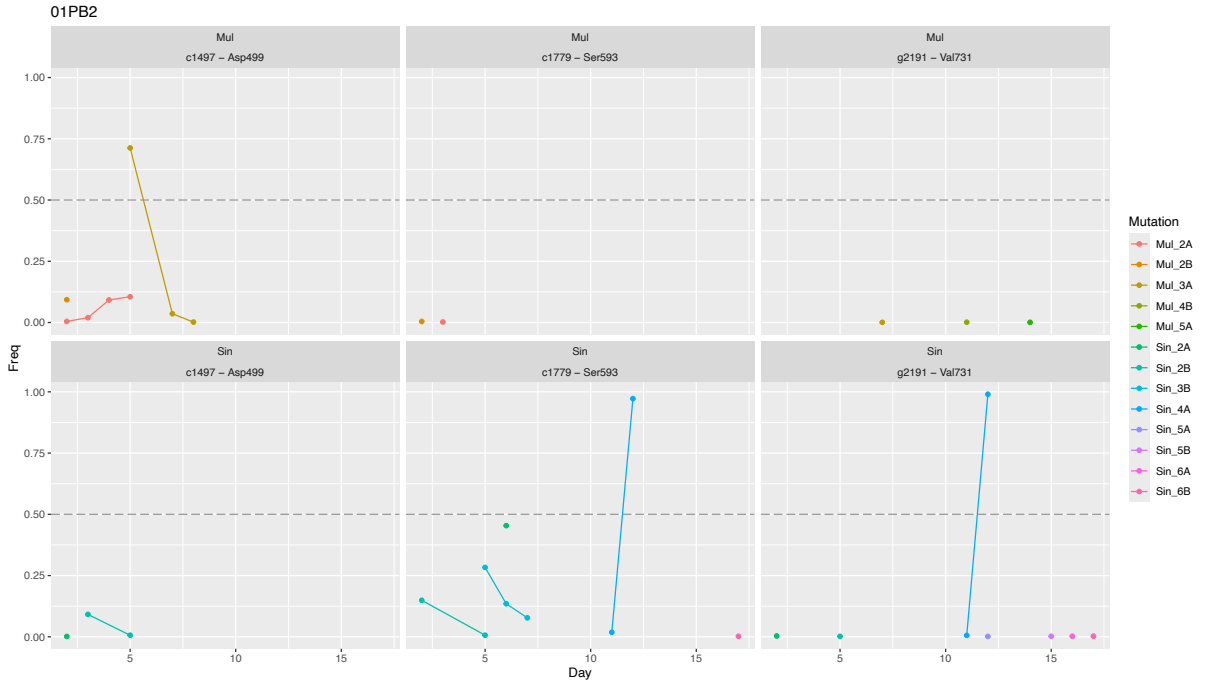


Figure 5.8: The proportion of reads displaying a mutation, the corresponding mutation is coloured and the host from which the sample was taken is annotated on the point. Similar graphs for each genomic segment are found in Supplemental 5.3

In segment two (PB1), one of the three consensus mutations, PB1-a881g/Gln294Arg, is transient with little change. The synonymous substitution PB1-t1500c (Gly500) however, fluctuates lots before fixation in the multistrain transmission group. Once in naïve hosts, cytosine is reported in >99.8% of genomes – demonstrably fixed in this population. However, the mutation is also shown in the three preceding transmission pairs, often enriched to high proportions. Indeed, it broaches the consensus level twice, in individual 3A and is maintained in 1-30% of viral genomes in these hosts as well as hosts 2A and 3B. Within host 3A of the multistrain group, reads displaying the mutant cytosine were present on three consecutive days: day nine of the experiment (82.58%), day ten (3.30%) and day 11 (92.77%). This wild variation is difficult to understand; it is unknown if this synonymous mutation has any effect on viral fitness or whether the patterns seen were caused by unrepresentative sampling on day ten. The mutation appears to arise *de novo* in host 2A on day 5 of the experiment, genomes displaying this mutation are then more frequent (from 4.09% to 14.63% of total genomes show PB1-t1500c) on the subsequent day. Though genomes with PB1-t1500c repeatedly fall below a frequency of 1% of viral genomes, the mutation persists with each transmission event until eventually being passed to naïve hosts in which cytosine becomes the dominant nucleotide at this site. In the singlestrain group, $\frac{21}{23}$ samples also present cytosine at this site in 0.1-0.25% of genomes, causing a nonsynonymous substitution. PB1-a1853g/Glu618Gly displays very low levels of variation (0.1-1% of genomes show the guanine mutation) before suddenly fixing in the multistrain naïve population. To note, 0.1-0.25% of genomes from each sample of the singlestrain transmission group also display the same mutation, suggesting either low level maintenance and/or a proclivity for variation at this site.

Segment three (PA) has five consensus-level mutations. PA-c825t and PA-g1180a/Asp394Asn, the only nonsynonymous mutation recorded in this genomic segment, are transient. Also on this genomic segment, PA-c201t shows lots of

interesting variation, as shown in the heatmap of Figure 5.10. Remembering that this mutation appeared in many of the global EIV sequences, putative phenotypic effects of PA-c201t ought to be investigated. The *de novo* appearance, maintenance and transmission of this variant could potentially indicate a beneficial adaptation for these viruses. The PA-t1221c mutation is maintained at low levels (0.1-0.3%) in almost all samples across both transmission experiments. The most variation is seen in Mul_6B, where after narrowly broaching consensus (50.36%) on day 16 of the experiment, genomes presenting this substitution become much rarer: 1.10% on day 17, 3.09% on day 18 and 0.14% on day 20. This may suggest a deleterious effect from this mutation, otherwise this strain is stochastically removed from the viral population within this host. PA-a1650g shows minimal enrichment the day before (2.43%) and after (2.61%) the peak (75.81%) in a single host.

As explored in previous chapters, two non-synonymous mutations in segment four lead to amino acid substitutions in haemagglutinin. HA-a431c/Gly144Asp appears *de novo* and is then never reported in frequencies higher than 0.2%, hence my assumption that this is merely a transient substitution. Appearing at consensus level in vaccinated hosts of both transmission groups, HA-a1401c/Arg467Ser is also observed in multiple samples at 1-10% of reads, all of which happen to be in vaccinated hosts across both transmission groups. Maintenance at low levels in the population indicates that the mutation is at least non-lethal for the virus. Indeed, referring back to *in silico* experiments of protein antigenicity, this mutation is estimated to increase the recognition of the surrounding protein structure by cells of the host adaptive immune system. It does not, however, persist to the naïve hosts and, in fact, is limited to hosts 2A and 2B in the Single transmission group.

Segment five's non-synonymous NP-g1445a (Ser482Asn) mutation defines the consensus F genome. In all naïve hosts of the singlestrain transmission group, >95% of the reads report adenine at this site. However, the opposite cannot be said for the vaccinated individuals; three of the 12 samples from vaccinated hosts display guanine below 85%. Vaccinated host 3B shows three consecutive days of sub-consensus enrichment of the guanine-adenine mutation: 43.49% adenine on day seven of the experiment, 1.28% on day eight and finally 15.07% on day nine. Before

03PA c201t, Asp 67

	Base			
	Adenine	Cytosine	Guanine	Thymine
Mul_2A_Vacc_25_Jul	0.03%	54.06%	0.01%	45.90%
Mul_2B_Vacc_25_Jul	0.01%	83.05%		16.94%
Mul_2A_Vacc_26_Jul	0.01%	66.82%		33.17%
Mul_2A_Vacc_27_Jul		78.94%		21.06%
Mul_2A_Vacc_28_Jul		94.70%		5.30%
Mul_3A_Vacc_28_Jul		99.87%		0.13%
Mul_3A_Vacc_30_Jul		99.94%	0.01%	0.05%
Mul_3B_Vacc_30_Jul	0.04%	31.20%		68.76%
Mul_3A_Vacc_31_Jul	0.01%	99.92%	0.01%	0.06%
Mul_3B_Vacc_31_Jul		37.44%		62.56%
Mul_4A_Vacc_31_Jul	0.01%	99.92%	0.01%	0.05%
Mul_3A_Vacc_01_Aug		99.87%	0.03%	0.10%
Mul_4B_Vacc_03_Aug		99.96%	0.01%	0.03%
Mul_5A_Naive_03_Aug		99.96%	0.01%	0.03%
Mul_5A_Naive_04_Aug		99.97%	0.03%	
Mul_5A_Naive_05_Aug		99.94%	0.01%	0.05%
Mul_5B_Naive_05_Aug		99.97%	0.01%	0.01%
Mul_5A_Naive_06_Aug		99.96%		0.04%
Mul_5B_Naive_06_Aug	0.03%	99.84%		0.14%
Mul_6B_Naive_06_Aug		99.97%		0.03%
Mul_5A_Naive_07_Aug		99.97%		0.03%
Mul_5B_Naive_07_Aug	0.03%	99.92%	0.01%	0.04%
Mul_6A_Naive_07_Aug	0.04%	99.88%		0.08%
Mul_6B_Naive_07_Aug		99.96%	0.01%	0.03%
Mul_6A_Naive_08_Aug	0.03%	99.89%	0.03%	0.05%
Mul_6B_Naive_08_Aug	0.03%	99.90%		0.08%
Mul_6A_Naive_09_Aug		99.99%		0.01%
Mul_6A_Naive_10_Aug	0.01%	99.90%	0.01%	0.07%
Mul_6B_Naive_10_Aug	0.01%	99.96%		0.03%

03PA c201t, Asp 67

	Base			
	Adenine	Cytosine	Guanine	Thymine
Sin_2A_Vacc_16_Oct	0.03%	85.84%		14.13%
Sin_2B_Vacc_16_Oct	0.01%	95.24%		4.75%
Sin_2B_Vacc_17_Oct	0.01%	95.15%	0.01%	4.83%
Sin_2B_Vacc_19_Oct		85.63%		14.37%
Sin_3A_Vacc_19_Oct	0.04%	0.40%		99.57%
Sin_3B_Vacc_19_Oct	0.05%	89.50%	0.01%	10.44%
Sin_2A_Vacc_20_Oct	0.01%	99.85%	0.04%	0.10%
Sin_3B_Vacc_20_Oct	0.01%	91.28%		8.70%
Sin_3B_Vacc_21_Oct		36.24%		63.76%
Sin_4A_Vacc_25_Oct	0.03%	0.49%	0.01%	99.47%
Sin_4A_Vacc_26_Oct	0.01%	0.34%		99.65%
Sin_5A_Naive_26_Oct		99.75%		0.25%
Sin_5B_Naive_26_Oct		99.94%		0.06%
Sin_5A_Naive_27_Oct		99.94%	0.01%	0.05%
Sin_5B_Naive_27_Oct		99.90%		0.10%
Sin_5B_Naive_29_Oct		99.96%		0.04%
Sin_6A_Naive_29_Oct	0.01%	99.93%	0.03%	0.03%
Sin_5A_Naive_30_Oct		97.84%	0.01%	2.15%
Sin_6A_Naive_30_Oct	0.01%	99.88%	0.01%	0.09%
Sin_6B_Naive_30_Oct	0.03%	99.86%		0.12%
Sin_6A_Naive_31_Oct		100.00%		
Sin_6B_Naive_31_Oct	0.01%	98.84%	0.01%	1.13%
Sin_6B_Naive_02_Nov		99.95%	0.01%	0.04%

Figure 5.9: Proportion of reads reporting a nucleotide at position 201 in genome segment 3. Cells are coloured according to frequency: red are between 0.1-1%, orange 1-10%, yellow 10-50% and green signifies consensus (>50%).

broaching the consensus level, mutation g1445a appears at very high proportions, falls back to almost being removed from the population then re-establishes itself. Only two sequences were recoverable from samples of the next hosts in the transmission chain (hosts 4A and 4B), both showing this mutation present at less than 0.3% indicating that these variant genomes may not have survived being transmitted. I must conclude then that either the mutation was passed to hosts 4A and 4B but viral loads were so low that genomes could not be sequenced or alternatively, the mutation died out in host 3B and spontaneously appeared *de novo* in the naïve hosts. This however would require the mutation to evolve in two naïve hosts (5A and 5B) on the same day, the first day that samples from these hosts could be sequenced.

The only synonymous mutation reported at the consensus level in segment six (NA-c690t) appears transiently in one individual; other samples from this individual and from both transmission experiments report a proportion of thymine reads at <0.1%. For this reason, I believe that this mutation is a random, transient occurrence. The two other mutations are also singletons. The first, NA-a1024g/Lys342Glu appears rarely in the preceding individual (Mul_2A presents guanine in 2.66% of genomes on day 7 of the experiment, Mul_3A 81.68% and 2.24% on days 10 and 11 respectively and host Mul_3B also presents guanine at a proportion of 1.09% on day 10). Finally, NA-t1385c/Ile462Thr is present in almost every sample at very low (0.1-1%) proportions, but appears in four of the five pairs in the Single group (1.4% in 3B on day 8, 96.37% in 4A on day 14, then both 5A:1.85% and 6A:1.54% on day 18 of the experiment).

Only one consensus-level mutation was observed in segment 7. M1-a418g broaches consensus level in Single_3B eight days after the beginning of the experiment (77.01%). However, the mutation is evident before and after appearing in the consensus sequence; the preceding day (day 7) 3.89% of reads showed a guanine mutation and the day after spiking (day 9) guanine is present at site 418 in 4.91% of reads. Unfortunately, these three days are the only sequences collected for individual Single_3B so tracking this transient, non-synonymous mutation beyond this one-day spike is impossible. Referring back to the consensus analysis of this mutation (Ch. 4 11.7), this amino acid substitution is predicted to have minimal effects on the twist angles of the structure. However, Lys and Glu have similar chemical properties.

In segment 8, both consensus mutations (NS-t84c and NS-t87c) only appear at levels >50% in individual Single_2B on a single day. However, the mutations are found at low levels in vaccinated hosts of both transmission chains over multiple days. That the same synonymous mutation appears at low levels in six vaccinated hosts at the beginning of each transmission chain (Single_2A, Single_2B, Multi_2A, Multi_2B, Multi_3A and Multi_3B) indicates some potential for neutral evolution. The mutation persists at proportions 1-15% of the viral population for five days in each transmission chain with no clear pattern of being enriched or purged.

Ultimately, mutations that appear in consensus sequences show a range of activity below the consensus level. Some variants gradually build in frequency before defining the consensus sequence while others remain at low proportions among the viruses, occasionally dominating the population as a result of founder effects or stochastic shifts in population composition. As seen in the consensus genomes, much of the diversity is generated within hosts with a history of EIV exposure (the "vaccinated" class) then viral genomes homogenise and sub-consensus mutations are either removed or forced down to very small proportions of the population. Overall, the noisiness of the data makes any inference challenging.

5.3.3 Sub-consensus Genetic Diversity

5.3.3.1 Abundance Indices

A preliminary count of the number of variants reported by LoFreq (which filters variants for strand bias and low quality), and how often they appear throughout the genome can give a top-down overview of the composition of viral genomes in the population. These counts and associated frequency metrics are reported in Table 3.

5.3.3.1.1 Abundance

Across the 13kb genome of all 8 genomic segments, the four epidemiological groups (Naïve in the multi group [N_M], Vaccinated in the multi group [V_M], Naïve in the single group [N_S] and finally Vaccinated in the single group [V_S]) showed similar numbers of sub-consensus variants. Roughly half of the positions across the genome showed evidence of variation above a 1% threshold shown as the Proportion in Table 12; this varied marginally between observation groups. The mean frequency of mutations, however, did show clear demarcation between transmission groups. Both vaccinated and naïve hosts in the multi transmission chain experienced mutations more frequently than those in the single group (average of $6.20e^{-3}$ compared to a frequency of $5.48e^{-3}$ in the single group). Differences in LFV frequencies between transmission groups (using Wilcoxon Rank tests), were significant, though minor. The heavy concentration of LFV that appear only at very low frequencies (<10%) likely confounds this comparison however.

5.3.3.1.2 Richness

Observing richness values, we can consider the number of polymorphisms per kilobase of genome (Table 5.3). To note, this can depend heavily on the read depth. Hence, the expected number of polymorphisms per kilobase is also dependent on the coverage and comparing this value between samples with different read depth distributions may be misleading. In our experiment, the richness of variants was similarly low between vaccinates in both transmission groups. More mutations per kb were seen in naïve hosts, especially in the multi group (N_M = 1.36, N_S = 1.05). This isn't a huge difference but indicates a lower persistence of sub-consensus mutations in vaccinated hosts than in naïve ones. As mentioned above though, the coverage of each read library can impact these calculations; the median read depth in the multistrain transmission experiment was 57,466 reads, lower than that of the singlestrain group, 102,517.

Table 5.3: Summary statistics of intra-host variant abundances. Differences may not be great, but indicate that N_M is the most diverse group and V_S the least.

	V_M	N_M	V_S	N_S
Variants	6150	6297	5833	5985
Common (10-50%)	38	32	35	26
Rare (1-10%)	129	77	67	61
Very Rare (0.1-1%)	5983	6188	5731	5898
Proportion	51.02%	49.53%	49.00%	48.99%
Frequency	$6.18e^{-3}$	$6.22e^{-3}$	$5.51e^{-3}$	$5.46e^{-3}$
Richness	1.36	0.75	1.05	0.77

5.3.3.2 Simpson Index

Simpson's Index was calculated for each genomic segment in each epidemiological group and the results for each are presented as a summary in Table 4. All 4 of the experimental groups

show roughly the same trend of highest diversity in the polymerase segments as well as segment 5 (NP). Overall, though, when all genomic segments are averaged, diversity is very similar across each epidemiological group.

Simpson's Index is very similar across each of the epidemic groups. Overall however, hosts in the multi transmission chain showed greater diversity than those in the single chain. However, as these probabilities hardly differ between groups, we infer that neither the vaccine status nor the transmission group of the host impact the diversity as measured by Simpson's Index.

Table 5.4: Sub-consensus diversity measures, summarised for each transmission group and vaccination status class. Cells are shaded in gradient, where the more saturated green represents greater diversity. *$\bar{x}$: arithmetic mean, σ^2: variance*

		V_M	N_M	V_S	N_S
Reads		28,652,069	39,290,539	32,892,890	33,269,595
$\bar{x}$ Reads		298,459	306,957	373,782	346,558
Frequency		$6.22e^{-3}$	$6.18e^{-3}$	$5.46e^{-3}$	$5.51e^{-3}$
Richness		0.75	1.36	0.77	1.05
Simpson		$6e^{-3}$	$5.9e^{-3}$	$5.2e^{-3}$	$5.1e^{-3}$
Shannon	H_S	5.76	5.99	5.62	5.65
	H_{SN}	0.33	0.34	0.32	0.32
	H_{SH}	1.40	1.58	1.42	1.56
π	$\bar{x}$ π	46%	62%	72%	71%
	σ^2	±17%	±14%	±9%	±13%
	π_e	$2e^{-3}$	$1.21e^{-3}$	$6.46e^{-3}$	$3.43e^{-3}$

5.3.3.3 Shannon Entropy

Most segments have similar diversity between groups (Table 5). Notable exceptions are segment 02PB1 and 04HA; Naïves in the multi group (N_M) have higher diversity in the PB1 polymerase segment, opposingly sub-consensus diversity in HA is highest in vaccinates in this group (V_M).

Smaller segments, 06NA and 08NS, *only* show sub-consensus diversity in vaccinated individuals of both transmission groups (V_M and V_S). Neuraminidase diversity is highest in multi groups V_M. Diversity in the non-structural protein, however, is highest in V_S individuals.

Table 5.5: Shannon Entropy averaged (H_S) and subsequently transformed. The first normalisation was to the read coverage (H_{SN}) then alternatively to the number of different genomes present (H_{SH})

Caller	Group	H_S	H_{SN}	H_{SH}
LoFreq	V_M	7.46	0.43	1.82
	N_M	7.75	0.44	2.05
	V_S	7.47	0.42	1.88
	N_S	7.47	0.42	2.07
FreeBayes	V_M	5.76	0.33	1.40
	N_M	5.99	0.34	1.58
	V_S	5.62	0.32	1.42
	N_S	5.65	0.32	1.56

5.3.3.3.1 Shannon Modelling

We see a slight decrease in entropy in the single group compared to the multi group (H_S difference of -0.18, p-value=0.0225, t value=13.265) when analysing GLMs constructed with Shannon entropy and sample metadata (transmission group and exposure history). The exposure history of the host, however, does not influence H_S in either transmission group. Entropy is much more consistent across the single group, and both are lower than the multi group individuals.

Lower sub-consensus diversity in viral populations from hosts in the single group indicates that viruses may have been under stronger pressures than in the alternative, multi, group. However, the lack of difference in H_S values between V_S and N_S groups also suggests that virus diversity remains suppressed in the N_S group. This effect is not seen when viruses from the V_M group transfer into N_M hosts; there, the population diversity increases notably upon entering unvaccinated hosts. The multi naïve group (N_M) has the highest entropy scores (5.99) in this group.

These scores indicate that viruses from hosts in the multivalent vaccine transmission chain display more diversity than those from the alternate transmission chain, throughout short outbreak periods (21 days). Additionally, the unvaccinated individuals at the end of this chain have the highest diversity overall. The low diversity in V_S hosts seems to be maintained on transmission to N_S hosts. This is the total opposite of the patterns of diversity seen in the consensus sequences.

These scores indicate that viruses from hosts in the multivalent vaccine transmission chain display more diversity than those from the alternate transmission chain, throughout the short outbreak periods (21 days). Additionally, the unvaccinated individuals at the end of this chain have the highest diversity overall. The temporary suppression of genetic diversity in V_S hosts seems to be maintained on transmission to N_S hosts; it may be assumed that if the experiment continued for longer, both viral populations would display similar levels of diversity once the population levels equalised. This is the total opposite of the patterns of diversity seen in the consensus sequences.

5.3.3.3.2 Viral Population Size in Relation to Sub-consensus Shannon Entropy

In order to observe whether viral population diversity was independent of the mere size of that population, the copy numbers from the qPCR explored in chapter 3 were used to test these relationships. These associations were tested in order to understand the relationships, if any, between viral population size and the sub-consensus diversity present within that population. A linear regression of H_S and the $\log_{10}$(copy numbers) with the host factors of group and vaccine status was used to examine these relations. The results indicate that only the transmission group a host was part of impacted the population diversity (H_S decreased by 0.21, p-value=0.0435, t value= -1.654 when observing transmission group independently). The addition of population size (as copy numbers) does not alter the results of the model and the model performs worse when $\log_{10}$(copy numbers) is included as a variable (ΔAIC = 3.84. The values themselves for the putative interaction of population size and transmission group shows a minute impact.

From this we can conclude that:

- Sub-consensus Shannon entropy of the viral populations was marginally impacted by whether the host belonged to the single or multi transmission group
- Host vaccination status did not influence sub-consensus diversity. The population diversity, as measured by Shannon Entropy did not correlate to the $\log_{10}$ of copy numbers from qPCR values (31.6% Spearman correlation). Further, inferences made by models including the population size performed *worse* than those that excluded qPCR data

5.3.3.4 Population nucleotide diversity (π)

Many ways to measure genomic diversity exist; Shannon Entropy is a mainstay of population ecology but, as shown in Zhao & Illingworth (2019), this measure can be influenced by the depth of reads. In order to overcome this potential flaw of H_S, I next explored the use of a metric that was able to account for read depth, the distance measure that is π nucleotide diversity. There is a slight association between π diversity and host factors. Figure 5.12B shows that the 3 polymerase-encoding genes tend to have the lowest levels of diversity, except in the Single Naive hosts where diversity in these segments ranges greatly. Then the opposite is seen in segments 5-8; in the Multiple Vaccinated group these segments have a range of diversity values which isn't seen in these segments in the other groups.
Wilcoxon rank tests were used to assess differences in diversity between host factors:

- Diversity in multiple/single groups: p-value = $1.4e^{-5}$
- Diversity in vaccinated/naive hosts: p-value = 0.023

So, we can be confident that π diversity is significantly different between the two transmission groups as well as between vaccinated and non-vaccinated hosts.

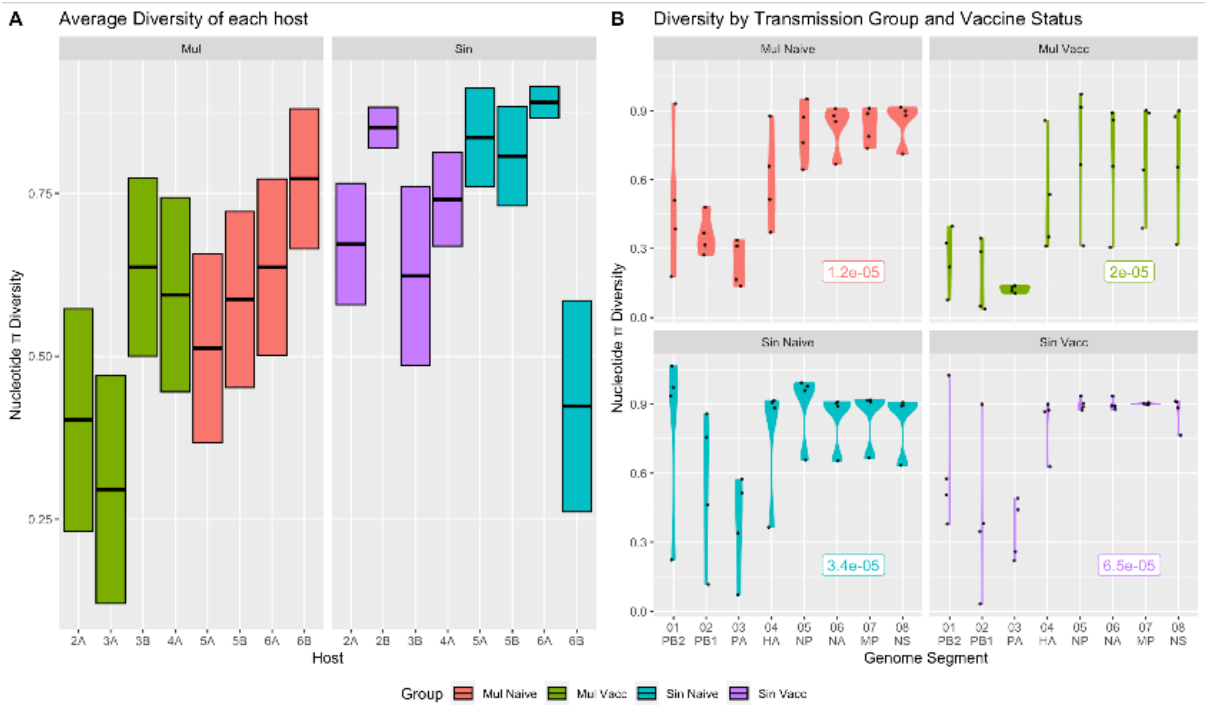


Figure 5.10: A) Average π diversity across each segment, for each host (averaged for all days when samples were collected on more than one day). There is a suggestion of lower diversity in vaccinated hosts than in naïve ones. B) Violins show the range of diversity with respect to each host class for each segment. Genomic segments five to eight consistently show higher sub-consensus diversity than the other segments.

With a GLM investigating π diversity of each segment in relation to the host group and host vaccination status, all possible trends were investigated (Figure 5.12). Firstly, host factors did impact the ranges of diversity observed: vaccinated hosts in the multi transmission group had lower diversity than naïves in this group. Conversely, samples from hosts in the single transmission group had greater diversity in both naive and vaccinated hosts (+0.09, +0.17 respectively). Statistically, we see that nucleotide π diversity correlates moderately (59.9%) to the number of reads.

This assures us that the π diversity is not merely being conflated with population size. As shown above, across epidemic groups, nucleotide π differs in accordance to host factors: the transmission chain each host was part of (Kruskal-Wallis chi^2= 17.009, df = 1, p-value =$3.72e^{-05}$) and the vaccination status (Kruskal-Wallis chi^2 = 3.8588, df = 1, p-value = 0.04948) both statistically distinguish π values.

5.3.3.4.1 Investigating π Diversity Across Genomic Segments

Most genomic segments were more diverse than the baseline 01PB2, chosen as the first and longest segment (Figure 5.12). Genes encoding the three proteins comprising the polymerase complex (01PB2, 02PB1 & 03PA) were the most conserved, though still with a significant amount of diversity (01PB2: 0.53, 02PB1: 0.36 & 03PA: 0.26). Segment 4, encoding haemagglutinin, showed higher diversity than PB2 (0.64) though the signal was insufficient to prove statistically significant. Lastly, the four smallest genome segments (5-8) had substantially more diversity, ranging from 0.79-0.80.

Like many studies of viral evolution, much interest is placed on antigenic proteins, in this case haemagglutinin (segment 4) and neuraminidase (segment 6). One may assume that because of their presentation on the surface of the virion and their role as targets of host adaptive immunity these genes would display the greatest amount of diversity, at both the consensus and sub-consensus level. That this is not the case is unexpected, and yet is seen when comparing Simpson's Index and Shannon Entropy as well as π nucleotide diversity. Results here show higher diversity in genomic segments encoding the polymerase (segments 1-PB2, 2-PB1 and 3-PA) proteins than in any other part of the EIV genome. Though these are the largest segments in the IAV genome, Shannon Entropy and nucleotide π diversity can account for sequence length, so the increased diversity cannot solely be caused by size. Furthermore, one would assume that the polymerase proteins require some of the greatest stability; they are integral to replication of the genome and both their heterotrimeric structure and their functions within host cells necessitate multiple protein-protein interactions.

5.3.3.4.2 Nucleotide Diversity and Population Size

Shedding data, as the $\log_{10}$(mean copy number), was then incorporated to detect any potential relationships between the mean π diversity of and the viral load of each host sample. However, to statistically test whether these variables correlated, I tried three correlatory tests (Spearman's, Pearson's and Kendall's to account for the potentially non-parametric relationships) using π diversity with the mean $\log_{10}$ of viral copy numbers.

No substantial association between shedding and viral population diversity was detected (14% Spearman correlation).Attempting another statistical test of investigate a potential relationship, a linear regression was built. Host factors *transmission group* and *vaccination status* were added to stratify data in the hopes that any signal specific to one subset only would be more visible. Alas, the model did show some weak linear relationship between the diversity and increasing viral load (0.1) but this was only marginally significant.

Neither a GLM nor GAM could find any statistically significant relationship between shedding and diversity, regardless of stratifying and classifying data. Model nomenclature is stated below and outputs are graphically represented in Figure

5.13, with full details of selection processes and model construction explained in Chapter 2 – Methodology (Ch 2 – Section 3.1):

$$log_{10}(mean\ copy\ number) \sim Group + Exposure\ History + Genome\ Segment$$

Removing this bias in measuring diversity allows comparative approaches, where diversity can be benchmarked against other datasets regardless of the viral load. Furthermore, it provides some reassurance that viral load does not act as substantial confounding factor in statistical modelling.

5.3.4 Bottleneck Analysis

5.3.4.1 Shared Variants

Following work from Hughes' (2012) investigation of viral population bottlenecks in EIV, here the number of variants shared between (A) two different hosts and (B) the same host over different days are compared (Figure 5.14). Of the 13,619bp EIV H3N8 genome, the number of sub-consensus variants we reliably called in a single sample ranged from 518 to 5988, with a mean of 4370 $\left(\frac{4370}{13619} = 32\%\right)$ bases that were above the threshold of least 1% in at least one sample.

Summarily, we would expect intra-host viral populations (sampled at different timepoints) to share more variants with each other than with viral populations sampled from individuals in a transmission pair. Then, viral populations between transmission pairs would likewise have more variants in common than those found in hosts with no epidemiological connection. The raw number of variants in each sample seems mostly unrelated to the 'epidemic factors' of that individual, i.e. transmission chain and vaccine status. The low p-value support for all variables in explaining the number of sub-consensus variants found (via a Wilcoxon Rank Sign test) implies a more random distribution of low-frequency mutations throughout the genome rather than being influenced by host factors.

However, when assessing the possibility of host factors influencing the proportion of variants shared between two hosts, the distribution of within-host variants differs largely from that of between-host variants. Within-host variants are more commonly shared than variants between individuals (Kruskal-Wallis chi^2 = 40.719, df = 1, p-value =$1.757e^{-10}$), as expected. The occurrence of shared variants when looking solely at within-host samples is not impacted by any host or epidemic factor. When observing viral populations between-hosts, however, some host variables do appear to influence common variants shared.

Given a transmission pair in the multi group, of two naïve individuals, on average 82.04% of variants will be shared. Transmission between two vaccinated hosts shows a slight decrease in the proportion of shared variants (76.06%, t value= -3.223, p=0.0013) which indicates a tighter bottleneck than seen in the naïve-naïve transmission event. Conversely, there is no measurable difference in shared variants when a vaccinated host infects a naïve one (t=0.601, p=0.5478).

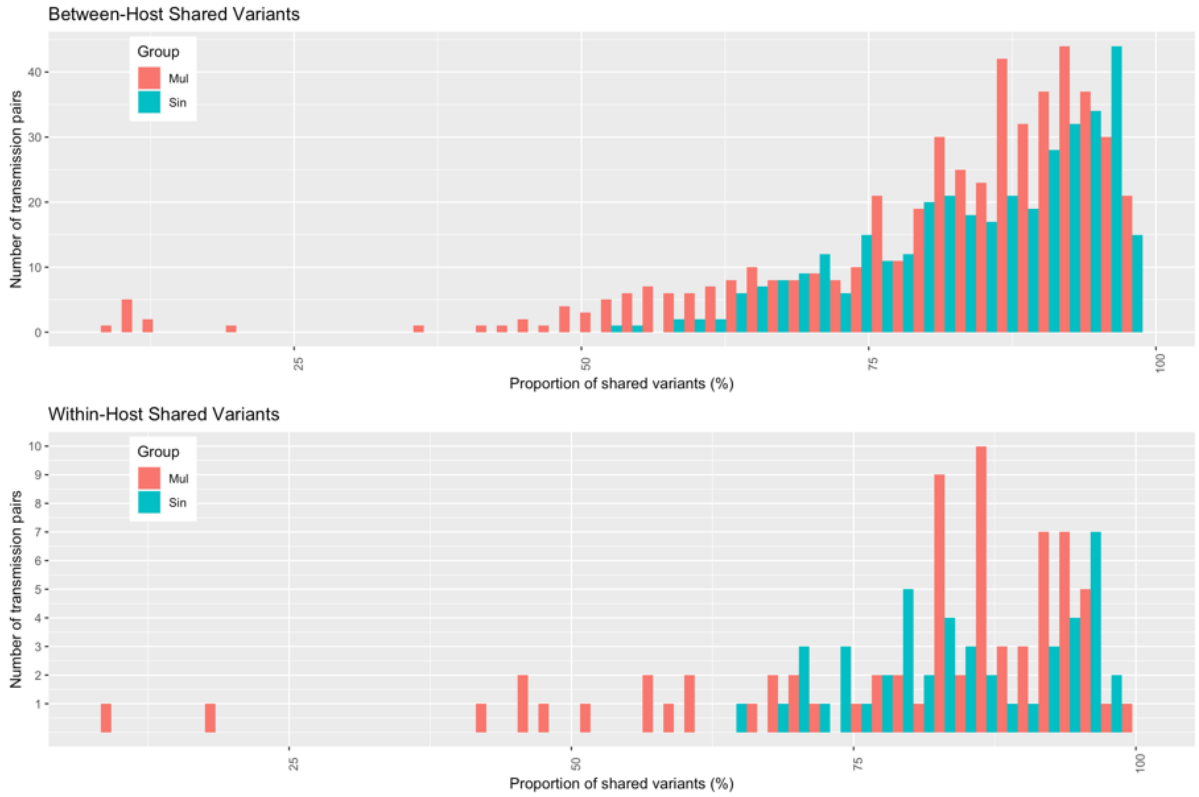


Figure 5.11: Number of LFVs that are found in more than one host divided into two graphs that show whether partnered samples were from the same host. The percentage of variants shared between each possible pairing of viral populations (i.e. donor and recipient) was assessed. Shared inter-host variation (A) ranges widely, from 50-100% of variants shared compared to intra-host variation (B), where most paired samples show greater similarity (mostly 75% or more of the variants are shared).

Switching to the single transmission chain, we see slightly more variants shared between naïve donor-recipient pairs (85.16%, p = 0.045). In this transmission group, we see the opposite trend between viral populations of vaccinated donors and recipients; these hosts now share an increased proportion of variants (1.49% greater, p = 0.06) compared to naïve-naïve pairs. Again, transmission from vaccinated to naïve hosts does not impact the shared variants in these data (p = 0.74).

A distance matrix of shared variants across the four epidemic groups shows, as expected, more similarity in the variants shared by samples from the same host on separate days than samples taken from two different hosts. Matrices of identity between the array of sub-consensus genomes are provided in Figure 5.15. A novel observation, however, is that common variants, that is variants found in more than one sample above the set frequency threshold of 1%, both within- and between-host are more often found in the single transmission chain (89.9% and 84.9% respectively) than the multi transmission chain (85.5% and 81.1%). This implies a wider transmission bottleneck in the multi chain, more lenient to allow greater diversity to pass from one host to the other, or to survive day-to-day during infection.

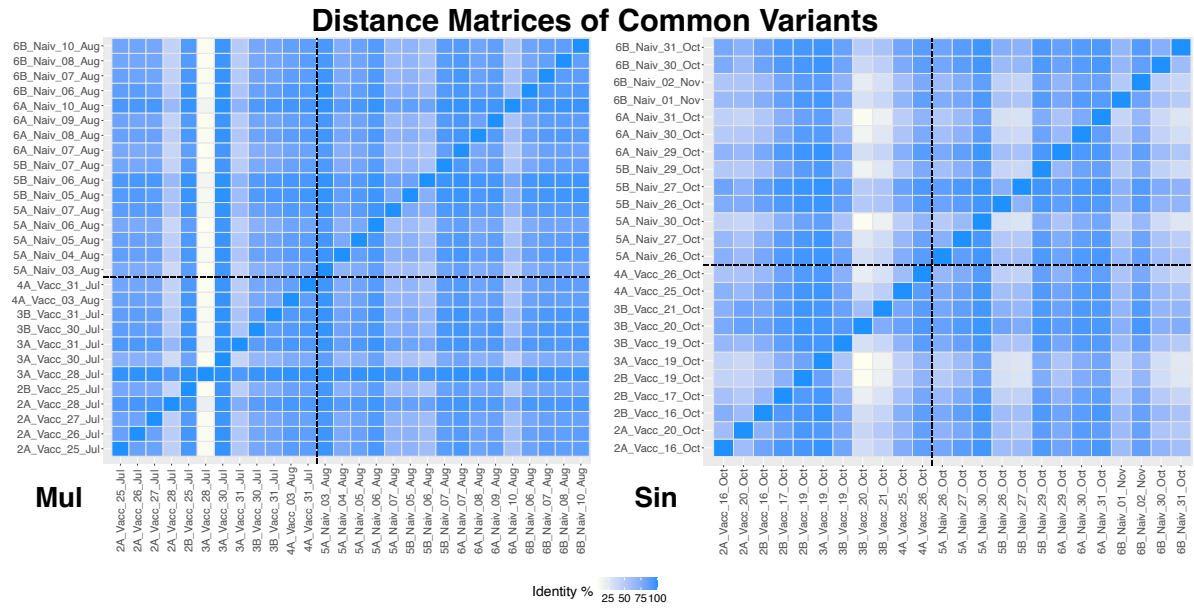


Figure 5.12: Two distance matrices, showing the proportion of common variants between the individuals of each transmission group. Dashed lines are added to divide each graph into vaccinated and naïve quadrants.

From observing the proportion of sub-consensus variants shared between two individuals in these transmission experiments, we can see that generally transmission pairs will share a vast majority of variants, samples taken from the same host over multiple days will have even more commonality in sub-consensus diversity. The vaccination status and transmission chain of the host can make a minor difference to this value: transmission between naïve-naïve pairs in the multi chain is higher than that of vaccinate-vaccinate pairs whereas in the single chain the opposite is true, and vaccinated pairs share more sub-consensus variants in common than pairs of naïve hosts.

5.3.4.2 Variants seen in multiple horses

Some variants are seen in many horses, regardless of whether they are epidemiologically connected or not. Figure 5.16 shows the entire EIV genome, each point coloured for the transmission group from which the sequences were obtained. The abundance of these shared variants is lowest in segments 1-3 (PB2, PB1 and PA), though this could simply be because of their disproportionate length in comparison to the other genomic segments. Roughly 50% of sites are shared between samples, irrespective of the transmission group or vaccination status of the host from which they were obtained. On testing with a Kruskal-Wallis test, variation in the abundance of shared mutations was not adequately explained by epidemiological group alone (Kruskal-Wallis chi^2 = 1.0511, df = 3, p-value = 0.7889).

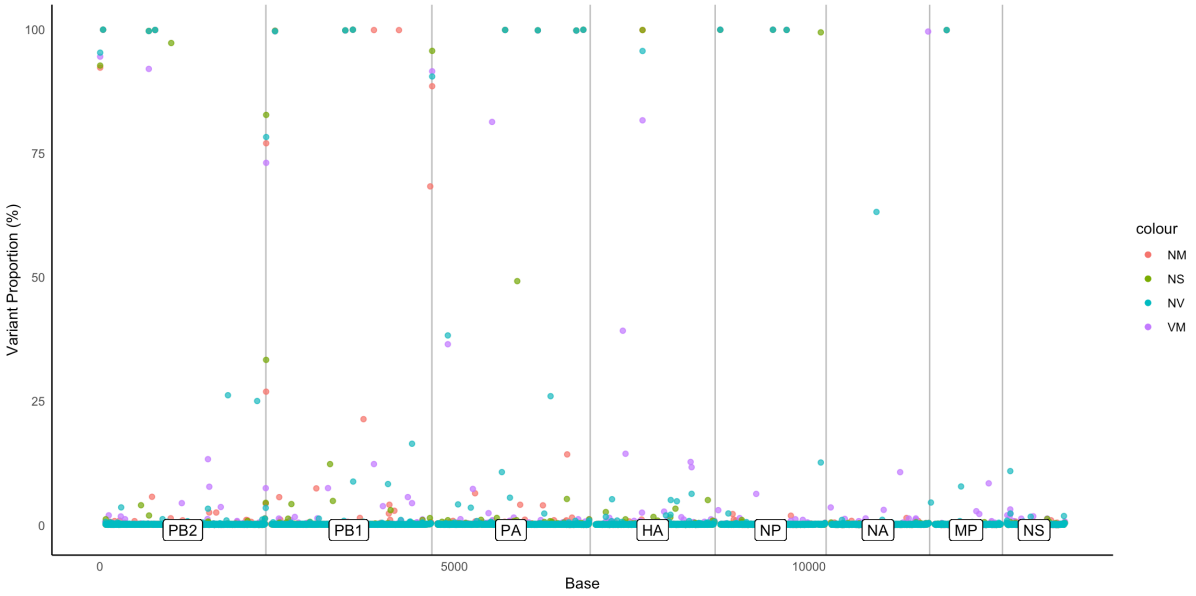


Figure 5.13: Location of LFV throughout the whole 13kb EIV genome.

5.3.4.3 Variants seen in the same horse on multiple days

As an additional investigation into the trajectory of LFV, I collated all variants reported by LoFreq and counted how many appeared across multiple days of sampling from an individual host. Figure 5.17 shows the number of days a variant was detected on the x-axis; from the height of bars, I infer that many LFV tend to appear on multiple days rather than just for a single day. Additionally, variants in the naïve hosts (pairs 5 and 6) are more likely to appear in three or more days than only appearing on one or two days. This is in stark contrast to the patterns seen in vaccinated hosts (pairs 2-4), though more sequences were obtained from naïve hosts than vaccinated ones which may skew this observation.

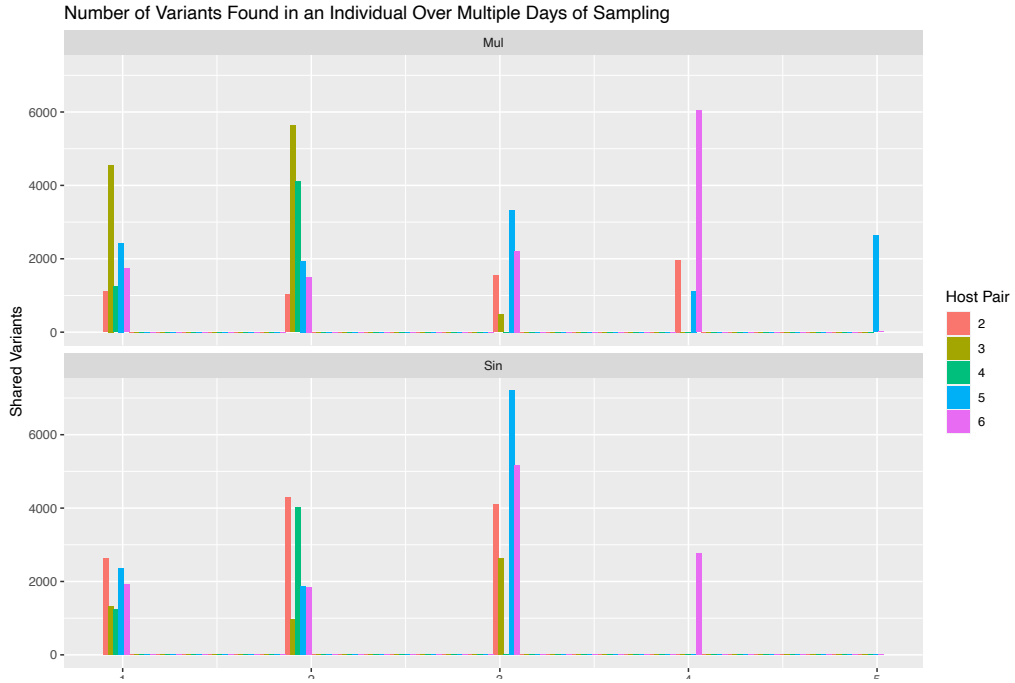


Figure 5.14: The LFV observed in an individual, and the number of days that it appeared in total (not necessarily consecutively). Many variants persist in within-host samples for multiple days This is seen especially in naïve hosts (horses comprising pairs 5 and 6).

5.3.4.4 Singletons

Finally, I conclude with examining the singletons (variants that only appear in a single sample): 1024 singletons appear across the entire dataset (Figure 5.18), only 1.07% of all reported LFV are above the threshold proportion of 1% $\left(\frac{1024}{951,353}\right)$. Variants that appear spuriously usually comprise a very low proportion of the genomes present, all of which are rarer than 10% of the genomes and the vast majority are found at close to the limit of detection – 98% of singletons are found below a frequency of 1.5% $\left(\frac{1006}{1024}\right)$, very close to the 1% frequency cut-off. Singletons may be due to highly unfit or lethal mutations, proving so deleterious to viral fitness that they are purged from the population within 24 hours. More likely, however, these nucleotide mutations are removed due to simple stochasticity and/or procedural error during sequencing. Already present at such low concentrations, the likelihood that a genome carrying a singleton mutation replicates successfully, retains the substitution and then remains at, or above, 1% concentration within the viral population is very low.

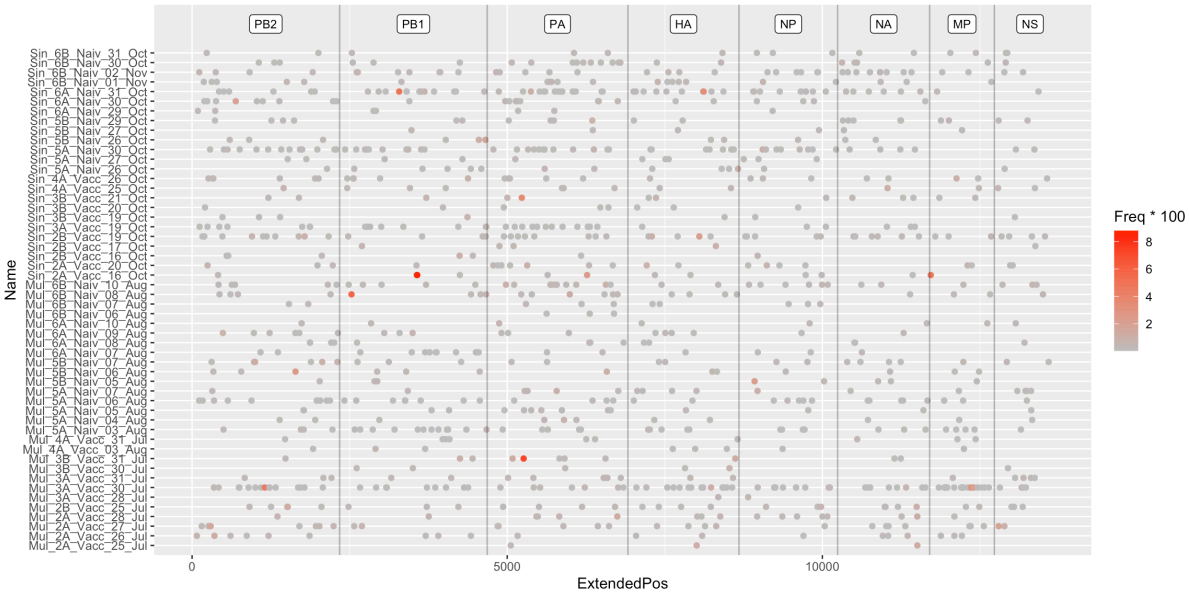


Figure 5.15: Across the entire 13kb EIV genome, LFV singletons are plotted at the nucleotide position they appear. Points are shaded corresponding to the frequency of variants, though the vast majority sit around the threshold of detection (1% frequency)

5.3.5 Beta-Binomial Calculations of Transmission Bottlenecks

Using a variant frequency threshold of 2% to estimate bottleneck sizes with Sobel Leonard's beta-binomial sampling procedure, the possible events in which a donor host could have infected a recipient host were averaged to give single values across each transmission chain. All estimated transmission events are shown in Figure 5.17, by arrows labelled with the size of transmission bottlenecks for each case where it would have been theoretically possible for hosts to infect one another. For example, no arrow connects hosts 3A and 4A in the Single transmission chain; there was no day when 3A shed sufficient virus during the period that it was co-housed with host 4A. Conversely in the Multi group, host 5A was actively shedding virus for two days before host 5B was infected and furthermore, 5B showed no sign of infection whilst co-housed with the preceding pair (4A and 4B). Hence, I inferred

this transmission event to originate from host 5A and therefore estimated the size of this putative bottleneck, explaining the horizontal arrow joining these hosts in the figure.

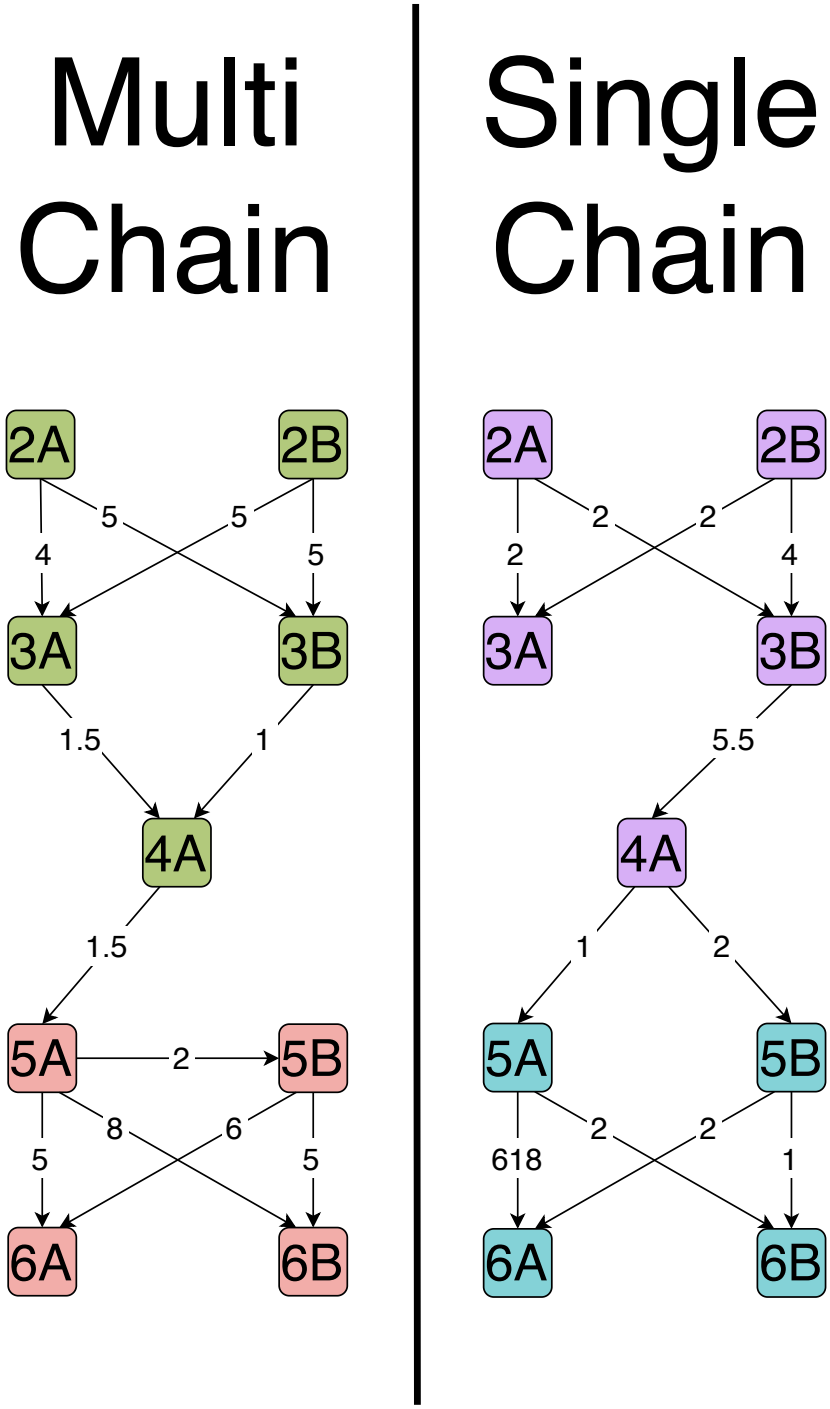


Figure 5.16: Estimated transmission bottleneck sizes between samples.

When filtering datasets by group, N_b differences of within- and between-host links are only statistically significant in the multi chain ($p = 0.028$), not the single transmission chain ($p = 0.14$). Now, comparing the class of hosts involved in the transmission event (vaccinate-vaccinate, vaccinate-naïve or naïve-naïve) we see no substantial differences in between-host N_b of hosts in either transmission chain (multi chain $p = 0.1692$ and single chain $p = 0.3018$). If we instead compare just the vaccination status of donor hosts or recipient hosts, rather than looking at both ends

of the transmission event as above, we still see no evidence that host factors impact the bottleneck size (Figure 5.20).

Bottleneck sizes do not differ between transmission groups (Kruskal-Wallis chi^2 = 4.9099, df = 1, p-value = 0.0267) but are associated to the immune status of hosts. Specifically, transmissions fall into one of three immune classes: vaccinate-vaccinate, vaccinate-naïve or naïve-naïve. Across the dataset, naïve-naïve transmissions tended to have marginally larger bottlenecks than vaccinate-vaccinate (p=0.067) and vaccinate-naïve (p=0.053) ones though it must be noted that the sparsity of vaccinate-naïve samples (n=4) results in low power for testing this and likely contributed to the p-values being near the borderline for significance. These differences in N_b are also linked to the group each host belonged to; naïve-naïve transmission events ceased to differ with either of the other classes when examined in the single transmission group alone. Now, to incorporate the upper and lower bounds of N_b estimates into tests in order to provide a more realistic response variable, we see different trends in the between-host N_b data. The only marked difference in bottleneck size is over transmission events from V_M to N_M hosts (p = 0.027).

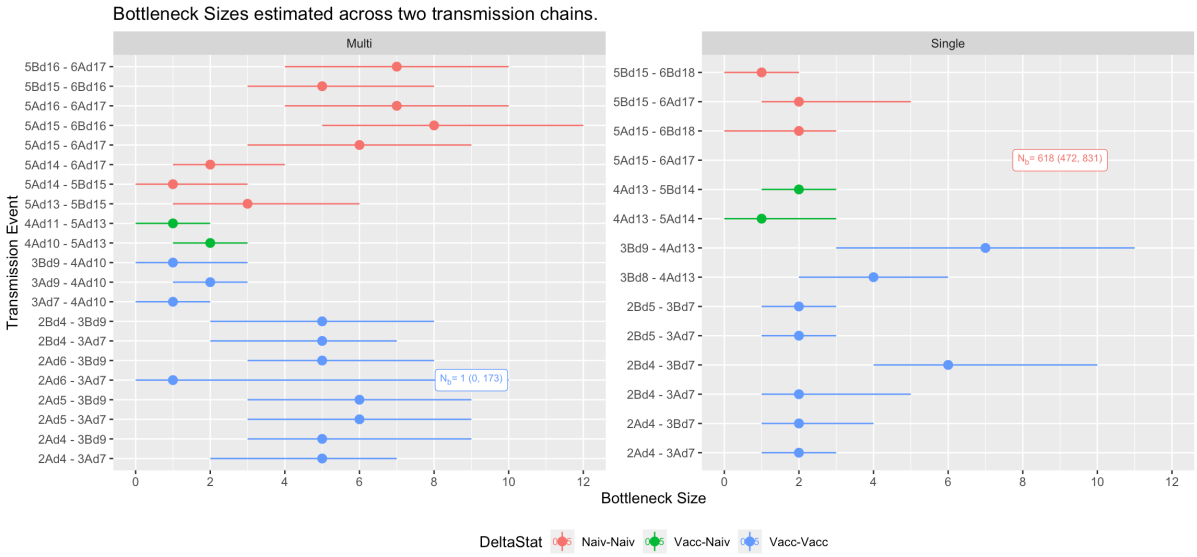


Figure 5.17: Estimated bottleneck sizes for each potential transfer event between hosts. Two outlying values far exceed the rest of the estimates and so for ease of visualising the estimated size (alongside upper and lower confidence intervals) are placed in textboxes at the corresponding event.

5.3.5.1 Host Factors and Bottleneck Sizes

As above, models were constructed under a Bayesian framework in *rstan* with priors estimated from a Cauchy distribution guided by four concurrent MCMC processes over 250,000 iterations:

$$N_b \sim Transmission\ Group \times \Delta Exposure\ History$$

When testing variables with a GLM, the nature of transmission events (vaccinate-vaccinate, vaccinate-naïve or naïve-naïve) proved statistically significant, as did the transmission group. However, the effects of these were minimal; as the N_b values themselves were so small, impacts bore little to no biological realism (Table 5.6).

Table 5.6: N_b values as calculated from the above model with n_{EFF} in brackets, as a proxy for confidence.

	Multi Group	Single Group
Vaccinate-Vaccinate	1.3 (88.27%)	3.5 (93.16%)
Vaccinate-Naïve	0.3 (88.15%)	3.6 (95.19%)
Naïve-Naïve	1.6 (93.16%)	5.1 (92.73%)

These estimates were then repeated below with the addition of sequence information.

5.3.5.2 Interactions & Correlations

As above, when examining measures of diversity, bottleneck sizes were then compared with the amount of virus shed by these hosts. Transmission events are stratified by the exposure status of both the donor and recipient hosts (vaccinated-vaccinated, vaccinated-naïve or naïve-naïve). The shedding of either donor or recipient host bears no meaningful association to the size of a transmission bottleneck size in these experimental data.

Correlations between the genetic distance (identity matrix) and the size of a transmission bottleneck appear minimal. Spearman correlations show a slightly negative association across the entire experiment (-6.24%). When stratified into corresponding transmission groups, correlations were similarly low (multi group = 15.20% and single group = 2.19%).

Desiring to examine these correlations closer, I next replicated the above Spearman tests in a GLM framework. Thus, first testing the impact of bottleneck sizes on genetic distance over both experimental transmission chains, there is a marginal decrease (17%, $p=3.66 \times 10^{-3}$) in sequence identity with a larger founder population in a transmission event. However, this difference is not influenced by the group in which the observed transmission event took place ($p = 0.187$), nor is it significantly impacted by the vaccination status of donor and recipient hosts ($p = 0.979$).

Table 5.7: Proportion of shared variants and the size of bottlenecks (in viral genomes) for each transmission event between hosts.

Group	Class	Events	Shared Variants	N_b
Multi	Vacc – Vacc	11	56.5% (±36.7%)	3.82 (±2.09)
	Vacc – Naïve	2	94.6% (±0%)	2.00 (±0)
	Naïve – Naïve	8	86.1% (±13.2%)	4.88 (±2.59)
Single	Vacc – Vacc	8	92.2% (±5.2%)	3.38 (±2.07)
	Vacc – Naïve	2	83.8% (±6.6%)	1.50 (±0.71)
	Naïve – Naïve	4	86.5% (±6.8%)	155.75 (±200)

In summary, Table 7 reports the average bottleneck sizes of transmission events in both experimental groups, alongside their proportion of shared sub-consensus variants. To note, a sparsity of samples in vaccinate-naïve transmission events skews these rows of results. Overall, both groups show that relatively few viral particles are involved in a single transmission event, though founder populations from naïve hosts appear larger than those from vaccinated ones.

5.4 Discussion

Due to the rapidity of IAV mutations, evolution and epidemic dynamics become intrinsically linked throughout pathogen spread (Kühnert et al., 2011).

Understanding the genetic diversity of even a small viral outbreak can reveal a great deal of information about the interplay of viral populations within and across host (and potentially vector) populations. Though viruses within a single infected host may have a huge capability for acquiring mutations, acute infections vastly limit the timeframe in which a *de novo* mutation can arise to an observable level within a population. Furthermore, should a variant emerge during infection, it is under time pressure to outcompete its progenitors and reproduce in quantities large enough for onward transmission (Xue et al., 2018). Experiments of chronic influenza infections provide enough time and raw mutational plasticity to allow the development of very diverse heterogeneous viral populations (Lumby et al., 2018). How important chronically-infected hosts are in the maintenance and generation of variant diversity remains to be answered.

5.4.1 Reporting sub-consensus viral genomes

In terms of the fastest tool, vSensus performs much faster per megabyte and per read in the library, though iVar is a close second. As a measure of accuracy, the frequency of variants in the population was compared with the frequencies reported in the published datasets. I don't think it's clear which is the better tool just from this, but knowing consistency is helpful. Measuring the time taken for analysis against the number of variants called is again, not a simple metric as it depends on the question one is trying to answer.

Observing the population diversity of viral genomes from each sample is one of the most common analyses performed with such deep-sequencing data of viruses. Testing the population richness and diversity then shows the kind of output one can expect using each tool.

Overall, the stand-out VCT for academic research of viral evolution are both FreeBayes and LoFreq. These tools have the highest accuracies except for DeepSNV (96.52% and 94.14% respectively). Despite higher accuracy (96.06%) and faster runtimes, DeepSNV comes with the large caveat that processing requires a control dataset. This necessity makes DeepSNV unsuitable for analyses of *de novo* viral sequences, and the lengthy set-up before running the tool is not included in the overall timing.

5.4.2 EIV within-host variation

Combining a range of diversity measures, we can confidently infer that host factors such as vaccination status do affect the sub-consensus diversity of viral populations. Between hosts, viral population diversity mostly differs based on the host's exposure and vaccination status. Unexpectedly, the two most commonly used diversity measures, Shannon Entropy (H_S) and nucleotide π detected different groups as the most diverse.

Shannon Entropy (and its normalised forms) is highest in naive hosts in the multi transmission group (N_M), whereas π diversity is highest in vaccinated individuals within the single transmission group (V_S). This difference in where diversity sits is unexpected, but as noted by Zhao et al. (2019) this may be skewed by the distribution of reads in each population; the N_M group has substantially more reads than the others.

Between the divisions of vaccine status, most measures show very little change. Averaging the diversity of all vaccinated hosts to all naïve ones, there is a stark difference in the richness of mutations (average naïves=1.205, average

vaccinates=0.76). Hosts of each transmission group do, however, show substantial differences in mutational richness, i.e. Shannon Entropy & π diversity. Within hosts, nucleotide π diversity either changes erratically over the duration of the experiment or remains relatively consistent. The occasions where we do see fluctuations in within-host diversity are mostly confined to the longer segments. Levels of within-host π diversity certainly appear stable in segments 6-8 of most hosts, retaining constantly high π diversity. I interpret this as these genes evolving at a constant rate, but unable to find any fitness advantages within their mutations. The longer segments however appear to be exploring mutational space more, and sudden peaks in their sub-consensus diversity imply periods of fast evolution. Alternatively, troughs in this diversity implies having honed-in on a particularly fit mutation, and thus generation of *de novo* mutations slows as the viral population accumulates a particularly fit mutation.

Further, neither π nor H_S diversity measures show a strong association with viral shedding. First, the relationship between each metric (H_S or π) and $\log_{10}$(copies) was investigated with correlatory tests (Spearman); Shannon Entropy doesn't show any correlation to the viral load (31%), nor does nucleotide π (14%).

Using a linear regression to quantify the relationships between variables the population size, as measured by $\log_{10}$(copy number), had no impact on the Shannon Entropy in any of the 4 epidemiological groups. By comparison, the nucleotide π diversity of hosts marginally increases with a larger population size in the multi group, though insignificantly (0.108, p=0.0675).

5.4.3 Transmission Bottlenecks of Naturally Transmitted EIV

To note, an important caveat of the above analyses is a reliance on the absolutism of transmission events within the confines of the experimental design. Though the transmission experiment only housed two pairs of hosts at a time, we cannot rule out that some of the virus shed remained infectious in the environment. Influenza viruses are capable of mechanical transmission, through fomites in the surrounding environment. One unfortunate consequence of this in the present study is that proving transmission occurred directly, exclusively between hosts in the mixing chamber, is not possible. Previous calculations of Influenza A Virus bottleneck sizes (N_b) vary, but mostly concur with the low (<10 virions) averages we report later.

- McCrone and Lauring (2018): 2-5 virions or up to 200 genomes (experimental ferret transmissions)
- Johnson and Ghedin (2020): 7-24 genomes in contact transmission, or 3-5 genomes with droplet transmission (human transmissions)
- Dimas Martins and Gjini (2020): 90 (±45) genomes in another ferret transmission study
- Sigal, Reid, and Wahl (2018): Their system requires an N_b of 20-100 genomes to adequately explain diversity
- LeClair and Wahl (2018): *in vitro* IAV transmissions barely worked with N_b of 1, but functioned well at N_b = 5

Overall though, the high proportion of shared variants between pairs of hosts indicates a generally loose transmission bottleneck in close-contact EIV infections. This enables viral populations to maintain a high level of the diversity generated in one host and transmit it to the subsequent host; essentially the mutations generated in a host have a good chance of surviving and passing to the next individual. This has

potential phylodynamic implications, variants generated *de novo* in hosts of the single chain have a better chance to be maintained and passed forward than mutations in the multi chain. To note, this analysis does not account for the proportion of each sub-consensus mutation and is simply counting the presence/absence of mutations at each possible nucleotide. Having estimated transmission events between co-housed hosts, we illustrate that under experimental outbreak conditions, low numbers of virions are involved in onward transmission of EIV. Non-parametric tests stated that neither of the examined host factors (transmission group or vaccination status) significantly impacted the bottleneck size.

Comparisons of viral populations within-hosts day-to-day using the same beta-binomial sampling methodology showed, as expected, much looser bottlenecks; greater numbers of and more diverse collections of virions link the viral populations of hosts from one day to the next.

Our data confirm that within-host IAV populations are highly dynamic, with multiple variants arising, persisting, and sometimes becoming transiently predominant or fixed, even during short chains of transmission. Future work linking minority variants between different animals will inform of the size of transmission bottlenecks during natural infection.

6 Discussion

In this thesis, I investigated the impact of prior Influenza A Virus (IAV) exposure on viral evolution at the within-host and inter-host level. To this end, I examined the virus population size and genomes of influenza viruses in infected horses that possessed different immunological histories and were linked by transmission.

6.1 Equine Influenza as a Model Virus

The use of equine influenza virus (EIV) as a model system for IAV outbreaks in mammals captures both direct and fomite-mediated transmission, with a virus known to jump species barriers. IAVs infect a broad range of mammalian and avian hosts and cross-species transmission occur sporadically, but with often dramatic consequences. Influenza in horses presents with similar symptoms to the disease in humans, and their movements are linked to anthropogenic activities. Though reports of active EIV infection in humans are both rare and sporadic, people working closely with horses often develop circulating H3N8 antibodies and this may be taken as evidence that EIV is at least able to colonise human hosts, who develop adaptive immune memory in response.

The spillover of H3N8 viruses into canine populations occurred concurrently with five nonsynonymous mutations in the haemagglutinin: Asn54Lys, Asn83Ser, Asn154Thr, Trp222Leu and Ile328Thr (Crawford et al., 2005). Though cross-species spillover is a multi-factorial event and cannot be attributed solely to protein conformational changes, such observations reveal that IAV can successfully adapt from avians to equines and then on to canine hosts.

6.2 Shedding of Equine Influenza Virus

The lower amounts of virus shed by both vaccinated and unvaccinated individuals of the single strain transmission chain speaks to the impact of prior exposure to immunogens that specifically match the strain hosts are challenged with. From this, I must then infer that the immunity of a host affects not only its' own viral load but also that of the hosts which it infects. This is important because, as discussed below, horses most likely to be moved around the country and interact with horses external to their day-to-day cohort (analogous to the notion of super-spreaders) are simultaneously likely to have been exposed to the greatest array of circulating EIV strains. However, these observations may be confounded by the ultimately different viruses at the end of each transmission chain.

Viral load can be used as a proxy for infectivity in epidemics of acute disease spread in a frequency-dependent manner. Once the quantity of viruses surpasses the threshold needed to establish infection, i.e. the minimum infectious dose (MID), it is generally assumed that the more virus present (whether in the environment outside of hosts, or viraemia for viruses spread by direct contact) the greater chance an exposed host has of becoming infected.

These findings on the amount of viruses shed and the influence of host adaptive immune status could additionally help parameterise epidemiological values such as the Critical Community Size, a term describing the proportion of susceptible hosts

needed in a population to prevent epidemic fade-out (Bartlett, 1960; Cliff et al., 2000) or p_{crit} which describes the proportion of the population which must be protected in order to stop outbreaks from occurring. Knowing that even vaccinated horses can shed sufficient virus to lead to further infections, the data generated here could also be incorporated into compartmental epidemic models. Vaccination decreases viral shedding, with the vaccine matching the challenge strain having a greater inhibitory effect. From this, we may therefore assume that distributing vaccines which closely match circulating strains would be a better public health measure than distributing vaccines that confer lower levels of immunity but to a greater range of viruses. However, a broader coverage would perhaps better account for unknown strains that may be in circulation. My work shows that providing some form of vaccine-mediated immunity is better at limiting the spread of virus than no immunisation and further, it reflects the real-world dynamics of EIV epidemiology, where imperfectly neutralised virus is still capable of transmission between horses.

Though still lacking the values to estimate CCS and other compartmental epidemiological models, these findings highlight the importance of including vaccinated horses in such models, acknowledging the potential contribution of such hosts to maintaining chains of transmission. Here, however, the relatively low viral load of univalent-vaccinated (V_S) hosts suggests a decreased capacity for shedding infectious virus. This in turn reduces the capability of EIV to transmit as effectively as in wholly naïve populations. First, by shedding a lower number of infectious particles, transmission events will require closer and/or longer duration of contact in order for recipients to receive the minimum infectious dose required to establish infection. If a successful viral transmission event becomes more difficult, each host is less likely to infect as many susceptible hosts as previous conditions permitted; the effective reproductive number (R_e) would fall as secondary transmissions from each host become increasingly rare. Clearly, the effects seen in the present transmission experiment are not enough to halt onward spread since all hosts in the experiment became infected. Importantly, unlike the natural epidemiology of EIV, data presented in this study come from an experiment designed specifically to facilitate continued transmission; hosts were kept in very close proximity until recipients showed signs of infection. Clearly, this creates artificial scenarios that would not be expected in the field. Yet as the focus lies on the evolutionary forces experienced by EIV, ensuring that each host became infected was paramount. *In situ* outbreaks rarely, if ever, see 100% of horses infected with EIV; understanding the limits of spread once a premises has been seeded by an index case could inform disease management strategies.

Upon infection, virus replication clearly occurs rapidly, though linking this to disease progression and the course of symptoms is unexplored as I did not have access to clinical information. In other hosts, associations between viral load and host disease presentation have been examined in populations of young adults (McKay et al., 2020) where minor correlations existed between viral load and patient body temperature. In contrast, paediatric patients showed that both symptoms and recovery time were correlated with the amount of virus present (Tran et al., 2023), when trialling a therapeutic probiotic intended to limit IAV infection by over-colonisation of nasal epithelia with commensal bacteria.

However, in human IAV infections, viral shedding (and thus transmission of virus) begins prior to the appearance of symptoms (Andrew et al., 2023; McKay et

al., 2020). Given the short period between exposure and detectable virus in infected horses in this experimental setting, viral transmission from pre-symptomatic horses is to be expected. The major epidemiologic consequence being that pre-symptomatic index case(s) could spread EIV through the population, as seen in previous FMDV outbreaks (Firestone et al., 2019; R. J. Orton et al., 2020; M. Woolhouse et al., 2001), delaying appropriate responses (e.g. quarantine/distancing, prophylaxis, alerting a vet), at which point it may be too late and EIV has spread to multiple other contacts. A paradigm of RNA virus evolution is that they are highly polymorphic and this is due to the error-prone polymerase which causes mutations to appear *de novo* throughout the genome. Thus, field populations are assumed to exhibit a high level of random mutations, with the number of mutations detected being proportionate to the size of the sampled population. Unexpectedly, the viruses sampled from naïve hosts were found to be largely homogenous, despite being much larger populations than those from vaccinated hosts. While one would expect that a larger population of viruses would allow for more variation to be generated, this, was not observed in either transmission experiment.

In real-world settings, the index cases of EIV are more likely to be protected (horses for sports/breeding are likely, or required, to be vaccinated) and so may display minimal symptoms, if at all. It may be hypothesised that those days in which shedding was not detected due to a lack of noticeable symptoms would allow continued, uncontrolled transmission throughout the population. Observations of influenza infections in human populations concur; many individuals that test positive for IAV infection show no symptoms or illness. Indeed, screening by Hayward et al. (2014) have revealed that 77% and 83% of IAV-positive individuals were asymptomatic, depending on whether the screen was carried out using serology and/or PCR amplification respectively . Arguably, the ability to go unnoticed by the host is the fittest (defined by genomic reproductive success) adaptation of some viruses; non-pathogenic infections are especially effective when spreading through populations of animals that display social behaviour. Hosts displaying visible physical symptoms, such as coughing or mucopurulent nasal discharge, may be avoided by other individuals of the same species thus reducing the amount of contact and potential for secondary transmission of the pathogen. Hence, asymptomatic infections could facilitate continued transmission, overcoming anti-social host behaviour which would otherwise limit contact rates.

Across the transmission experiment, hosts shed substantially different quantities of EIV genomes based on their exposure histories. Recognising the caveat that qPCR quantifies only the number of vRNA copies and not replication-competent virions, the copy numbers in samples from unvaccinated hosts show a greater number of viruses compared to vaccinated ones. However, only naïve hosts in the multivalent transmission group (N_M) show significantly greater quantities of virus shed compared to vaccinated hosts. Hosts without any history of exposure in the Single group (V_S) did not shed substantially different amounts of virus to the hosts with exposure histories. This may be caused by wide-ranging values from samples in this group and/or the smaller founder population that these hosts receive; as V_S hosts shed low amounts of virus, the hosts they infected are expected to have

received a small infective dose. Thus, even if viruses in both transmission groups were equally fit, those in N_M hosts were seeded with a larger initial viral population leading to their significantly greater viral shedding. To note, as I will explore below, the viruses infecting N_M hosts were genetically distinct from those sampled from N_S hosts, with 2 non-synonymous and 1 synonymous mutations separating the two populations.

Reported values of viral shedding concur with those published in other studies of influenza A infections. Average daily shedding closely resembles that seen in other experimental infections of horses (Murcia et al., 2010, 2013) and, as anticipated, this was higher than in samples collected during an outbreak (Hughes et al., 2012). Likewise, in accordance with studies in horses and pigs (Lloyd et al., 2011), vaccinated hosts shed less virus than the unvaccinated hosts. These figures also concur with those reported in human IAV infections (To et al., 2010; Ward et al., 2004). As explored in the introduction of Chapter 3 (3.1.2), the amount of virus shed by a host is expected to correlate to the infectivity of that individual, i.e. the ß parameter or Force of Infection (Heesterbeek, 2002; Matthews & Woolhouse, 2005) in compartmental epidemiological models. A strengthened Force of Infection indicates an increased chance for secondary transmission from an infected host, thus reducing the amount of time susceptible hosts need to be exposed to a shedding host to acquire a dose sufficient to establish EIV infection. When considering the virus, this greater ß parameter leads to a larger viral population in circulation in the local environment; this likely increases the impact of selective forces and decreases those of stochastic fluctuations in mutant genome levels.

Observing the transmission of EIV in such a controlled environment allows for examination of the transmission dynamics in a way hitherto understudied in equine populations. By quantifying shed virus in each transmission event, the most striking result was how quickly infection is established in a host. Moreover, hosts become infected and then are able to spread infection very rapidly. All hosts began shedding detectable amounts of virus by two days post-contact with infected hosts, providing a very short interval between successive transmission events, i.e. the serial interval. In terms of viral evolution, the rapidity of this transmission poses a double-edged sword: fast spread of course benefits the virus in the short term, as infections can be established and transmitted before adaptive immune responses can be fully mounted. However, as the virus transmits so quickly, any diversity generated in the course of the host's infection may be lost unless the maintained virus is able to spread and/or re-infect other hosts. Reporting the serial interval of EIV amongst populations with heterogeneous exposure histories grants novel insight into its' epidemiology within horse populations.

6.3 Ensuing Work

Conclusions may be drawn on the assumption that a higher viral load correlates to a virus better able to replicate, i.e. the virus replicates more and/or faster thus increasing the quantity of genomes collected by each nasal swab. Hosts may shed different quantities of virus for myriad reasons, however, many confounding variables interfere with this assumption. Thus, even though viral load is occasionally used as a proxy for fitness and/or competitiveness later in the analysis, it is far from

an ideal measure. A lingering question from this study is whether shedding can be used as a legitimate approximate of viral fitness. To support the work presented here, proof of the correlation between viral load and replicative fitness could be carried out *in vitro*. Potential avenues to test this include comparing parallel growth curves of different genotypes in cell culture or competitive co-culture of differing viral genotypes in the same culture to assess phenotypic strengths. These would allow for estimates of competitiveness, or at least comparisons of fitness within such a closed system (Domingo et al., 2019), and give an indication of what Wargo & Kurath (2012) distinguish as both replicative and transmission fitness.

However, the virus was able to infect vaccinated hosts in both transmission chains, so clearly it was fit enough for continued natural transmission. Hosts with prior immunity were able to become infected and then shed sufficient quantities of virus to lead to secondary infection of other hosts with similar exposure histories. Concerningly, even horses which would be expected to have very strong adaptive immune responses, due to their recent exposure to the challenge virus, were capable of transmitting infectious virus to other hosts with histories of exposure. V_S hosts had five exposures to inactivated stock of the challenge virus across a period of 40 weeks, 40 weeks before being vaccinated in the transmission experiment and could still transmit and be infected by EIV, showing a lack of sterilising immunity generated by the memory response. Under such conditions, contrary to what is generally accepted, this ability to reinfect hosts despite strong immune memory suggests that selective pressures originating from host immunity may not be particularly strong in guiding EIV evolution. Though the conditions of the experimental transmission were highly controlled and would not be expected to resemble those in natural settings, horses are also unlikely to have such comprehensive history of exposure to a circulating strain. However, immunity developed from natural infection, rather than exposure to inactivated whole-virus formulations could establish a stronger memory response for these hosts.

Additionally, were the experiment to be repeated I would endeavour to quantify the amount of virus in the immediate surroundings of infected hosts, such as nearby surfaces or even suspended in air, as in studies by Neira et al (2016). Understanding the viral load of a host plus the number of viral genomes (and thus assumedly infectious particles) in the vicinity of an infected host would reveal infectious particles present in the environment. Fomites are known to be important environmental vectors of transmission in seasonal human IAV; evidence for mechanical spread of EIV was reported in Australian and South African outbreaks (Cullinane & Newton, 2013). However, during this transmission experiment, after each transfer of hosts, the entire mixing chamber was thoroughly disinfected thereby removing any possibility of fomite transmission. How this might affect viral evolution is less known; once virus is shed into the environment it cannot mutate/evolve until entering another host.

Further areas of investigation ought to include the determination of the minimum infectious dose (MID) of H3N8 EIV in horses. Knowing the number of viral particles needed to establish a successful infection, *in vitro* or *in vivo*, would allow for a proper quantification of the force of infection (ß) needed for epidemic spread ($R_e \geq 1$). Such experiments would, however, involve prohibitively high numbers of horses to test which raises ethical concerns. Many factors such as MID exert an

influence on host-pathogen interactions, including host innate and adaptive immunity, host intrinsic barriers to infection (e.g. skin and mucus), co-infection, viral strain and viral fitness to name but a few. To estimate MID while accounting for these potentially confounding variables would necessitate large numbers of horses to be infected and observed in controlled environments. Alternatively, scaled-down experiments involving serial dilutions of viral cultures applied to tissue explants and/or 3D tissue cultures may grant some insight into cellular MID estimates.

Although viral genomes were collected from nasal swabs, matching disease symptoms to shedding would also help in the construction of epidemiological models; how do clinical symptoms correlate to the pattern of viral shedding? Influenza has a reputation for being most contagious in pre-symptomatic hosts (Bell et al., 2006; Hayden et al., 1998; Webb et al., 2010), but this remains to be proven in equine populations. While correlating disease progression with daily patterns of viral load has not yet been undertaken for EIV, going forward I would endeavour to use existing datasets of loads and published details of symptomatic horses in a meta-analyses to craft inferences on this relationship.

6.4 Consensus Analyses

As seen especially through the emergence of variants of concern (VoC) during the COVID-19 pandemic, viruses, like all life forms, do not evolve in a straightforward, linear manner but rather generate an array of variants upon which selective forces and stochasticity can act. Single Nucleotide Polymorphisms (SNP) in viral genomes can have positive, negative or neutral impacts on the overall replicative fitness of that individual virion. The trajectory of mutations within a viral population at the within-host and between-host scales can reveal insights into how diversity can be generated, transmitted and maintained at a global level.

Examining the genetic similarities between the EIV strains which comprise the whole-virus inactivated inocula and that of the challenge strain established just how different the immune responses of hosts could be. Memory responses of the adaptive immune system rely on the recognition and presentation of antigens to, primarily, memory B cells. The four strains comprising the multi-strain inoculum regiment shared 92.25% genetic identity, meaning an average difference of 1,056 nucleotides between the inactivated viruses used to provide exposure histories to the V_M horses.

In comparison, on examining three commercial EIV vaccines (the recombinant Canarypox vaccine ProteqFlu, plus two inactivated virus vaccines Equilis Prequenza and Equip F (NOAH, 2016)), the strains formulating each shared varying sequence similarity. The immunogens used in the trivalent Equip F vaccine differed the most from the other vaccines and was also had the most diverse range of immunogens of any vaccine, including a now extinct H7N7 virus; similarity between the genomes of inactivated viruses marked only 70%, stimulating recipients with three very different viruses. Though some meta-analyses on vaccine formulation have been carried out (Elliott et al., 2023), the impacts of such broad immunisation on viral evolution remains a mystery. According to the results presented above, exposure to 3 highly distinct immunogens may not be as effective as other formulations. These three EIV vaccines are approved for use in the UK and EU, and though the manufacturer of

each (ProteqFlu – Boehringer Ingelheim, Equilis Prequenza – Intervet International B.V. and Equip F – Zoetis UK) declined to provide sales figures, many of the 850,000 horses in the UK (BETA, 2024) are assumed to receive at least one of these if they have been vaccinated.

Though all advertise protection from EIV, the three contain different vaccine strains and most importantly differ in their formulation. Equilis Prequenza and Equip F are both inactivated, whole-virus vaccines comprised of two and three unique EIV strains respectively. ProteqFlu uses two EIV strains, but only incorporates a single genomic segment in its composition: the haemagglutinin, as encoded by segment four, is carried by a recombinant canarypox virus. We hence see two vaccines stimulating immune responses to an entire, replication-incompetent virus and one vaccine training horse adaptive immune systems exclusively to haemagglutinin.

While direct comparisons of vaccine efficacy have not been performed, the differences in the strength (as measured by antibody titre upon challenge) and breadth (range of EIV strains that elicit a memory response) of immunity provided are unlikely to be equal. In fact, given the results from the investigations above, the use of multiple heterogeneous immunogens in vaccinations may not be providing the best protection for individuals or for populations against circulating EIV. This is especially due to the distance between some vaccine strains and those currently circulating; this is represented graphically in Figure 6.1.

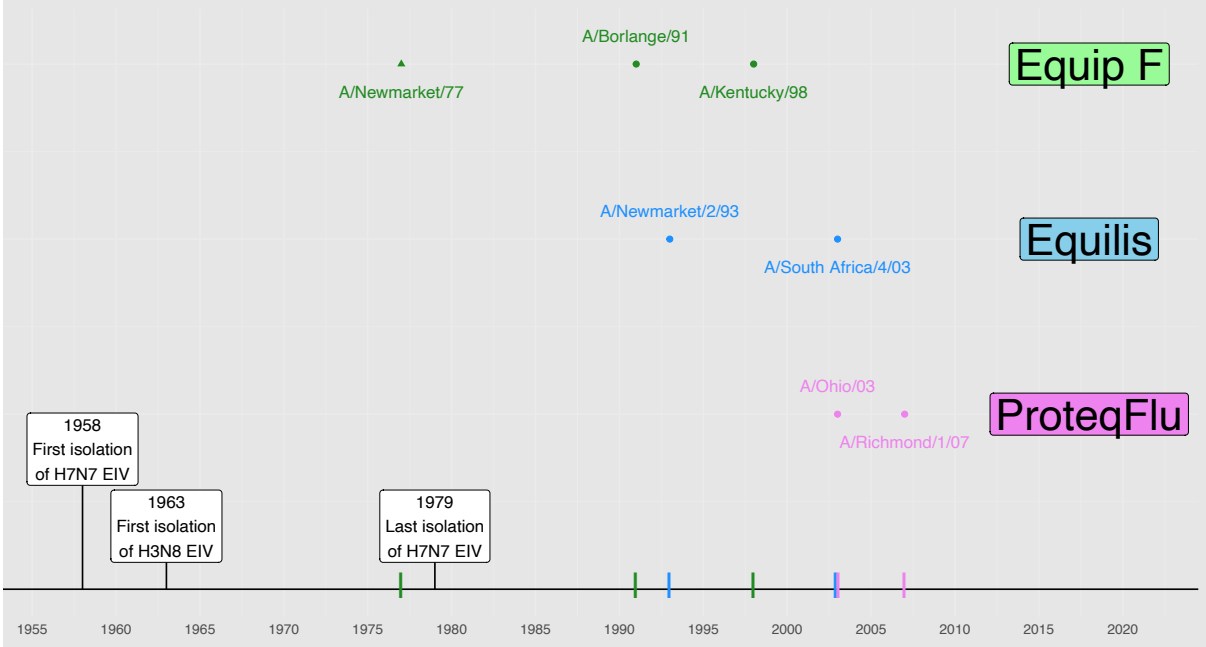


Figure 6.1: Brief timeline showing the three main commercial EIV vaccines sold in the UK, and the original strain upon which they are based. A/Newmarket/77 is represented by a triangle in order to show its unique inclusion as an H7N7 virus.

Despite no longer circulating in the wild, one of the commercial vaccines (Equip F) continues to add inactivated H7N7 viruses into the formulation, a decision that I believe may do more harm than good. Though tests explicitly comparing the impacts of cross-immunity from H7N7 to H3N8 viruses have not been carried out in horses, differences between these viruses may be exemplar of the imprinting described by Gostic and others (Gostic et al., 2016, 2019; Kelvin & Zambon, 2019). If the host's first infection is from a virus with group 1 haemagglutinins (e.g. H1), subsequent infection with viruses presenting group 2 (e.g. H3) proteins will elicit weakened immune responses compared to infections with other group 1 viruses. H7 and H3,

however, both fall within group two of the haemagglutinin subtypes. A similar trend is seen in neuraminidase groupings, yet as HA is around four-fold more common on the surface of virions (Bouvier & Palese, 2008), the immune-dampening is less pronounced. However, N7 and N8 are grouped separately in neuraminidase trees (N7 is a group two neuraminidase, N8 is group one). Though the effects on host immunity caused by imprinting has been shown to be more dramatic in haemagglutinin, differences in divergent neuraminidase proteins may also hamper immune memory responses.

However, conclusions drawn from another equine influenza transmission study (Park et al., 2009) included the observation that even vaccines with mismatched or outdated immunogens can still provide moderate protection at the population level when sufficiently distributed. Uncertainty in circulating strains and manufacturing constraints continue to encourage the formulation of vaccines conferring immunity to multiple viral strains.

Viral mutations appear in horses and may appear or disappear regardless of their impact. As this experiment could be seen as a small-scale EIV outbreak, with a small population and short duration, the many singleton nucleotide substitutions reported here describe well the trajectory of mutations through a population, most notably in representing the stochastic nature of what happens to mutant genomes. Whether or not a mutation becomes fixed in a population is determined by both selective forces and random chance. As the hosts with exposure histories are expected to mount rapid memory immune responses, viruses in these hosts ought to be under greater selective pressures in attempts at removal by adaptive immune cells and molecules. During the transmission experiment, diversity among consensus sequences was seen almost exclusively in these historically exposed hosts (V_M and V_S classes), suggesting the existence of strong pressures forcing genetic diversification and rapid removal of any less-fit variants. I thus understand that in these controlled settings, the fixation of consensus mutations is less likely in populations infecting hosts with previous IAV exposure. As selective immune pressures are lower in the naïve hosts, virus genomes are able to sustain neutral, or even slightly unfit, mutations without being purged as severely as replicating viruses in hosts with rapid, specific immune activation. Below the consensus level, diversity (i.e. Shannon entropy and π nucleotide) is highest in viral populations infecting naïve hosts, with no previous exposure to IAV.

The host environments in which EIV replicates are not all equal; differing exposure histories, not to mention possible host heterogeneities, can dictate the quantity and quality of antibodies present in mucus. Such host environments do not necessarily affect evolutionary rates of EIV, but they can create situations that encourage immune selection or reduce the effect of stochastic removal of genomes.

Unexpectedly, the output of mutations was almost entirely even, i.e. 10 synonymous and 11 nonsynonymous mutations. As nonsynonymous mutations are more likely to impact phenotypes than synonymous ones, and due to codon redundancy, synonymous mutations are generally expected to occur more frequently. That nonsynonymous mutations appear more often could indicate an exploration of the fitness landscape; strong selective pressures in vaccinated hosts may drive the proliferation of nonsynonymous mutations in order to develop a fitter phenotype.

6.4.1 Global EIV Sequences

Many of the 21 consensus mutations observed across the experiment were also observed in publicly available EIV sequences. From this, and the mutations reported in both transmission chains, some level of convergent evolution or hypervariability in those sites must be suspected. In contrast, the mutations that appeared most frequently (PB1-t1500c/Gly500, PB1-a1853g/Glu618Gly and NP-g1445a/Ser482Asn) and became fixed at the end of each transmission chain were rarely reported in global EIV sequences. Indeed, the nonsynonymous mutation PB1-a1853g/Glu618Gly, seen in all naïve hosts in the multi group, is never observed in the field, which I believe indicates that if there is any phenotype associated with this mutation it must provide very minimal benefits to the virus.

Conversely, some mutations appeared spuriously in the study without becoming fixed, but have been observed at the global epidemiological scale. The success of viruses carrying these mutations in real-world settings does not reflect their appearance within this transmission study; the two settings are substantially different. The experiment explicitly aimed to ensure EIV transmission among horses, keeping individuals indoors to encourage transmission. Hence, mutations that benefit wild viruses may be unsuitable in such tightly-controlled environments.

Were this study to be furthered, without consideration of cost or time, I would direct an investigation into measuring fitness effects of these haplotypes. Thirteen viruses, each representative of one of the reported haplotypes, would be cultured and used to establish growth curves. Measuring the replicative speed and efficiency would provide a much clearer marker of the genomic fitness that each constellation of mutations provides to the virus.

6.4.2 Genetic Linkage

Hitherto, the 21 consensus mutations and the 13 haplotypes they group into have been discussed independently. However, genetic linkage is a major consideration in evolution that warrants further investigation, which was not feasible in the present study. Though many of the mutations reported from this dataset appear only once (16 of 21 SNPs are singletons), 17 appear in conjunction with at least one other consensus mutation, either with other mutations generated *de novo* in the sample, or mutations that had prior fixed in the population. This highlights the importance of considering mutations interconnectedly; a moderately beneficial mutation may not be selected for if it is accompanied by a second mutation conferring detrimental effects to the viral genome. How then can we understand the dynamics and impacts of co-occurring mutations?

Linkage equilibria studies can elucidate the genotypic/phenotypic effects of two or more mutations independently as well as considering any interactions between them. This form of analysis can also explore the potential for genetic reassortment, whereby combinations of genetic segments from co-infecting heterogeneous parental virions can assemble within a single progeny virion. Understanding such linkage effects across the ~13kb genome was not attempted during this study; despite the potential biological relevance, nucleotide mutations appeared in frequencies too low for any study of interactivity. There are only 15

samples in this dataset that appear more than once and share more than one mutation, making any inference of mutation linkage difficult. Additionally, linkage of mutations at a scale large enough to impact viral fitness may be incredibly unlikely to occur over the 20-day period of sample collection. Finally, the short-read Illumina sequencing procedure used in the experiment made studying such linkage interactions incredibly difficult. A previous model study of such epistatic relationships between SNPs examined the inter-connectedness of both HA and NA activity on the proliferation of HA mutations (Liu et al., 2022). This example utilised an alternative deep-sequencing technology, NovaSeq 6000, followed by analyses of the read library with the variant call tool 'DeepSNV', which I reviewed in Chapter 5.4.1.

Of the eleven non-synonymous mutations observed over the course of the experiment, eight appeared only once. Many, however, were predicted to have some impact on protein structure and function. Predictions were based on physio-chemical differences between amino acid residues, spatial displacement caused by residue substitutions and mining IAV literature for annotations of homologous proteins.

Three examples of substitutions in protein functional sites appear in Polymerase Basic 1, Haemagglutinin and Neuraminidase proteins. The Gln294Arg substitution in PB1 falls in a highly exposed portion of the catalytic RNA-dependent RNA-polymerase (RdRp) region. HA Gly144Asp is the middle of a triad of residues forming antigenic site A, as labelled in human H3 proteins (Both et al., 1983; Caton et al., 1982). NA Lys342Glu is at a site that, while not directly involved in protein activity, is a critical binding site for antibodies. Antibodies raised to H11N9 viruses by mice in laboratory settings (A/Tern/Australia/G70C/1975) were observed, *in silico*, binding to the 3D neuraminidase structure, including bonds between site 342 and the antibody light-chain.

To further explore each of the mutations observed during the transmission experiment, I sought evidence of these substitutions in published EIV datasets, searching through only samples with full genome sequences. Seeking an indication of the functional or epidemiological consequences of these mutations, I instead found a complete lack of reported genomes containing these PB1 and NA mutations. Likewise, the HA mutation (Gly144Asp) was reported only once among 384 genomes. Despite putative antigenic and/or functional changes conferred by these mutations, they rarely, if ever, appeared in EIV sequences generated to date.

6.4.3 Protein Structures Predicted Well

Previously unfamiliar with the intricacies of structural biology, the opportunity to utilise new *in silico* modelling procedures enabled me to estimate the structures of all 10 major proteins in the EIV proteome. Experimental structure-resolution is an expensive, laborious process. Alternatively, homology modelling, i.e. using translated genomic sequences in conjunction with modelling software, such as the cloud-based platform ColabFold (Mirdita et al., 2022), can estimate hitherto unresolved proteins using a database of pre-existing solved structures.

At time of writing, only three EIV structures have been resolved experimentally, all of which represent haemagglutinin (A/Equine/Newmarket/2/93 [H3N8] PDB:4UNW, A/Equine/Richmond/07 [H3N8] PDB:4UO0 and

A/Equine/NY/49/73 [H7N7] PDB:6N5A). Having a full complement of the internal and external proteins of EIV would allow for further exploration of host-virus interactions and the putative phenotypic effects of non-synonymous mutations upon protein function. With the development of *in silico* modelling and machine-learning procedures, 3D protein structures can now be estimated for many viruses with relatively low computational and time costs (Abbas et al., 2023; Mirdita et al., 2022) and once developed, have been used extensively for predicting protein structures (Evans et al., 2021; Varadi et al., 2022). Such methods of predicting protein structures, based solely on genomic sequence data, have been used for an array of viral species and proteins. A meta-review of AlphaFold's usage in virology by Gutnik et al. (2023) discusses prediction of proteins from SARS-CoV-2, Mpox and HSV-1 viruses among those of numerous bacteriophages.

Having established high-confidence models of protein structures, I then moved to *in silico* testing the putative effects of non-synonymous mutations on structure and function. The impact of amino acid substitutions on local morphology, as measured by two angles of rotation (the Ramachandran angles Φ and Ψ), gave a rudimentary value of potential changes to protein function. Though only viewing morphological changes, the placement of such mutations may reveal phenotypic changes to proteins; for example, the Gly144Asp mutation observed in haemagglutinin sits within part of the cluster of residues known as antigenic site A. Due to differences in hydropathy, molecular weight and charge between the two residues, the substitution is expected to alter the plane of this site, potentially changing its' antigenic presentation. Similarly, a mutation seen in neuraminidase (Lys342Glu) lies on the surface of the protein. Though not at an active site, on searching homologous proteins, this site was shown to be part of a binding site targeted by anti-neuraminidase antibodies.

These two examples indicate that even just single point mutations to amino acids may be sufficient to influence viral fitness. Rather than focusing on the exact impact of these mutations, I highlight the potential of phenotypic mutations to arise even in short transmission experiments. Though an experimental transmission and under highly controlled conditions, this study provides evidence to suggest that detectable viral evolution can occur during even short outbreaks among small populations.

6.4.4 Summarising Consensus Findings

In the absence of immune pressures, the fastest-replicating viruses may be expected to dominate. The homogenisation of sequences seen in the unvaccinated individuals of both transmission chains indicates this. Having infected a host without a primed adaptive immune response, the fittest virus (J in the Multi and F in the Single chain) is whichever can outcompete other EIV variants. Whatever effects that the J (PB1 t1500c synonymous and a1853g/Glu618Gly substitutions) or F (NP g1445a/Ser482Asn mutation) haplotypes have on viral replication remain to be seen; this would require testing the experimental fitness of these two viruses.

Further interest lies in the synonymous mutation PA c201t. This mutation appeared independently in both transmission chains and was also recorded in 15% of global full-genome EIV sequences. Though these variants $\left(\begin{smallmatrix}199&200&\mathbf{201}\\G&A&\mathbf{C}\end{smallmatrix}\right.$ and $\left.\begin{smallmatrix}199&200&\mathbf{201}\\G&A&\mathbf{T}\end{smallmatrix}\right)$ are the only codons to possibly encode the Aspartic acid seen at residue 67, perhaps

the mutant form (GAT) is favoured in cellular translation. Other potential explanations for this apparent preference may include anti-sense codons (CTG or CTA) unfavourable to host cells, or post-translation modifications caused by this synonymous mutation. Such seemingly innocuous changes can be due to differences in host cellular machinery; any codon preferences present in host ribosomes will be imprinted onto viruses reproducing in that host. Observation of host-specific codon biases have been reported in coronaviruses (Kumar 2021), rotaviruses (Kattoor 2015) and influenza A viruses (Wen 2019). Tests for quantifying this genomic bias, including Relative Synonymous Codon Usage (RSCU) and Effective Number of Codons (ENCs), were employed in the above studies to show discrete adaptive differences in viruses in their natural hosts (e.g. avian) compared to replication in spillover hosts (e.g. swine). Viral codon usage patterns can reflect fitness adaptations; vRNA with features that are otherwise rare in host cells is more likely to be recognised as alien, possibly stimulating innate immune responses. Adopting codon usage patterns that match host cells may be a form of immune evasion by replicating viral genomes.

6.5 Quantifying Transmission Bottlenecks

Using the beta-binomial model proposed by Sobel Leonard et al. (2017), the proportion of variant reads detected in two epidemiologically-connected samples (either samples from a single individual over multiple days or a donor-recipient pair) it was possible to estimate the number of viral genomes needed to affect particular transmission events . Concurrent with other published IAV data (Dimas Martins & Gjini, 2020; Johnson & Ghedin, 2020; LeClair & Wahl, 2018; McCrone & Lauring, 2018; Sigal et al., 2018), the actual number of distinct genomes involved in a transmission event tends to be very small, at most five viruses.

Additionally, transmission bottlenecks differ between groups in the experiment with events where the donor has a history of prior EIV exposure being, on average, smaller. Whereas, when naïve hosts transmit virus, the bottlenecks include both more genomes and a slightly greater range of diversity. Extrapolating this finding, in an outbreak setting we would then assume that, like seen in the viral load, when hosts with previous exposure to EIV are involved in a transmission event, the number of viral genomes they transmit will be small and unable to fully represent the range of diversity generated in that host.

Like any bottlenecking event in nature, inter-host transmission limits the amount of diversity that can be maintained in the population as a whole. Regardless of how fit or how many advantageous mutations are able to develop over the course of a single infection, if those highly-competitive variants are unable to spread to subsequent hosts then that beneficial variant will be ultimately lost. Having observed minor differences between the bottlenecks of different classes, despite the greater diversity seen in vaccinated hosts across both groups, the low number of viable viruses that actually transmit means that much of this diversity could be lost.

6.6 Real-World EIV Epidemiology

Premises can only be infected with EIV from a limited number of routes: introduction by horse, cross-species transmission from a non-equine vector or

environmental exposure to the infectious virus. We expect most EIV outbreaks to be maintained primarily by horse-horse transmission, either a new infected individual enters the population (e.g. trading or purchasing) or a horse native to the premises acquires infection elsewhere (e.g. a sporting or other agricultural event) and then returns to the focal premises. Though fomite and environmental transmission are implicated sources of infection, the relatively short period of viability for IAV on most surfaces (up to 24 hours (Thompson & Bennett, 2017; Wißmann et al., 2021)) makes these routes more likely to contribute to infection within a premises, rather than seed infection of a new premises. Conversely, virus entering an equine population from another host species assumes circulation of EIV amongst the external (likely wild bird) population and a plethora of cross-species contact events in order for enough infectious virus to establish infection in the equine index case. In most countries where wild horses are much rarer than owned/managed horses, under both introductory circumstances, the moving horse is more likely to be vaccinated than a stationary individual, due to economic importance, sporting guidelines or trade requirements. Given this, I then assume that EIV is most commonly introduced to a premises by horses with *some* history of vaccination.

From the above analyses, we see that hosts with a history of viral exposure are more likely to a) shed lower amounts of infectious virus and b) foster an intra-host environment that drives viral diversification. Hence, the index case of each outbreak is expected to transmit a small, diverse population of viral variants. Few genomes pass through each transmission bottleneck but carry a good representation of the diversity generated in the index case. This, in a compartmental epidemiological model, would then make for a low force of transmission (β).

Importantly, this experiment investigated the effects of not only different host exposure histories on viral populations, but the knock-on effects this may cause further along a transmission chain. Despite best practices, horse populations are never going to be fully protected by EIV vaccines; whether due to unforeseen changes in circulating virus, socio-economic disparities in vaccine availability or host heterogeneities, a premises will never have complete immunity to EIV. However, they have likely been exposed to many EIV strains throughout their life. This, then amplified by inter-herd heterogeneities, creates an interconnected population of horses with varying levels of immunity and contact. Therefore, knowing how influenza A viruses are affected by the infection of hosts with adaptive immune memory of prior IAV exposure and how these changes persist (or vanish) upon infecting a host with no immunological memory can provide a realistic picture of the epidemiological landscape. Assuming that the owner/rider has economic access to routine vaccination, the immunity of each individual horse is both knowable and controllable. Without in-depth epidemiological investigation, the time since contact with an infected individual and the vaccine status of that individual plus hosts that preceded it, are unknown.

Though horses in more economically-developed countries are now rarely used as working animals, those in sporting roles often have high levels of veterinary care. They are also most likely to travel and contact individuals outside of their normal group, features traditionally associated with individuals classed as “super-spreaders” in epidemic models. Quite how EIV is affected by replication in these well-protected super-spreaders is not fully understood, but in the experiment

presented here one class of hosts modelled horses with life histories of multiple exposures, the V_M class, and so this represents a good proxy for drawing inferences on viral evolution.

6.7 An Immune System Exposed to a Plethora of Influenza Viruses

Throughout the above experiment, the horses in the "vaccinated" class had exposure histories of either one or four equine influenza strains. However, even in the Multi group, where subjects had been exposed to four different EIV strains before being challenged, all of the viruses to which they were exposed were H3N8, which shares 90-95% genetic identity with the haemagglutinin of the challenge virus. Nonetheless, phenomena that don't rely on haemagglutinins of differing groups, like epitope masking, are beginning to be understood for their potential to impair a full-strength humoral response to IAV infection. Epitope masking describes a process wherein cross-reactive antibodies physically block new antibodies from binding to viral surface proteins (Zarnitsyna et al., 2015), even though these newly developed antibodies may have better neutralising activity. B cells activated by a memory response produce antibodies much faster and in greater quantities than B cells from newly elicited clonal expansion; this epitope masking can therefore dampen the ability of newly-developed antibodies to neutralise influenza virions. Studying these interactions between hosts and pathogens on a molecular scale is only possible with high-resolution structural models, such as those I developed here. Further exploration of this immune phenomena using the protein structures presented here could aid in the development of strain-agnostic EIV vaccines.

Following from work on mice that had been exposed to different strains of mouse-adapted influenza (J. H. Kim et al., 2009b), studies have shown immune-boosting and/or reactivation of responses to the IAV strain that individual was first exposed to even when infected with an unrelated strain of IAV. Information on the strength of adaptive immune memory responses were collected from hosts modelled as having had previous exposure history (V_M and V_S classes) as serial radial haemolysis values. Due to time constraints, and a lack of similar data collected during the transmission experiment, these were not explored over the course of this thesis.

6.8 Game Theory

The incorporation of game theory into epidemiological modelling has become more common as computational biology has developed, having especially snowballed in popularity since the emergence of SARS-CoV-2. However, the majority of studies to date have focused on individuals or sub-populations of decision-making people; little attention has been directed to the study of epizootics.

Wild horse populations are far rarer than human-managed horses in the majority of countries within agricultural and/or sport settings and this creates unique epidemiological structures that translate (albeit clumsily) into game theory. Rather than human players making decisions for themselves, those people managing equine populations (e.g. farmers, jockeys, breeders etc.) make decisions on behalf of the horses, a process as yet underexplored in game theory. The decisions to vaccinate, move and report infection of horses lay with owners, which may alter the parameters

that feed into such game theory models, for example risk of infection, risks & costs of vaccination and the knowledge base (self-learned or imitation) from which decisions are made.

By measuring the impacts of adaptive immunity on viral shedding and evolution, this work aids in characterising some parameters that may influence decision-making processes of individuals (owners) and populations (via policy). Illuminating the effects of vaccination on the viral populations within-hosts (i.e. a decrease in the quantity of virus shed and application of strong selective immune pressures) should thus encourage more regular vaccine uptake. On an individual level, monetary and other costs (namely, risks of adverse reactions) of vaccination may serve to downplay the benefits of such prophylactic behaviours. Indeed, in agricultural populations reactive vaccination protocols are often employed by individuals managing horses (Wilson 2021).

6.9 Study Limitations

Of the transmission experiment itself, some methodological constraints may have limited certain aspects of the analyses. To begin with, the subjects (Welsh mountain ponies) were reared under controlled conditions and thus had known exposure histories. However, it should be appreciated that they were 24 unique individuals, bringing inherent differences in responses to infections beyond those implemented in the exposure programme. Individual differences in innate, and potentially even intrinsic, immunity could vastly alter the host environments in which viruses find themselves, leading to different evolutionary drivers. This is especially evident when considering NS1, a viral immune-deregulatory protein that acts to interfere with intrinsic cellular immunity. Heterogeneity amongst the cellular machinery of individuals could, thus, impact EIV fitness.

Secondly, a missed opportunity from this experiment was sampling exclusively from the nasal mucosa. Pathogen populations can be keenly influenced by the spatial heterogeneity present within hosts; a common example is the tropism for lower vs upper respiratory tracts evident in influenza infections. Partitioning by differing tropisms can cause populations to diverge; without input from each other they can create very different population structures even if the microclimates provide the exact same conditions and selective pressures. To demonstrate the potential for such variety in horse respiratory tracts, veterinary clinicians recommend a 2-3 metre endoscope for equine bronchoscopy – along which conditions (such as the temperature, humidity and presentation of immune cells and molecules to name a few) can vary greatly for colonising viruses. This, therefore, can mean that EIV populations within a host can face wholly different adaptive landscapes, driving evolution in separate, distinct directions. Of course, daily sampling of multiple areas of a host's airway is untenable due to welfare reasons.

There was no recording of symptoms while horses were observed and, additionally, the pairing of individuals for purposes of welfare unfortunately interferes somewhat with details of the transmission events. Though the direction of the transmission chain is known, having four hosts present at each transmission event meant simplifying the viruses in paired hosts X_A and X_B to have acted as one population (X). This then omits the possibility of one host being infected by two (or even three) individuals and assumes that both donor hosts transmitted the virus

evenly and in identical proportions. Clearly, this is unrepresentative of real epidemic dynamics, and somewhat confounds estimations of transmission bottleneck sizes.

Sera samples were collected from the vaccinated hosts for measuring the strength of their responses during the regiment of exposure to viruses. Unfortunately, due to time constraints of the study and the low quality of these samples, this avenue of research was explored only at the shallowest level.

Like much predictive modelling, the protein structural estimations and associated *in silico* mutagenesis experiments were never qualified or verified with any *in vitro* studies. However, to account for this I did try to compare the models with actual lab experiments whenever possible to at least add some credence to the claims put forth, such as the comparison of predicted H3N8 protein properties with those of IAV proteins quantified *in vitro*, and the attempts to contrast 3D structural models with fully-resolved crystal structures of proteins from other IAV.

As discussed above, fomites and environmental transmission of EIV can play a role in epidemic dynamics within premises. Hence, having the subjects held together in a single compound may lead to slightly overinflated values that would only be observed in cases where horses are in constant indoor contact. This perhaps explains continual transmission events, even from hosts that shed very little amounts of virus.

7 Closing Remarks

Altogether, I present an intricate look at the evolutionary dynamics of EIV through short, experimental transmission chains at both the consensus and sub-consensus level, further supported by daily quantification of the viral populations shed by hosts. This comprehensive view of viruses as they experience transmission bottlenecks moving between hosts allows for observation of the ways in which diversity generated in one viral genome can be maintained or removed from the population *en masse*. Furthermore, differing host environments were created by exposing some hosts to whole, inactivated virus in order to stimulate immunological memory, simulating having previously encountered such viruses. Whether hosts experienced these simulated life-history exposures to influenza viruses, and if so by how much did these previous immunogens differ from the challenge virus, created a three-class system in which to analyse viral evolution.

Appendices

A)

```
> summary(mod_sumlogCopies_split)

Model Info:
 function:     stan_gamm4
 family:       gaussian [identity]
 formula:      totCopies ~ s(dpc, k = 8) + Status + group
 algorithm:    sampling
 sample:       50000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 36

Estimates:
                    mean    sd    10%   50%   90%
(Intercept)         11.1    0.7  10.2  11.1  12.1
StatusVacc           1.6    0.9   0.5   1.6   2.6
groupSin            -2.1    0.9  -3.2  -2.1  -1.1
s(dpc).1           -33.5   23.2 -63.6 -33.1  -4.1
s(dpc).2            48.1   17.1  25.9  48.5  69.4
s(dpc).3            57.3   16.1  36.3  58.0  77.1
s(dpc).4           -20.2    5.8 -27.5 -20.2 -12.7
s(dpc).5             7.3    5.0   0.8   7.4  13.7
s(dpc).6            10.0    9.2  -1.9  10.3  21.4
s(dpc).7           -30.7   17.9 -53.6 -31.0  -6.6
sigma                2.5    0.4   2.1   2.5   3.0
smooth_sd[s(dpc)1]  47.1   22.5  24.9  42.7  73.8
smooth_sd[s(dpc)2] 119.9  109.0  21.0  84.2 271.8
```

```
Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 10.8    0.6 10.1  10.8  11.6

MCMC diagnostics
                    mcse Rhat n_eff
(Intercept)         0.0  1.0  36422
StatusVacc          0.0  1.0  36413
groupSin            0.0  1.0  37329
s(dpc).1            0.2  1.0  21699
s(dpc).2            0.1  1.0  20441
s(dpc).3            0.1  1.0  15009
s(dpc).4            0.0  1.0  31730
s(dpc).5            0.0  1.0  20009
s(dpc).6            0.1  1.0  17793
s(dpc).7            0.1  1.0  16186
sigma               0.0  1.0  14298
smooth_sd[s(dpc)1] 0.2  1.0   8152
smooth_sd[s(dpc)2] 0.9  1.0  14472
mean_PPD            0.0  1.0  42107
log-posterior       0.1  1.0   6091
```

Status	Group	DPC								
		0	1	2	3	4	5	6	7	8
Naive	Mul	4.037	2.825	13.496	19.132	18.026	17.290	17.128	5.571	2.735
Naive	Sin	1.601	3.390	17.005	17.886	13.994	10.178	13.132	3.188	0.479
Vacc	Mul	3.066	2.589	22.052	15.681	19.004	20.615	16.491	11.939	2.760
Vacc	Sin	2.459	2.144	19.341	15.641	15.096	18.287	13.931	7.194	0.894

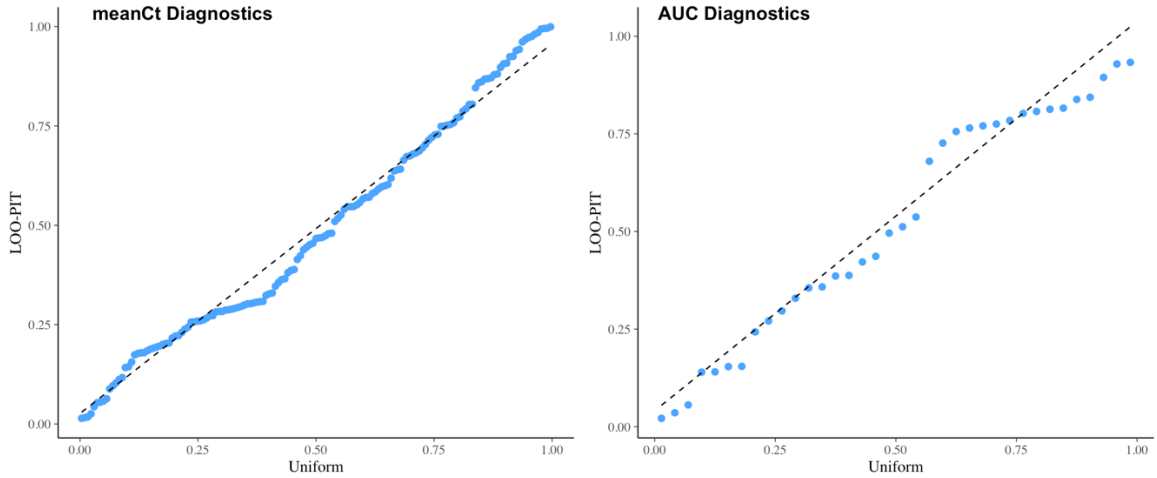
B)

Model 1: mean Ct values ~ Status + Group + Days Post-contact (smoothed with 8 pivots)

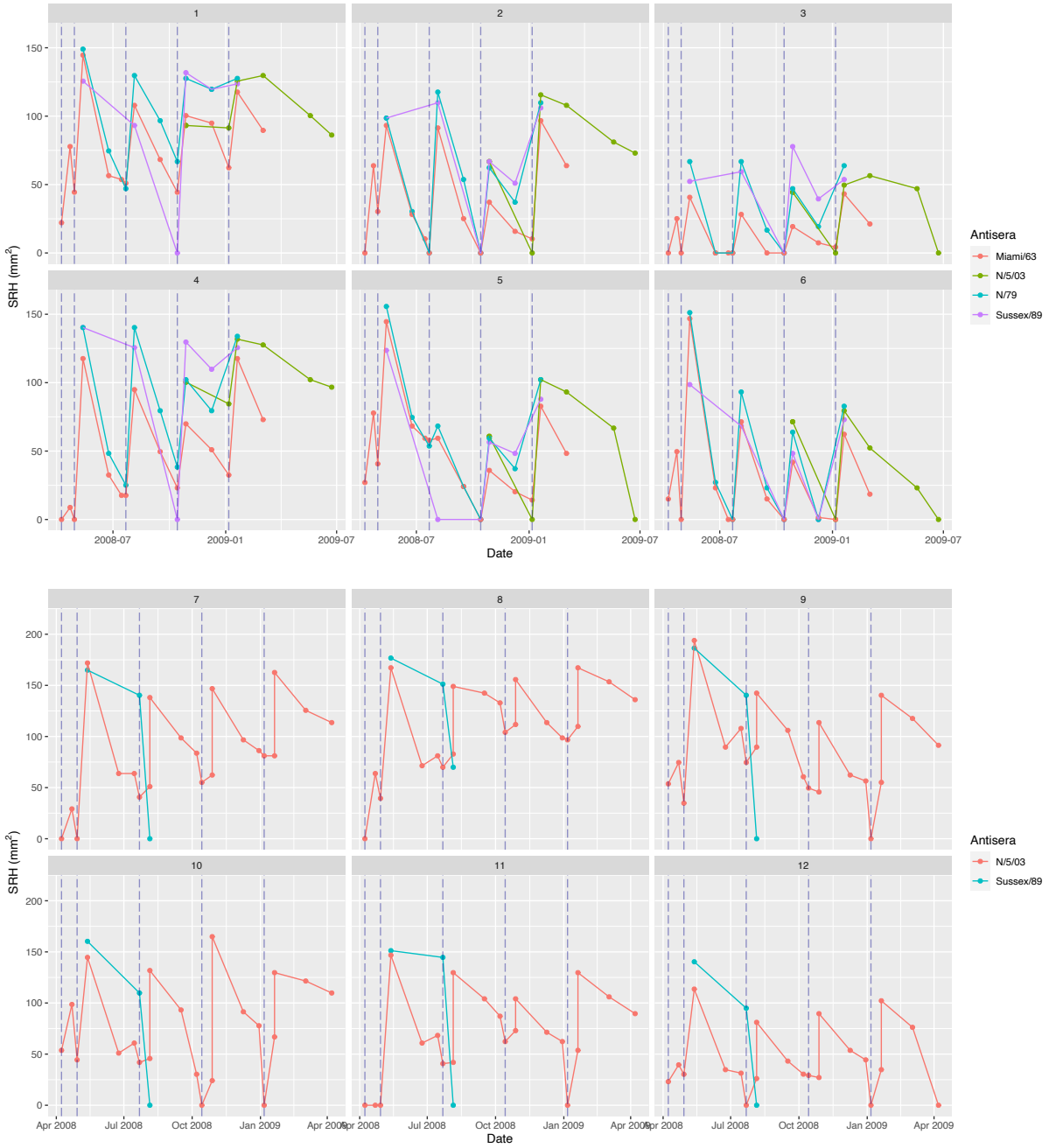
LOOIC model 1: 890.6

Model 2: total Ct (AUC) ~ Status + Group + Days Post-contact (smoothed with 8 pivots)

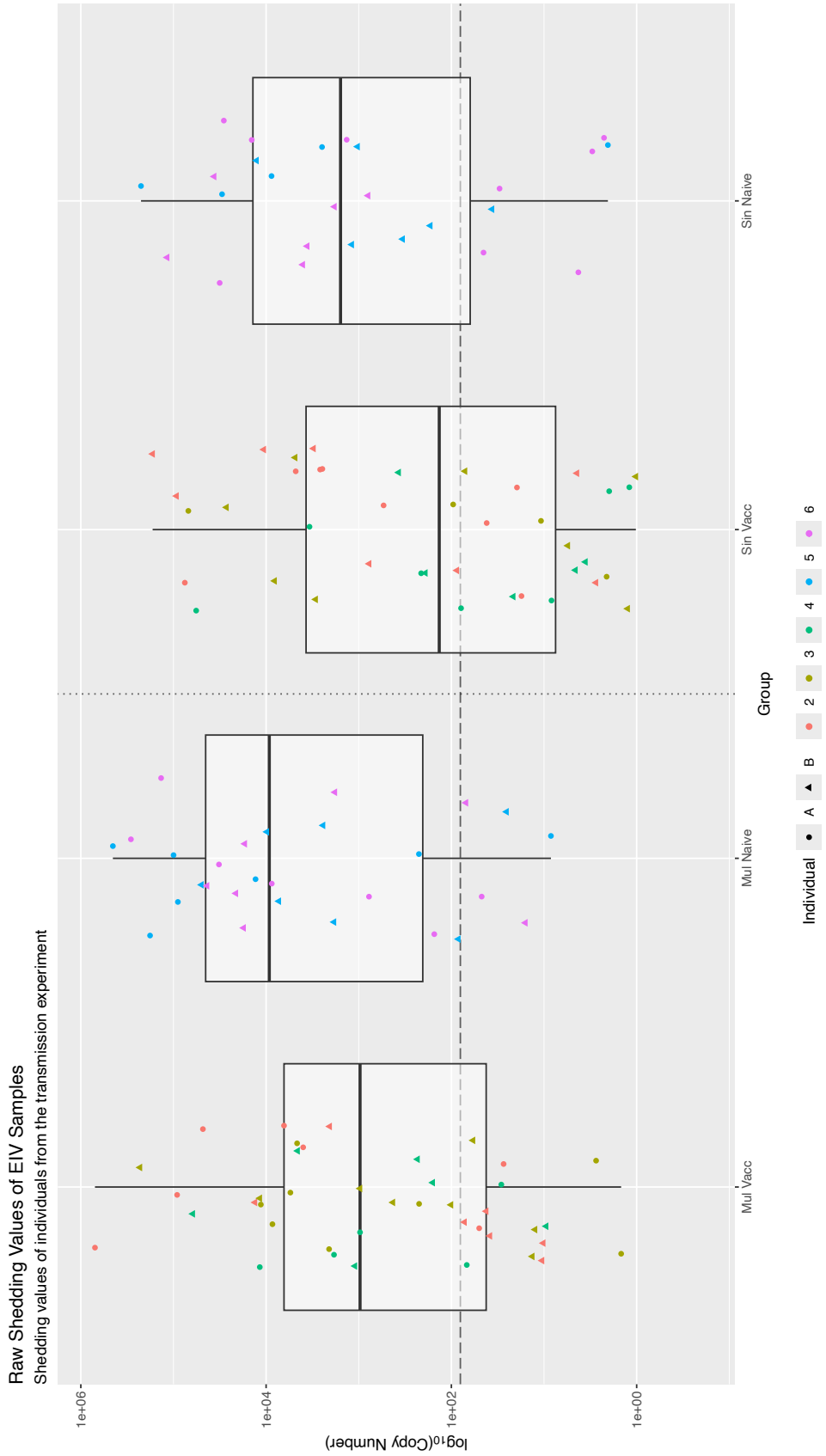
LOOIC model 2: 395.7



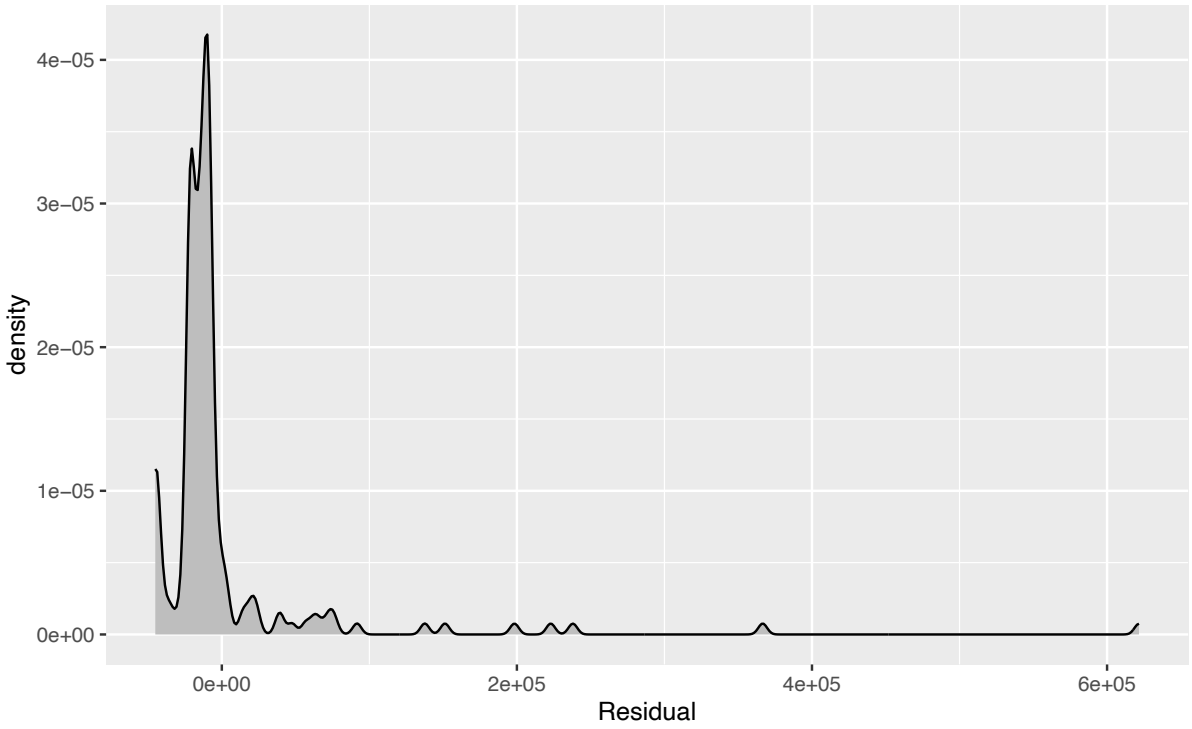
Supplementary 2.1: A) Regression Table and raw data used in the GAM analysis of viral shedding. B) In determining whether to use the average (meanCt) or the summed (AUC) amount of virus shed when modelling, models were constructed in parallel and then compared with a Leave-One-Out Information Criterion (LOOIC). Diagnostic plots show the distribution of residuals.



Supplementary 3.1: Serial Radial Haemolysis (SRH) experiments from vaccinated horses in the Multi (top) and Single (bottom) transmission groups. Dashed lines denote the date of each vaccination. Each point shows the degree of circulating anti-EIV antibodies as represented by the size of SRH plaques in response to exposing antisera to viral cultures.



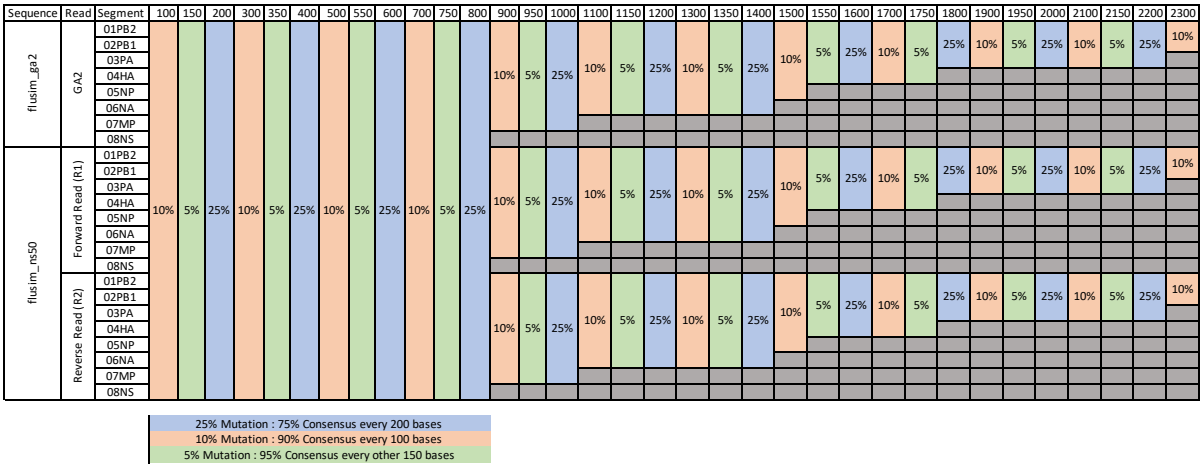
Supplementary Figure 3.2: Viral loads obtained from each nasal swab, coloured for the host's position in the transmission chain. Shapes differentiate paired hosts from each other (A or B).



Supplementary 3.3: Plot of the residuals from models created using the mean copy numbers of samples, demonstrating non-normality in the distribution of residuals. In addition, skew was calculated at 5.83 and the kurtosis of the curve was 42.58. Hence, non-parametric tests were used in the analyses that followed.

Protein	UniProt ID	UniProt Accession	PDB Reference
01PB2	>sp|P03429	P26105	6QNW_3
02PB1	>sp|P03432	P16505	6QNW_2
03PA	>sp|P03429	P13169	6QNW_1
04HA	>tr|Q82847	P17001	4UNW
05NP	>tr|Q1K9H2	P67915	2IQH
06NA	>sp|A0A0C4WXC5	Q07582	5HUK
07M1	>sp|P03485	Q77ZK7	1EA3
07M2	>sp|P0DOF5	Q77ZK8	2L0J
08NEP	>sp|P03508	Q77ZM4	1PD3
08NS1	>tr|Q20NS3	Q20NS3	4OPH

Supplementary 4.1: Published proteins that were used to map regions of the proteins translated from sequence data collected in the transmission experiment.



Supplementary 5.1: The simulated genomes used to test Variant Call Tools. Nucleotides are numbered on the first row and cells are filled and labelled to show the frequency of mutant reads compared to consensus reads.

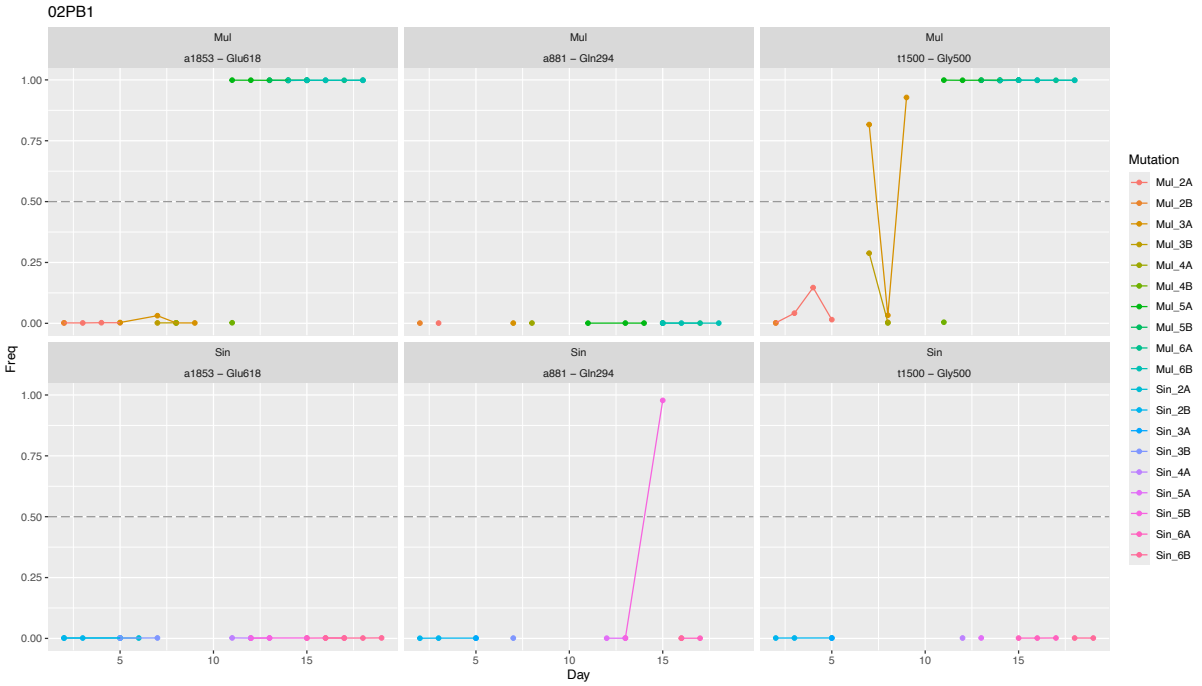
Dataset	Bioproject	Reference Genome	NCBI taxID	Sequence
SimData	-	A/Equine/Newmarket/5/03 (H3N8)	568375	ga2
SimData	-	A/Equine/Newmarket/5/03 (H3N8)	568375	ns50
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3359624
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3359625
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3359626
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3359627
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3359628
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3360141
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3360142
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3360143
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3360144
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3360149
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3360151
McCrone2016	PRJNA317621	A/WSN/1933 (H1N1)	382835	SRR3360152
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121274
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121281
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121301
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121368
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121380
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121409
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121620
McCrone2018	PRJNA412631	A/New York/WC-LVD-15-031/2015 (H3N2)	1895544	SRR6121630
Han2021	PRJNA722099	A/Brisbane/10/2007 (H3N2)	476294	SRR14242319
Han2021	PRJNA722099	A/Brisbane/10/2007 (H3N2)	476294	SRR14242328
Han2021	PRJNA722099	A/Brisbane/10/2007 (H3N2)	476294	SRR14242338
Han2021	PRJNA722099	A/Brisbane/10/2007 (H3N2)	476294	SRR14242374

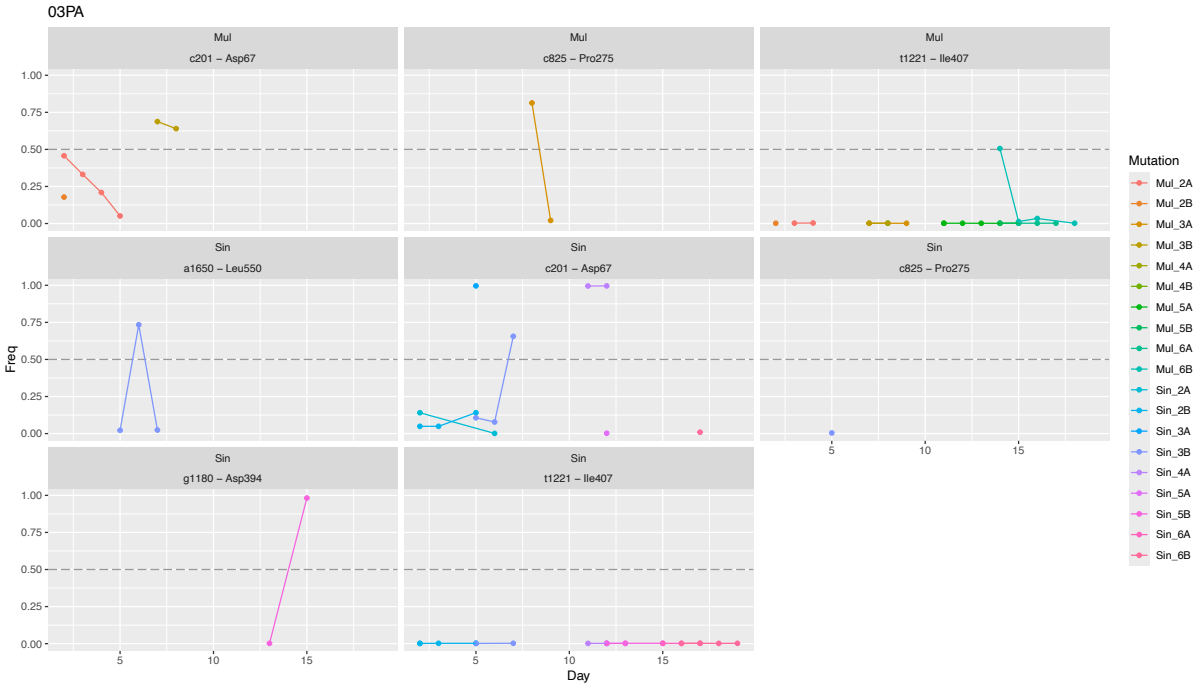
Han2021	PRJNA722099	A/California/04/2009 (H1N1)	641501	SRR6121449
Han2021	PRJNA722099	A/California/04/2009 (H1N1)	641501	SRR6121456
Han2021	PRJNA722099	A/California/04/2009 (H1N1)	641501	SRR6121594
Han2021	PRJNA722099	A/California/04/2009 (H1N1)	641501	SRR6121605
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443362
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443363
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443366
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443370
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443375
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443376
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443382
Poelvoorde2022	PRJNA692424	A/Victoria/1003/2012 (H3N2)	2044087	SRR13443383
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443356
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443379
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443387
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443390
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443391
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443394
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443397
Poelvoorde2022	PRJNA692424	A/Bretagne/7608/2009 (H1N1)	1506405	SRR13443399

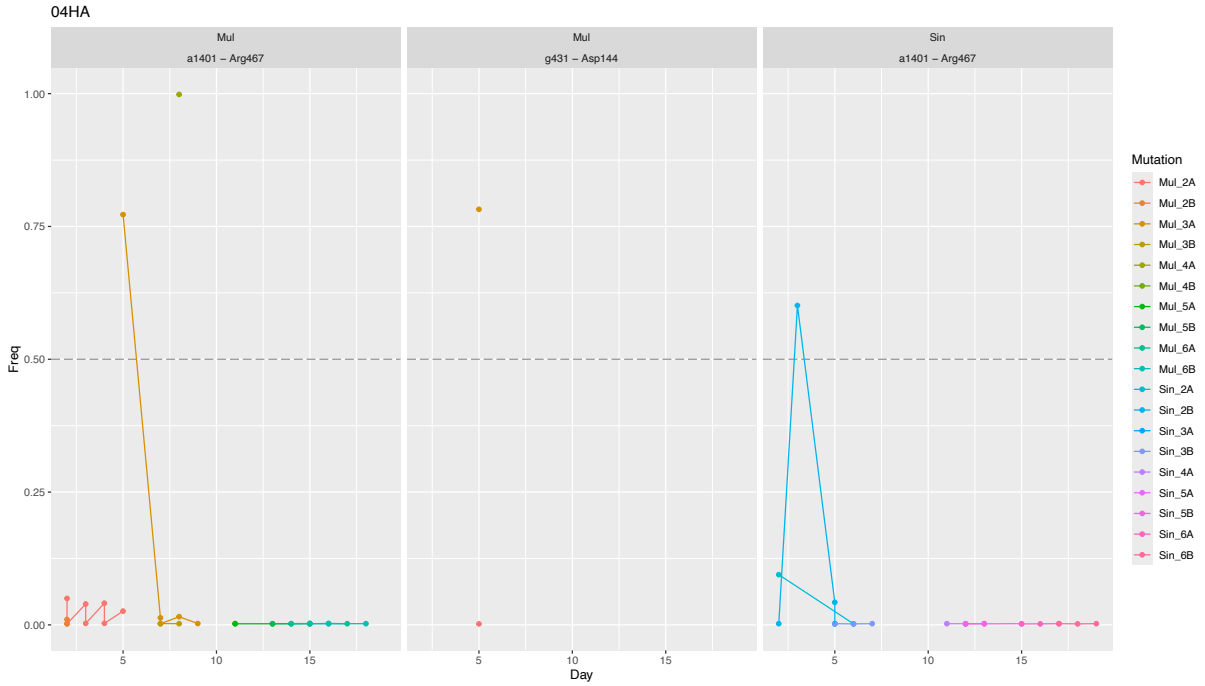
Supplementary 5.2: Sequences used to test Variant Call Tools, obtained from previously published data or sequences using the ART simulator.

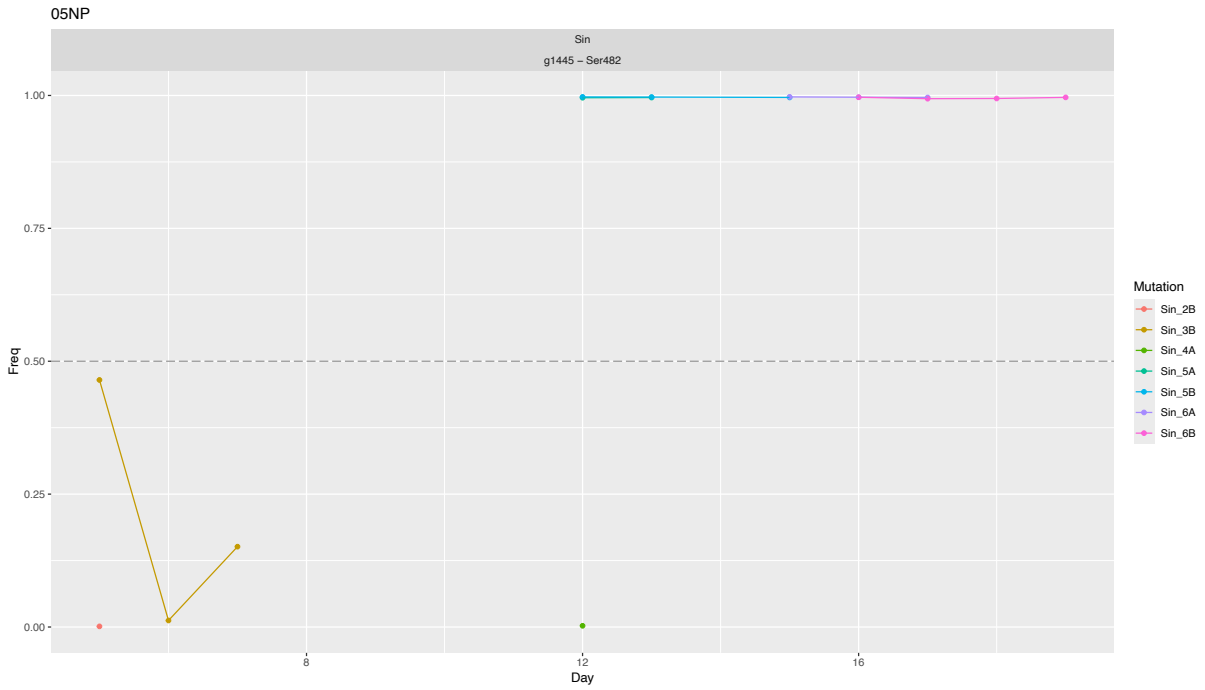
Supplementary 5.3: Mutation sub-consensus frequency seen on genomic segments 2-7. Graphs attempt to show the trajectory of mutations throughout the experiment, hence only mutations that broach the consensus level (as illustrated by dashed line) were examined. Further, mutations that appear in consensus sequences but are only detectable within the sub-consensus reads on that day are not shown. Both segment 8 mutations (110 and 113) for example were only seen above the limit of detection on the day in which they appeared at the consensus level.

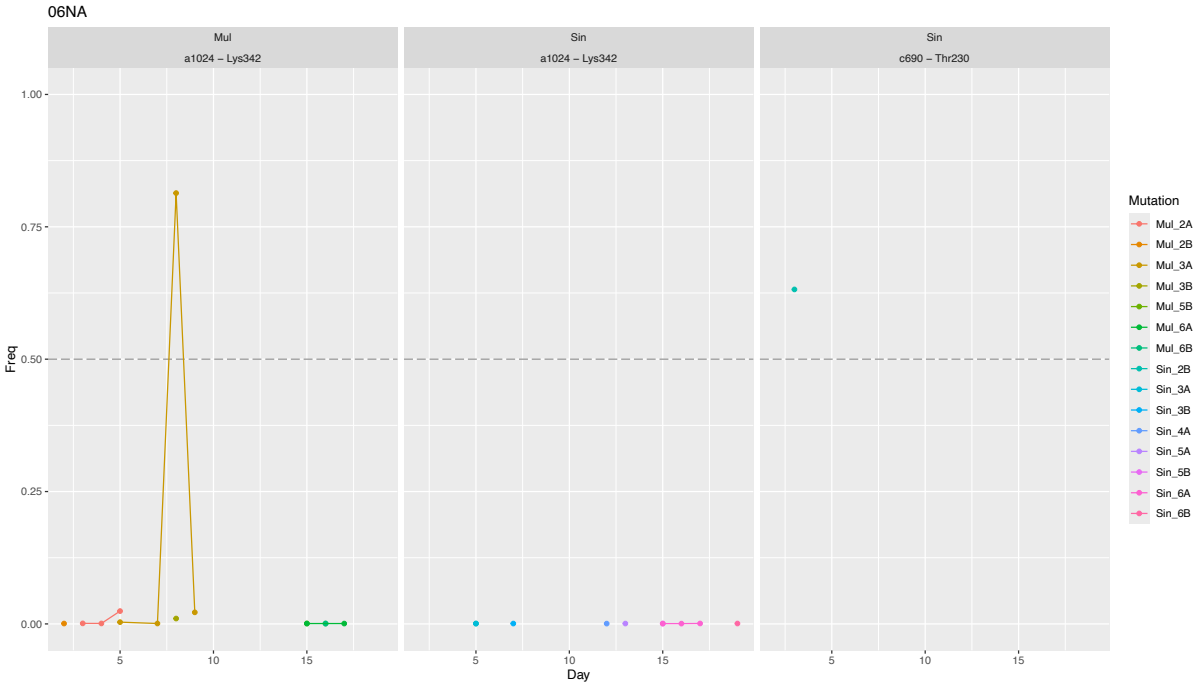
7

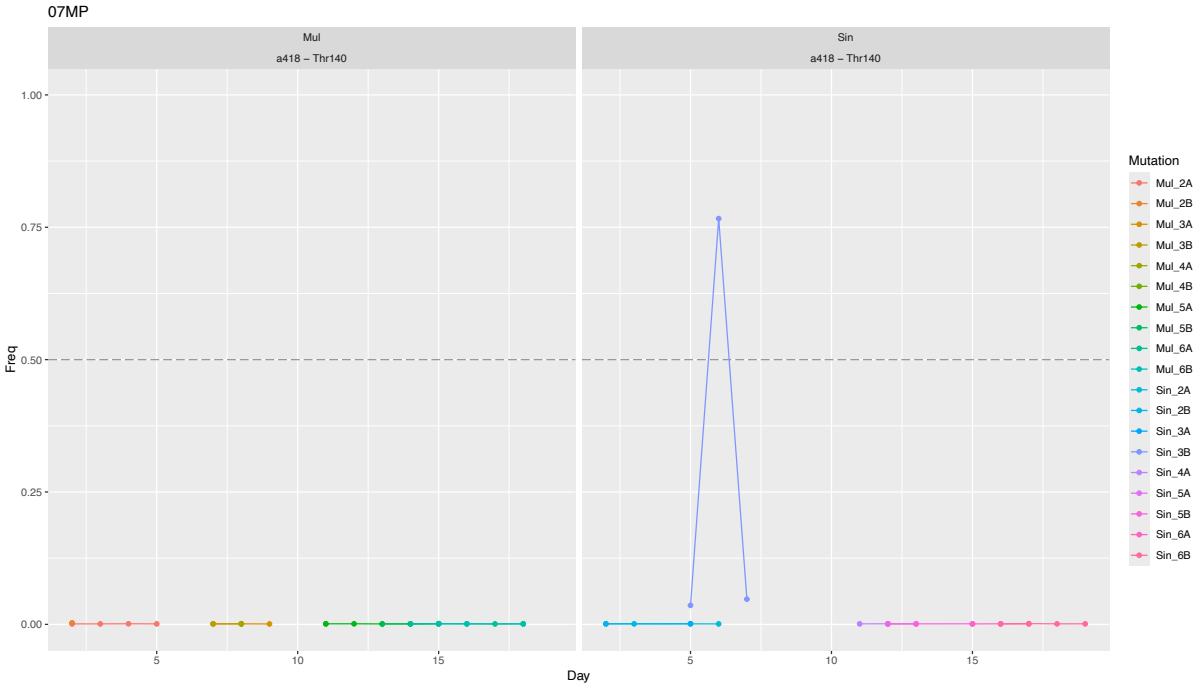

02PB1
Mul
a1853 – Glu618
Mul
a881 – Gln294
Mul
t1500 – Gly500
Sin
a1853 – Glu618
Sin
a881 – Gln294
Sin
t1500 – Gly500
Freq
1.00
0.75
0.50
0.25
0.00
5
10
15
Day
Mutation
Mul_2A
Mul_2B
Mul_3A
Mul_3B
Mul_4A
Mul_4B
Mul_5A
Mul_5B
Mul_6A
Mul_6B
Sin_2A
Sin_2B
Sin_3A
Sin_3B
Sin_4A
Sin_5A
Sin_5B
Sin_6A
Sin_6B


03PA
Mul
c201 – Asp67
Mul
c825 – Pro275
Mul
t1221 – Ile407
Sin
a1650 – Leu550
Sin
c201 – Asp67
Sin
c825 – Pro275
Sin
g1180 – Asp394
Sin
t1221 – Ile407
Freq
1.00
0.75
0.50
0.25
0.00
5
10
15
Day
Mutation
Mul_2A
Mul_2B
Mul_3A
Mul_3B
Mul_4A
Mul_4B
Mul_5A
Mul_5B
Mul_6A
Mul_6B
Sin_2A
Sin_2B
Sin_3A
Sin_3B
Sin_4A
Sin_5A
Sin_5B
Sin_6A
Sin_6B


04HA
Mul
a1401 – Arg467
Mul
g431 – Asp144
Sin
a1401 – Arg467
Freq
1.00
0.75
0.50
0.25
0.00
5
10
15
Day
Mutation
Mul_2A
Mul_2B
Mul_3A
Mul_3B
Mul_4A
Mul_4B
Mul_5A
Mul_5B
Mul_6A
Mul_6B
Sin_2A
Sin_2B
Sin_3A
Sin_3B
Sin_4A
Sin_5A
Sin_5B
Sin_6A
Sin_6B


05NP
Sin
g1445 – Ser482
Freq
1.00
0.75
0.50
0.25
0.00
8
12
16
Day
Mutation
Sin_2B
Sin_3B
Sin_4A
Sin_5A
Sin_5B
Sin_6A
Sin_6B


06NA
Mul
a1024 – Lys342
Sin
a1024 – Lys342
Sin
c690 – Thr230
Freq
1.00
0.75
0.50
0.25
0.00
5
10
15
Day
Mutation
Mul_2A
Mul_2B
Mul_3A
Mul_3B
Mul_5B
Mul_6A
Mul_6B
Sin_2B
Sin_3A
Sin_3B
Sin_4A
Sin_5A
Sin_5B
Sin_6A
Sin_6B


07MP
Mul
a418 – Thr140
Sin
a418 – Thr140
Freq
1.00
0.75
0.50
0.25
0.00
5
10
15
Day
Mutation
Mul_2A
Mul_2B
Mul_3A
Mul_3B
Mul_4A
Mul_4B
Mul_5A
Mul_5B
Mul_6A
Mul_6B
Sin_2A
Sin_2B
Sin_3A
Sin_3B
Sin_4A
Sin_5A
Sin_5B
Sin_6A
Sin_6B

List of References

Abbas, U. L., Chen, J., & Shao, Q. (2023). Assessing Fairness of AlphaFold2 Prediction of Protein 3D Structures. *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 1–10. https://doi.org/10.1145/3584371.3612943

Abdel-Moneim, A. S., Abdel-Ghany, A. E., & Shany, S. A. (2010). Isolation and characterization of highly pathogenic avian influenza virus subtype H5N1 from donkeys. *Journal of Biomedical Science*, *17*(1), 25. https://doi.org/10.1186/1423-0127-17-25

Abdel-Moneim, A. S., Shehab, G. M., & Abu-Elsaad, A.-A. S. (2011). Molecular evolution of the six internal genes of H5N1 equine influenza A virus. *Arch Virol*, *156*(7), 1257–1262. https://doi.org/10.1007/s00705-011-0966-3

Adeyefa, C. A. O., James, M. L., & Mccauley, J. W. (1996). Antigenic and genetic analysis of equine influenza viruses from tropical Africa in 1991. *Epidemiol Infect*, *117*(2), 367–374. https://doi.org/10.1017/s0950268800001552

Aeschbacher, S., Santschi, E., Gerber, V., Stalder, H. P., & Zanoni, R. G. (2015). Development of a real-time RT-PCR for detection of equine influenza virus. *Schweizer Archiv Fur Tierheilkunde*, *157*(4), 191–201. https://doi.org/10.17236/sat00015

Alford, R. H., Kasel, J. A., Lehrich, J. R., & Knight, V. (1967). Human Responses to Experimental Infection with Influenza A/EQUI 2 Virus. *American Journal of Epidemiology*, *86*(1), 185–192. https://doi.org/10.1093/oxfordjournals.aje.a120723

Alves Beuttemmüller, E., Woodward, A., Rash, A., Dos Santos Ferraz, L. E., Fernandes Alfieri, A., Alfieri, A. A. A. F. A. A., Elton, D., Beuttemmüller, E. A., Woodward, A., Rash, A., Eduardo, L., Alfieri, A. A. A. F. A. A., Alfieri, A. A. A. F. A. A., Elton, D., Alves Beuttemmüller, E., Woodward, A., Rash, A., Dos Santos Ferraz, L. E., Fernandes Alfieri, A., … Elton, D. (2016). Characterisation of the epidemic strain of H3N8 equine influenza virus responsible for outbreaks in South America in 2012. *Virology Journal*, *13*(1), 45. https://doi.org/10.1186/s12985-016-0503-9

Amat, J. A. R., Patton, V., Chauché, C., Goldfarb, D., Crispell, J., Gu, Q., Coburn, A. M., Gonzalez, G., Mair, D., Tong, L., Martinez-Sobrido, L., Marshall, J. F., Marchesi, F., & Murcia, P. R. (2021). Long-term adaptation following influenza A virus host shifts results in increased within-host viral fitness due to higher replication rates, broader dissemination within the respiratory epithelium and reduced tissue damage. *PLOS Pathogens*, *17*(12), 1–25. https://doi.org/10.1371/journal.ppat.1010174

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. In *Nature Medicine* (Vol. 26, Issue 4, pp. 450–452). https://doi.org/10.1038/s41591-020-0820-9

Andrew, M. K., Pott, H., Staadegaard, L., Paget, J., Chaves, S. S., Ortiz, J. R., McCauley, J., Bresee, J., Nunes, M. C., Baumeister, E., Raboni, S. M., Giamberardino, H. I. G., McNeil, S. A., Gomez, D., Zhang, T., Vanhems, P., Koul, P. A., Coulibaly, D., Otieno, N. A., … Lina, B. (2023). Age Differences in Comorbidities, Presenting Symptoms, and Outcomes of Influenza Illness Requiring Hospitalization: A Worldwide Perspective From the Global Influenza Hospital Surveillance Network. *Open Forum Infectious Diseases*, *10*(6), 1–10. https://doi.org/10.1093/ofid/ofad244

Aoki, F. Y., & Boivin, G. (2009). Influenza virus shedding—Excretion patterns and effects of antiviral treatment. *Journal of Clinical Virology*, *44*(4), 255–261. https://doi.org/10.1016/J.JCV.2009.01.010

Arevalo, C. P., Le Sage, V., Bolton, M. J., Eilola, T., Jones, J. E., Kormuth, K. A., Nturibi, E., Balmaseda, A., Gordon, A., Lakdawala, S. S., & Hensley, S. E. (2020). Original antigenic sin priming of influenza virus hemagglutinin stalk antibodies. *Proceedings of the National Academy of Sciences*, *117*(29), 17221–17227. https://doi.org/10.1073/pnas.1920321117

Armero, A., Berthetm, N., & Avarre, J.C. (2021). Intra-Host Diversity of SARS-CoV-2 Should Not Be Neglected: Case of the State of Victoria, Australia. *Viruses*, *13*(1), 133. https://doi.org/10.3390/v13010133

Arnold, M. (2021). *mattarnoldbio/alphapickle: v1.4.1*. Zenodo. https://doi.org/10.5281/zenodo.5752375

Back, H., Treiberg, L., Gröndahl, G., Ståhl, K., Pringle, J., Zohari, S., Berndtsson, L. T., Gröndahl, G., Ståhl, K., Pringle, J., & Zohari, S. (2016). The first reported Florida clade 1 virus in the Nordic countries, isolated from a Swedish outbreak of equine influenza in 2011. *Vet Microbiol*, *184*, 1–6. https://doi.org/10.1016/j.vetmic.2015.12.010

Baele, G., Suchard, M. A., Bielejec, F., & Lemey, P. (2016). Bayesian codon substitution modelling to identify sources of pathogen evolutionary rate variation. *Microbial Genomics*. https://doi.org/10.1099/mgen.0.000057

Balasuriya, U. B. R. (2020). RNA extraction from equine samples for equine influenza virus. In *Methods in Molecular Biology* (Vol. 2123, pp. 369–382). Humana Press Inc. https://doi.org/10.1007/978-1-0716-0346-8_28

Bean, B., Moore, B. M., Sterner, B., Peterson, L. R., Gerding, D. N., & Balfour, H. H. (1982). Survival of Influenza Viruses on Environmental Surfaces. *The Journal of Infectious Diseases*, *146*(1), 47–51. http://www.jstor.org/stable/30109645

Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Dallas, T. A., Eskew, E. A., Farrell, M. J., Guth, S., Han, B. A.,

Simmons, N. B., & Carlson, C. J. (2020). Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization. *BioRxiv Preprint*, *May 23*, 1–47. https://doi.org/10.1101/2020.05.22.111344

Bell, D., Nicoll, A., Fukuda, K., Horby, P., Monto, A., Hayden, F., Wylks, C., Sanders, L., & Van Tam, J. (2006). Non-pharmaceutical interventions for pandemic influenza, international measures. *Emerging Infectious Diseases*, *12*(1), 81–87. https://doi.org/10.3201/EID1201.051370

Belser, J. A., Eckert, A. M., Huynh, T., Gary, J. M., Ritter, J. M., Tumpey, T. M., & Maines, T. R. (2020). A Guide for the Use of the Ferret Model for Influenza Virus Infection. In *American Journal of Pathology* (Vol. 190, Issue 1, pp. 11–24). Elsevier Inc. https://doi.org/10.1016/j.ajpath.2019.09.017

Bendall, E. E., Callear, A. P., & Getz, A. (2023). Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible SARS-CoV-2 variants. *Nat Commun*, *14*, 272. https://doi.org/10.1038/s41467-023-36001-5

Bera, B. C., Virmani, N., Kumar, N., Anand, T., Pavulraj, S., Rash, A., Elton, D., Rash, N., Bhatia, S., Sood, R., Singh, R. K., & Tripathi, B. N. (2017). Genetic and codon usage bias analyses of polymerase genes of equine influenza virus and its relation to evolution. *BMC Genomics*, *18*(1), 652. https://doi.org/10.1186/s12864-017-4063-1

Bergstrom, C. T., McElhany, P., & Real, L. A. (1999). Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proceedings of the National Academy of Sciences*, *96*(9), 5095–5100. https://doi.org/10.1073/pnas.96.9.5095

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bessière, P., & Volmer, R. (2021). From one to many: The within-host rise of viral variants. *PLOS Pathogens*, *17*(9), e1009811. https://doi.org/10.1371/journal.ppat.1009811

Bessonov, N., Bocharov, G., Meyerhans, A., Popov, V., & Volpert, V. (2020). Nonlocal Reaction–Diffusion Model of Viral Evolution: Emergence of Virus Strains. *Mathematics*, *8*(1), 117. https://doi.org/10.3390/math8010117

BETA. (2024). *National Equestrian Form (NEF19)*. https://beta-uk.org/equestrian-trade-news/

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., & Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology and Evolution*, *30*(6), 306–313. https://doi.org/10.1016/j.tree.2015.03.009

Biek, R., & Real, L. A. (2010). The landscape genetics of infectious disease emergence and spread. *Molecular Ecology*, *19*(17), 3515–3531. https://doi.org/10.1111/j.1365-294X.2010.04679.x

Blanco-lobo, P., Rodriguez, L., Reedy, S., Oladunni, F. S., Nogales, A., Murcia, P. R., Chambers, T. M., & Martinez-Sobrido, L. (2019). A bivalent live-attenuated vaccine for the prevention of equine influenza virus. *Viruses*, *11*(10), 933. https://doi.org/10.3390/v11100933

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boni, M. F., de Jong, M. D., van Doorn, H. R., & Holmes, E. C. (2010). Guidelines for identifying homologous recombination events in influenza A virus. *PloS One*, *5*(5), e10434. https://doi.org/10.1371/journal.pone.0010434

Bonilla, F. A., & Oettgen, H. C. (2010). Adaptive immunity. *Journal of Allergy and Clinical Immunology*, *125*(2, Supplement 2), S33–S40. https://doi.org/https://doi.org/10.1016/j.jaci.2009.09.017

Borchers, K., Daly, J., Stiens, G., Kreling, K., Kreling, I., & Ludwig, H. (2005). Characterisation of three equine influenza A H3N8 viruses from Germany (2000 and 2002): evidence for frozen evolution. *Vet Microbiol*, *107*(1–2), 13–21. https://doi.org/10.1016/j.vetmic.2005.01.010

Both, G. W., Sleigh, M. J., Cox, N. J., & Kendal, A. P. (1983). Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *Journal of Virology*, *48*(1), 52–60. https://doi.org/10.1128/jvi.48.1.52-60.1983

Boukharta, M., Zakham, F., Touil, N., Elharrak, M., & Ennaji, M. M. (2014). Cleavage site and Ectodomain of HA2 sub-unit sequence of three equine influenza virus isolated in Morocco. *BMC Res Notes*, *7*, 448. https://doi.org/10.1186/1756-0500-7-448

Bountouri, M., Fragkiadaki, E., Ntafis, V., Kanellos, T., & Xylouri, E. (2011). Phylogenetic and molecular characterization of equine H3N8 influenza viruses from Greece (2003 and 2007): evidence for reassortment between evolutionary lineages. *Virol J*, *8*, 350. https://doi.org/10.1186/1743-422X-8-350

Bouvier, N. M., & Palese, P. (2008). The biology of influenza viruses. *Vaccine*, *26*(SUPPL. 4), D49–D53. https://doi.org/10.1016/J.VACCINE.2008.07.039

Bryant, N. A., Rash, A. S., Woodward, A. L., Medcalf, E., Helwegen, M., Wohlfender, F., Cruz, F., Herrmann, C., Borchers, K., Tiwari, A., Chambers, T. M., Newton, J. R., Mumford, J. A., & Elton, D. M. (2011). Isolation and characterisation of equine influenza viruses (H3N8) from Europe and North America from 2008 to 2009. *Vet Microbiol*, *147*(1–2), 19–27. https://doi.org/10.1016/j.vetmic.2010.05.040

Callinan, I. D. F. (2008). Report of the Equine Influenza Inquiry. In *Equine influenza : the August 2007 outbreak in Australia* (p. 345). Commonwealth of Australia [Canberra]. https://nla.gov.au/nla.obj-961843962

Campbell, F., Cori, A., Ferguson, N., & Jombart, T. (2019). Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Computational Biology*, *15*(3). https://doi.org/10.1371/journal.pcbi.1006930

Campbell, F., Strang, C., Ferguson, N., Cori, A., & Jombart, T. (2018a). When are pathogen genome sequences informative of transmission events? *PLoS Pathogens*, *14*(2), e1006885. https://doi.org/10.1371/journal.ppat.1006885

Campbell, F., Strang, C., Ferguson, N., Cori, A., & Jombart, T. (2018b). When are pathogen genome sequences informative of transmission events? *PLoS Pathogens*, *14*(2), e1006885. https://doi.org/10.1371/journal.ppat.1006885

Canini, L., Holzer, B., Morgan, S., Hemmink, J. D., Clark, B., Woolhouse, M. E. J., Tchilian, E., & Charleston, B. (2020). Timelines of infection and transmission dynamics of H1N1pdm09 in swine. *PLoS Pathogens*, *16*(7). https://doi.org/10.1371/journal.ppat.1008628

Cassini, A., Colzani, E., Pini, A., Mangen, M.-J. J., Plass, D., McDonald, S. A., Maringhini, G., van Lier, A., Haagsma, J. A., Havelaar, A. H., Kramarz, P., & Kretzschmar, M. E. (2018). Impact of infectious diseases on population health using incidence-based disability-adjusted life years (DALYs): results from the Burden of Communicable Diseases in Europe study, European Union and European Economic Area countries, 2009 to 2013. *Eurosurveillance*, *23*(16), 1–20. https://doi.org/10.2807/1560-7917.ES.2018.23.16.17-00454

Caton, A. J., Brownlee, G. G., Yewdell, J. W., & Gerhard, W. (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell*, *31*(2 PART 1), 417–427. https://doi.org/10.1016/0092-8674(82)90135-0

Cauldwell, A. V., Long, J. S., Moncorgé, O., Barclay, W. S., Cauldwell, A. V., Long, J. S., Moncorge, O., Moncorgé, O., & Barclay, W. S. (2014). Viral determinants of influenza A virus host range. *The Journal of General Virology*, *95*(Pt 6), 1193–1210. https://doi.org/10.1099/vir.0.062836-0

Chambers, T. M. (2020). Equine Influenza. *Cold Spring Harb Perspect Med*. https://doi.org/10.1101/cshperspect.a038331

Chao, L. (1990). Fitness of RNA virus decreased by Muller's ratchet. *Nature*, *348*(6300), 454–455. https://doi.org/10.1038/348454a0

Chauché, C. M. (2017). *Molecular Evolution of Equine Influenza Virus Non-Structural Protein 1*. University of Glasgow.

Chen, F., & Cui, J. (2017). Cross-species epidemic dynamic model of influenza. *Proceedings - 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and*

Informatics, CISP-BMEI 2016, 1567–1572. https://doi.org/10.1109/CISP-BMEI.2016.7852965

Chen, J. M., Sun, Y. X., Chen, J. W., Liu, S., Yu, J. M., Shen, C. J., Sun, X. D., & Peng, D. (2009). Panorama phylogenetic diversity and distribution of type A influenza viruses based on their six internal gene sequences. *Virology Journal*, *6*(1), 137. https://doi.org/10.1186/1743-422X-6-137

Chen, Y. Q., Wohlbold, T. J., Zheng, N. Y., Huang, M., Huang, Y., Neu, K. E., Lee, J., Wan, H., Rojas, K. T., Kirkpatrick, E., Henry, C., Palm, A. K. E., Stamper, C. T., Lan, L. Y. L., Topham, D. J., Treanor, J., Wrammert, J., Ahmed, R., Eichelberger, M. C., … Wilson, P. C. (2018). Influenza Infection in Humans Induces Broadly Cross-Reactive and Protective Neuraminidase-Reactive Antibodies. *Cell*, *173*(2), 417-429.e10. https://doi.org/10.1016/j.cell.2018.03.030

Chlanda, P., Schraidt, O., Kummer, S., Riches, J., Oberwinkler, H., Prinz, S., Kräusslich, H.-G., & Briggs, J. A. G. (2015). Structural Analysis of the Roles of Influenza A Virus Membrane-Associated Proteins in Assembly and Morphology. *Journal of Virology*, *89*(17), 8957–8966. https://doi.org/10.1128/jvi.00592-15

Clark, A. M., Nogales, A., Martinez-Sobrido, L., Topham, D. J., & DeDiego, M. L. (2017). Functional evolution of influenza virus ns1 protein in currently circulating human 2009 pandemic h1n1 viruses. *Journal of Virology*, *91*(17). https://doi.org/10.1128/JVI.00721-17

Cleaveland, S., Haydon, D. T., & Taylor, L. (2007). Overviews of pathogen emergence: Which pathogens emerge, when and why? In *Current Topics in Microbiology and Immunology* (Vol. 315, pp. 85–111). https://doi.org/10.1007/978-3-540-70962-6_5

Cohn, S. K. (2020). The dramaturgy of epidemics. *Bulletin of the History of Medicine*, *94*(4), 578–589. https://doi.org/10.1353/bhm.2020.0083

Crawford, P. C., Dubovi, E. J., Castleman, W. L., Stephenson, I., Gibbs, E. P. J. J., Chen, L., Smith, C., Hill, R. C., Ferro, P., Pompey, J., Bright, R. A., Medina, M.-J., Johnson, C. M., Olsen, C. W., Cox, N. J., Klimov, A. I., Katz, J. M., & Donis, R. O. (2005). Transmission of Equine Influenza Virus to Dogs. *Science*, *310*(5747), 482–485. https://doi.org/10.1126/science.1117950

Cullinane, A., & Newton, J. R. (2013a). Equine influenza-A global perspective. In *Veterinary Microbiology* (Vol. 167, Issues 1–2, pp. 205–214). https://doi.org/10.1016/j.vetmic.2013.03.029

Cullinane, A., & Newton, J. R. (2013b). Equine influenza—A global perspective. *Veterinary Microbiology*, *167*(1–2), 205–214. https://doi.org/10.1016/j.vetmic.2013.03.029

Daly, J. M., Lai, A. C. K., Binns, M. M., Chambers, T. M., Barrandeguy, M., & Mumford, J. A. (1996). Antigenic and genetic evolution of equine H3N8 influenza A viruses. *J Gen*

Virol, *77 (Pt 4)*(4), 661–671. https://doi.org/10.1099/0022-1317-77-4-661

Daly, J. M., Yates, P. J., Newton, J. R., Park, A., Henley, W., Wood, J. L. N., Davis-Poynter, N., & Mumford, J. A. (2004). Evidence supporting the inclusion of strains from each of the two co-circulating lineages of H3N8 equine influenza virus in vaccines. *Vaccine*, *22*(29–30), 4101–4109. https://doi.org/10.1016/j.vaccine.2004.02.048

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Das, K., Aramini, J. M., Ma, L. C., Krug, R. M., & Arnold, E. (2010). Structures of influenza A proteins and insights into antiviral drug targets. *Nature Structural and Molecular Biology*, *17*(5), 530–538. https://doi.org/10.1038/nsmb.1779

Daversa, D. R., Fenton, A., Dell, A. I., Garner, T. W. J., & Manica, A. (2017). Infections on the move: How transient phases of host movement influence disease spread. In *Proceedings of the Royal Society B: Biological Sciences* (Vol. 284, Issue 1869). https://doi.org/10.1098/rspb.2017.1807

Dayhoff, M. O., & Foundation, N. B. R. (1979). *Atlas of Protein Sequence and Structure* (Issue v. 5). National Biomedical Research Foundation. https://books.google.co.uk/books?id=BIRFAQAAIAAJ

De, A., Sarkar, T., & Nandy, A. (2016). Bioinformatics studies of Influenza A hemagglutinin sequence data indicate recombination-like events leading to segment exchanges. *BMC Research Notes*, *9*, 222. https://doi.org/10.1186/s13104-016-2017-3

De Fine Licht, H. H. (2018). *Does pathogen plasticity facilitate host shifts?* https://doi.org/10.1371/journal.ppat.1006961

De Maio, N., Worby, C. J., Wilson, D. J., & Stoesser, N. (2018a). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Computational Biology*, *14*(4), 1–23. https://doi.org/10.1371/journal.pcbi.1006117

De Maio, N., Worby, C. J., Wilson, D. J., & Stoesser, N. (2018b). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Computational Biology*, *14*(4), 1–23. https://doi.org/10.1371/journal.pcbi.1006117

De Maio, N., Wu, C. H., & Wilson, D. J. (2016a). SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLoS Computational Biology*, *12*(9), 1–23. https://doi.org/10.1371/journal.pcbi.1005130

De Maio, N., Wu, C.-H., & Wilson, D. J. (2016b). SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLOS Computational Biology*, *12*(9), e1005130. https://doi.org/10.1371/journal.pcbi.1005130

De Vlugt, C., Sikora, D., & Pelchat, M. (2018). Insight into Influenza: A Virus Cap-Snatching. *Viruses*, *10*(11). https://doi.org/10.3390/v10110641

Dee, K., Goldfarb, D. M., Haney, J., Amat, J. A. R., Herder, V., Stewart, M., Szemiel, A. M., Baguelin, M., & Murcia, P. R. (2021). Human Rhinovirus Infection Blocks Severe Acute Respiratory Syndrome Coronavirus 2 Replication Within the Respiratory Epithelium: Implications for COVID-19 Epidemiology. *The Journal of Infectious Diseases*, *224*(1), 31–38. https://doi.org/10.1093/infdis/jiab147

Dee, K., Schultz, V., Haney, J., Bissett, L. A., Magill, C., & Murcia, P. R. (2022). Influenza A and Respiratory Syncytial Virus Trigger a Cellular Response That Blocks Severe Acute Respiratory Syndrome Virus 2 Infection in the Respiratory Tract. *The Journal of Infectious Diseases*. https://doi.org/10.1093/infdis/jiac494

Delport, W., Poon, A. F. Y. Y., Frost, S. D. W. W., & Kosakovsky Pond, S. L. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, *26*(19), 2455–2457. https://doi.org/10.1093/bioinformatics/btq429

Deng, J., Sheng, Z., Zhou, K., Duan, M., Yu, C., & Jiang, L. (2009). Construction of Effective Receptor for Recognition of Avian Influenza H5N1 Protein HA1 by Assembly of Monohead Glycolipids on Polydiacetylene Vesicle Surface. *Bioconjugate Chemistry*, *20*(3), 533–537. https://doi.org/10.1021/bc800453u

Diallo, A. A., Souley, M. M., Issa Ibrahim, A., Alassane, A., Issa, R., Gagara, H., Yaou, B., Issiakou, A., Diop, M., Ba Diouf, R. O., Lo, F. T., Lo, M. M. M. M., Bakhoum, T., Sylla, M., Seck, M. T., Meseko, C., Shittu, I., Cullinane, A., Settypalli, T. B. K., … Cattoli, G. (2021). Transboundary spread of equine influenza viruses (H3N8) in West and Central Africa: Molecular characterization of identified viruses during outbreaks in Niger and Senegal, in 2019. *Transbound Emerg Dis*, *68*(3), 1253–1262. https://doi.org/10.1111/tbed.13779

Dias, A., Bouvier, D., Crépin, T., McCarthy, A. A., Hart, D. J., Baudin, F., Cusack, S., & Ruigrok, R. W. H. (2009). The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature*, *458*(7240), 914–918. https://doi.org/10.1038/nature07745

Didelot, X., Fraser, C., Gardy, J., Colijn, C., & Malik, H. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, *34*(4), 997–1007. https://doi.org/10.1093/molbev/msw275

Dimas Martins, A., & Gjini, E. (2020). Modeling Competitive Mixtures With the Lotka-Volterra Framework for More Complex Fitness Assessment Between Strains. *Frontiers in Microbiology*, *11*(September), 1–11. https://doi.org/10.3389/fmicb.2020.572487

Ding, X., Qin, L., Meng, J., Peng, Y., Wu, A., & Jiang, T. (2021). Progress and Challenge in Computational Identification of Influenza Virus Reassortment. *Virologica Sinica*, *36*(6), 1273–1283. https://doi.org/10.1007/s12250-021-00392-w

Diskin, E. R., Friedman, K., Krauss, S., Nolting, J. M., Poulson, R. L., Slemons, R. D., Stallknecht, D. E., Webster, R. G., & Bowman, A. S. (2020). Subtype Diversity of Influenza A Virus in North American Waterfowl: a Multidecade Study. *Journal of Virology*, *94*(11). https://doi.org/10.1128/jvi.02022-19

Domingo, E. (2020). Long-term virus evolution in nature. In *Virus as Populations* (Vol. 2507, Issue 1, pp. 225–261). Elsevier. https://doi.org/10.1016/B978-0-12-816331-3.00007-6

Domingo, E., de Ávila, A. I., Gallego, I., Sheldon, J., & Perales, C. (2019). Viral fitness: history and relevance for viral pathogenesis and antiviral interventions. *Pathogens and Disease*, *77*(2), ftz021. https://doi.org/10.1093/femspd/ftz021

Domingo, E., Higuera, I. de la, Moreno, E., Ávila, A. I. de, Agudo, R., Arias, A., & Perales, C. (2017). Quasispecies Dynamics Taught by Natural and Experimental Evolution of Foot-and-mouth Disease Virus. In *Foot and Mouth Disease Virus: Current Research and Emerging Trends* (pp. 147–170). Caister Academic Press. https://doi.org/10.21775/9781910190517.07

Domingo, E., & Perales, C. (2019). Viral quasispecies. *PLoS Genetics*, *15*(10), 1–20. https://doi.org/10.1371/journal.pgen.1008271

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1). https://doi.org/10.1186/1471-2148-7-214

DuBois, R. M., Aguilar-Yañez, J. M., Mendoza-Ochoa, G. I., Oropeza-Almazán, Y., Schultz-Cherry, S., Alvarez, M. M., White, S. W., Russell, C. J., Aguilar-Yanez, J. M., Mendoza-Ochoa, G. I., Oropeza-Almazan, Y., Schultz-Cherry, S., Alvarez, M. M., White, S. W., Russell, C. J., Aguilar-Yañez, J. M., Mendoza-Ochoa, G. I., Oropeza-Almazán, Y., Schultz-Cherry, S., … Russell, C. J. (2011). The Receptor-Binding Domain of Influenza Virus Hemagglutinin Produced in Escherichia coli Folds into Its Native, Immunogenic Structure . *Journal of Virology*, *85*(2), 865–872. https://doi.org/10.1128/jvi.01412-10

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., Bielejec, F., Caddy, S. L., Cotten, M., D'Ambrozio, J., Dellicour, S., Di Caro, A., Diclaro, J. W., Duraffour, S., Elmore, M. J., … Rambaut, A. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, *544*(7650), 309–315. https://doi.org/10.1038/nature22040

Dunham, E. J., Dugan, V. G., Kaser, E. K., Perkins, S. E., Brown, I. H., Holmes, E. C., & Taubenberger, J. K. (2009). Different Evolutionary Trajectories of European Avian-Like and Classical Swine H1N1 Influenza A Viruses. *Journal of Virology*, *83*(11), 5485–5494. https://doi.org/10.1128/jvi.02565-08

Dunning, J., Thwaites, R. S., & Openshaw, P. J. M. (2020). Seasonal and pandemic influenza: 100 years of progress, still much to learn. *Mucosal Immunology*, *13*(4), 566–573. https://doi.org/10.1038/s41385-020-0287-5

Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., & Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Research*, *49*(W1), W216–W227. https://doi.org/10.1093/nar/gkab225

Duxbury, E. M. L., Day, J. P., Vespasiani, D. M., Thüringer, Y., Tolosana, I., Smith, S. C. L., Tagliaferri, L., Kamacioglu, A., Lindsley, I., Love, L., Unckless, R. L., Jiggins, F. M., & Longdon, B. (2019). Host-pathogen coevolution increases genetic variation in susceptibility to infection. *ELife*, *8*. https://doi.org/10.7554/eLife.46440

Ekiert, D. C., Friesen, R. H. E., Bhabha, G., Kwaks, T., Jongeneelen, M., Yu, W., Ophorst, C., Cox, F., Korse, H. J. W. M., Brandenburg, B., Vogels, R., Brakenhoff, J. P. J., Kompier, R., Koldijk, M. H., Cornelissen, L. A. H. M., Poon, L. L. M., Peiris, M., Koudstaal, W., Wilson, I. A., & Goudsmit, J. (2011). A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science (New York, N.Y.)*, *333*(6044), 843–850. https://doi.org/10.1126/science.1204839

Elena, S. F., Sanjuán, R., Bordería, A. V., & Turner, P. E. (2001). Transmission bottlenecks and the evolution of fitness in rapidly evolving RNA viruses. *Infection, Genetics and Evolution*, *1*(1), 41–48. https://doi.org/10.1016/S1567-1348(01)00006-5

Elliott, S., Olufemi, O. T., & Daly, J. M. (2023). Systematic Review of Equine Influenza A Virus Vaccine Studies and Meta-Analysis of Vaccine Efficacy. *Viruses*, *15*(12). https://doi.org/10.3390/V15122337

Emini, E. A., Hughes, J. V, Perlow, D. S., & J., B. (1985). Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol*, *55*(3), 836–839.

Endo, A., Pecoraro, R., Sugita, S., & Nerome, K. (1992). Evolutionary pattern of the H 3 haemagglutinin of equine influenza viruses: multiple evolutionary lineages and frozen replication. *Arch Virol*, *123*(1–2), 73–87. https://doi.org/10.1007/BF01317139

Evans, R., O\textquoterightNeill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Ž\'\idek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool,

K., Jain, R., Clancy, E., … Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*. https://doi.org/10.1101/2021.10.04.463034

Feng, K. H., Gonzalez, G., Deng, L., Yu, H., Tse, V. L., Huang, L., Huang, K., Wasik, B. R., Zhou, B., Wentworth, D. E., Holmes, E. C., Chen, X., Varki, A., Murcia, P. R., & Parrish, C. R. (2015). Equine and Canine Influenza H3N8 Viruses Show Minimal Biological Differences Despite Phylogenetic Divergence. *Journal of Virology*, *89*(13), 6860–6873. https://doi.org/10.1128/JVI.00521-15

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., & Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature 2005 437:7056*, *437*(7056), 209–214. https://doi.org/10.1038/nature04017

Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature 2006 442:7101*, *442*(7101), 448–452. https://doi.org/10.1038/nature04795

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., & Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, *260*(5551), 500–507. https://doi.org/10.1038/260500a0

Firestone, S. M., Hayama, Y., Bradhurst, R., Yamamoto, T., Tsutsui, T., & Stevenson, M. A. (2019). Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. *Scientific Reports*, *9*(1), 4809. https://doi.org/10.1038/s41598-019-41103-6

Firestone, S. M., Hayama, Y., Lau, M. S. Y., Yamamoto, T., Nishi, T., Bradhurst, R. A., Demirhan, H., Stevenson, M. A., & Tsutsui, T. (2020). Transmission network reconstruction for foot- and-mouth disease outbreaks incorporating farm-level covariates. *PLoS ONE*, *15*(7 July), 1–17. https://doi.org/10.1371/journal.pone.0235660

Flanagan, M. L., Parrish, C. R., Cobey, S., Glass, G. E., Bush, R. M., & Leighton, T. J. (2012). Anticipating the Species Jump: Surveillance for Emerging Viral Threats. In *Zoonoses and Public Health* (Vol. 59, Issue 3, pp. 155–163). https://doi.org/10.1111/j.1863-2378.2011.01439.x

Flannery, B., Zimmerman, R. K., Gubareva, L. V, Garten, R. J., Chung, J. R., Nowalk, M. P., Jackson, M. L., Jackson, L. A., Monto, A. S., Ohmit, S. E., Belongia, E. A., McLean, H. Q., Gaglani, M., Piedra, P. A., Mishin, V. P., Chesnokov, A. P., Spencer, S., Thaker, S. N., Barnes, J. R., … Fry, A. M. (2016). Enhanced Genetic Characterization of Influenza A(H3N2) Viruses and Vaccine Effectiveness by Genetic Group, 2014--

2015. *The Journal of Infectious Diseases*, *214*(7), 1010–1019. https://doi.org/10.1093/infdis/jiw181

Forster, T. (2021). Illustrations of the Atmospherical Origin of Epidemic Diseases, and of its relation to their Predisponent Constitutional Causes, Exemplified by Historical Notices and Cases, and on the Twofold Means of Prevention, Mitigation, and Cure, and of the Powerful. In *Spiritualism, Mesmerism and the Occult, 1800--1920 Vol 1* (pp. 51–58). Routledge.

Francis, T. (1960a). On the Doctrine of Original Antigenic Sin. *Proceedings of the American Philosophical Society*, *104*(6), 572–578. http://www.jstor.org/stable/985534

Francis, T. (1960b). On the Doctrine of Original Antigenic Sin Author (s): Thomas Francis , Jr . Published by : American Philosophical Society Stable URL : http://www.jstor.org/stable/985534 REFERENCES Linked references are available on JSTOR for this article : You may need. *Proceedings of the American Philosophical Society*, *104*(6), 572–578.

Frost, S. D. W., Magalis, B. R., & Kosakovsky Pond, S. L. (2018). Neutral theory and rapidly evolving viral pathogens. *Molecular Biology and Evolution*, *35*(6), 1348–1354. https://doi.org/10.1093/molbev/msy088

Frothingham, R. (1999). Evolutionary bottlenecks in the agents of tuberculosis, leprosy, and paratuberculosis. *Medical Hypotheses*, *52*(2), 95–99. https://doi.org/10.1054/MEHY.1997.0622

Fuhrmann, L., Jablonski, K. P., & Beerenwinkel, N. (2021). Quantitative measures of within-host viral genetic diversity. *Current Opinion in Virology*, *49*, 157–163. https://doi.org/10.1016/j.coviro.2021.06.002

Gallagher, M. E., Brooke, C. B., Ke, R., & Koelle, K. (2018a). Causes and Consequences of Spatial Within-Host Viral Spread. *Viruses*, *10*(11), 627. https://doi.org/10.3390/v10110627

Gallagher, M. E., Brooke, C. B., Ke, R., & Koelle, K. (2018b). Causes and consequences of spatial within-host viral spread. *Viruses*, *10*(11), 627. https://doi.org/10.3390/v10110627

Ganti, K., Bagga, A., DaSilva, J., Shepard, S. S., Barnes, J. R., Shriner, S., Koelle, K., & Lowen, A. C. (2021). Avian Influenza A Viruses Reassort and Diversify Differently in Mallards and Mammals. *Viruses*, *13*(3). https://doi.org/10.3390/v13030509

Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. In *Nature Reviews Genetics* (Vol. 19, Issue 1, pp. 9–20). https://doi.org/10.1038/nrg.2017.88

Garrison, E., & Marth, G. (2012a). *Haplotype-based variant detection from short-read sequencing*. 1–9. https://doi.org/10.48550/ARXIV.1207.3907

Garrison, E., & Marth, G. (2012b). Haplotype-based variant detection from short-read sequencing. *Q-Bio.GN*, 1–9. https://doi.org/https://doi.org/10.48550/arXiv.1207.3907

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). *The Proteomics Protocols Handbook* (J. M. Walker, Ed.; pp. 571–607). Humana Totowa, NJ.

Gelbart, M., Harari, S., Ben-Ari, Y., Kustin, T., Wolf, D., Mandelboim, M., Mor, O., Pennings, P. S., & Stern, A. (2020). Drivers of within-host genetic diversity in acute infections of viruses. *PLoS Pathogens*, *16*(11), 1–21. https://doi.org/10.1371/journal.ppat.1009029

Geoghegan, J. L., & Holmes, E. C. (2018). Evolutionary Virology at 40. *Genetics*, *210*(4), 1151–1162. https://doi.org/10.1534/genetics.118.301556

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., & Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, *3*(1), 811. https://doi.org/10.1038/ncomms1814

Gerstung, M., Papaemmanuil, E., & Campbell, P. J. (2014). Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics*, *30*(9), 1198–1204. https://doi.org/10.1093/bioinformatics/btt750

Ghafari, M., Lumby, C. K., Weissman, D. B., & Illingworth, C. J. R. (2020). Inferring Transmission Bottleneck Size from Viral Sequence Data Using a Novel Haplotype Reconstruction Method. *Journal of Virology*, *94*(13). https://doi.org/10.1128/jvi.00014-20

Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T., Leitner, T., & Tanaka, M. M. (2017). Inference of Transmission Network Structure from HIV Phylogenetic Trees. *PLoS Computational Biology*, *13*(1), 1005316. https://doi.org/10.1371/journal.pcbi.1005316

Gibb, R. J. (2020). *Understanding and predicting effects of global environmental change on zoonotic disease* [University College London]. https://www.researchgate.net/publication/346979374_Understanding_and_predicting_effects_of_global_environmental_change_on_zoonotic_disease

Gocnikova, H., & Russ, G. (2007). Influenza a virus PB1-F2 protein. *Acta Virol*, *51*(2), 101–108.

Gong, L. I., Suchard, M. A., & Bloom, J. D. (2013). Stability-mediated epistasis constrains the evolution of an influenza protein. *ELife*, *2013*(2). https://doi.org/10.7554/ELIFE.00631

Gore, S., Sanz García, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Mading, S., Mak, L., Mukhopadhyay, A.,

Oldfield, T. J., Patwardhan, A., Peisach, E., … Kleywegt, G. J. (2017). Validation of Structures in the Protein Data Bank. *Structure*, *25*(12), 1916–1927. https://doi.org/10.1016/j.str.2017.10.009

Gostic, K. M., Ambrose, M., Worobey, M., & Lloyd-Smith, J. O. (2016). Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science*, *354*(6313), 722–726. https://doi.org/10.1126/science.aag1322

Gostic, K. M., Bridge, R., Brady, S., Viboud, C., Worobey, M., & Lloyd-Smith, J. O. (2019). Childhood immune imprinting to influenza A shapes birth year-specific risk during seasonal H1N1 and H3N2 epidemics. *PLOS Pathogens*, *15*(12), e1008109. https://doi.org/10.1371/journal.ppat.1008109

Grear, D. A., Hall, J. S., Dusek, R. J., & Ip, H. S. (2018). Inferring epidemiologic dynamics from viral evolution: 2014–2015 Eurasian/North American highly pathogenic avian influenza viruses exceed transmission threshold, in wild birds and poultry in North America. *Evolutionary Applications*, *11*(4), 547–557. https://doi.org/10.1111/eva.12576

Greatorex, J. S., Digard, P., Curran, M. D., Moynihan, R., Wensley, H., Wreghitt, T., Varsani, H., Garcia, F., Enstone, J., & Nguyen-Van-Tam, J. S. (2011). Survival of influenza A(H1N1) on materials found in households: implications for infection control. *PLoS One*, *6*(11), e27932. https://doi.org/10.1371/journal.pone.0027932

Gregori, J., Perales, C., Rodriguez-Frias, F., Esteban, J. I., Quer, J., & Domingo, E. (2016a). Viral quasispecies complexity measures. *Virology*, *493*, 227–237. https://doi.org/10.1016/j.virol.2016.03.017

Gregori, J., Perales, C., Rodriguez-Frias, F., Esteban, J. I., Quer, J., & Domingo, E. (2016b). Viral quasispecies complexity measures. *Virology*, *493*, 227–237. https://doi.org/https://doi.org/10.1016/j.virol.2016.03.017

Gregori, J., Perales, C., Rodriguez-Frias, F., Esteban, J. I., Quer, J., & Domingo, E. (2016c). Viral quasispecies complexity measures. *Virology*, *493*, 227–237. https://doi.org/https://doi.org/10.1016/j.virol.2016.03.017

Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Van Rompay, K. K. A., Isern, S., Michael, S. F., Coffey, L. L., Loman, N. J., Andersen, K. G., Goes De Jesus, J., Main, B. J., … Andersen, K. G. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology*, *20*(8), 8. https://doi.org/10.1186/s13059-018-1618-7

Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., Goes De Jesus, J., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Rompay, K. K. A. Van, Isern, S., Michael, S. F., Coffey, L. L., Loman, N. J., &

Andersen, K. G. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology*, *20*(8). https://doi.org/10.1186/s13059-018-1618-7

Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., Goes De Jesus, J., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Rompay, K. K. A. Van, Isern, S., Michael, S. F., Coffey, L. L., Loman, N. J., Andersen, K. G., De Jesus, J. G., Main, B. J., … Andersen, K. G. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology*, *20*(8), 8. https://doi.org/10.1186/s13059-018-1618-7

Gutnik, D., Evseev, P., Miroshnikov, K., & Shneider, M. (2023). Using AlphaFold Predictions in Viral Research. *Current Issues in Molecular Biology*, *45*(4), 3705–3732. https://doi.org/10.3390/cimb45040240

Hall, M., Woolhouse, M., & Rambaut, A. (2015). Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Computational Biology*, *11*(12), 1–36. https://doi.org/10.1371/journal.pcbi.1004613

Hallgren, J., Tsirigos, K. D., Pedersen, M. D., Armenteros, J. J. A., Marcatili, P., Nielsen, H., Krogh, A., & Winther, O. (2022). *DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks*. https://doi.org/10.1101/2022.04.08.487609

Han, A. X., Felix Garza, Z. C., Welkers, M. R. A. A., Vigeveno, R. M., Tran, N. D., Le, T. Q. M. T. H. T. Q. M., Pham Quang, T., Dang, D. T., Tran, T. H. T. N. A. T. H., Ha, M. T., Nguyen, T. H., Le, Q. T., Le, T. Q. M. T. H. T. Q. M., Hoang, T. B. N., Chokephaibulkit, K., Puthavathana, P., Nguyen, V. V. C. K., Nghiem, M. N., Nguyen, V. V. C. K., … Russell, C. A. (2021). Within-host evolutionary dynamics of seasonal and pandemic human influenza A viruses in young children. *ELife*, *10*(8), e68917. https://doi.org/10.7554/eLife.68917

Han, G.-Z., & Worobey, M. (2011). Homologous recombination in negative sense RNA viruses. *Viruses*, *3*(8), 1358–1373. https://doi.org/10.3390/v3081358

Haney, J., Vijayakrishnan, S., Streetley, J., Dee, K., Goldfarb, D. M., Clarke, M., Mullin, M., Carter, S. D., Bhella, D., & Murcia, P. R. (2022). Coinfection by influenza A virus and respiratory syncytial virus produces hybrid virus particles. *Nature Microbiology 2022 7:11*, *7*(11), 1879–1890. https://doi.org/10.1038/s41564-022-01242-5

Hapuarachchi, H. C., Koo, C., Kek, R., Xu, H., Lai, Y. L., Liu, L., Kok, S. Y., Shi, Y., Chuen, R. L. T., Lee, K.-S., Maurer-Stroh, S., & Ng, L. C. (2016). Intra-epidemic evolutionary dynamics of a Dengue virus type 1 population reveal mutant spectra

that correlate with disease transmission. *Scientific Reports*, *6*(1), 22592. https://doi.org/10.1038/srep22592

Hartshorn, K. L. (2020). Innate Immunity and Influenza A Virus Pathogenesis: Lessons for COVID-19. *Front Cell Infect Microbiol*, *10*, 563850. https://doi.org/10.3389/fcimb.2020.563850

Harvey, W. T., Benton, D. J., Gregory, V., Hall, J. P. J. J., Daniels, R. S., Bedford, T., Haydon, D. T., Hay, A. J., McCauley, J. W., & Reeve, R. (2016). Identification of Low- and High-Impact Hemagglutinin Amino Acid Substitutions That Drive Antigenic Drift of Influenza A(H1N1) Viruses. *PLoS Pathogens*, *12*(4), 1–23. https://doi.org/10.1371/journal.ppat.1005526

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160–174. https://doi.org/10.1007/BF02101694

Hausser, J., & Strimmer, K. (2008). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, *10*, 1469–1484. http://arxiv.org/abs/0811.3579

Hausser, J., & Strimmer, K. (2021). *Estimation of Entropy, Mutual Information and Related Quantities* (1.3.1). https://cran.r-project.org/package=entropy

Hayden, F. G., Fritz, R. S., Lobo, M. C., Alvord, W. G., Strober, W., & Straus, S. E. (1998). Local and systemic cytokine responses during experimental human influenza A virus infection. Relation to symptom formation and host defense. *The Journal of Clinical Investigation*, *101*(3), 643–649. https://doi.org/10.1172/JCI1355

Haydon, D. T., Cleaveland, S., Taylor, L. H., & Laurenson, M. K. (2002). Identifying reservoirs of infection: A conceptual and practical challenge. In *Emerging Infectious Diseases* (Vol. 8, Issue 12, pp. 1468–1473). https://doi.org/10.3201/eid0812.010317

Hayward, A. C., Fragaszy, E. B., Bermingham, A., Wang, L., Copas, A., Edmunds, W. J., Ferguson, N., Goonetilleke, N., Harvey, G., Kovar, J., Lim, M. S. C., McMichael, A., Millett, E. R. C., Nguyen-Van-Tam, J. S., Nazareth, I., Pebody, R., Tabassum, F., Watson, J. M., Wurie, F. B., … Zambon, M. (2014a). Comparative community burden and severity of seasonal and pandemic influenza: Results of the Flu Watch cohort study. *The Lancet Respiratory Medicine*, *2*(6), 445–454. https://doi.org/10.1016/S2213-2600(14)70034-7

Hayward, A. C., Fragaszy, E. B., Bermingham, A., Wang, L., Copas, A., Edmunds, W. J., Ferguson, N., Goonetilleke, N., Harvey, G., Kovar, J., Lim, M. S. C., McMichael, A., Millett, E. R. C., Nguyen-Van-Tam, J. S., Nazareth, I., Pebody, R., Tabassum, F., Watson, J. M., Wurie, F. B., … Zambon, M. (2014b). Comparative community burden and severity of

seasonal and pandemic influenza: results of the Flu Watch cohort study. *The Lancet Respiratory Medicine*, *2*(6), 445–454. https://doi.org/10.1016/S2213-2600(14)70034-7

Health and Social Care, D. of. (2020). *UK Pandemic Preparedness*. https://www.gov.uk/government/publications/uk-pandemic-preparedness

Heesterbeek, J. A. P. A. P. (2002). A brief history of R0 and a recipe for its calculation. *Acta Biotheoretica*, *50*(3), 189–204. https://doi.org/10.1017/cbo9780511525667.003

Hemmink, J. D., Morgan, S. B., Aramouni, M., Everett, H., Salguero, F. J., Canini, L., Porter, E., Chase-Topping, M., Beck, K., Loughlin, R. Mac, Carr, B. V., Brown, I. H., Bailey, M., Woolhouse, M., Brookes, S. M., Charleston, B., & Tchilian, E. (2016). Distinct immune responses and virus shedding in pigs following aerosol, intra-nasal and contact infection with pandemic swine influenza A virus, A(H1N1)09. *Veterinary Research*, *47*(1), 103. https://doi.org/10.1186/s13567-016-0390-5

Henry, C., Palm, A.-K. E., Krammer, F., & Wilson, P. C. (2018). From Original Antigenic Sin to the Universal Influenza Virus Vaccine. *Trends in Immunology*, *39*(1), 70–79. https://doi.org/10.1016/j.it.2017.08.003

Hicks, J. T., Lee, D.-H., Duvvuri, V. R., Kim Torchetti, M., Swayne, D. E., & Bahl, J. (2020). Agricultural and geographic factors shaped the North American 2015 highly pathogenic avian influenza H5N2 outbreak. *PLOS Pathogens*, *16*(1), e1007857. https://doi.org/10.1371/journal.ppat.1007857

Hidano, A., & Gates, M. C. (2019a). Assessing biases in phylodynamic inferences in the presence of super-spreaders. *Veterinary Research*, *50*(1), 74. https://doi.org/10.1186/s13567-019-0692-5

Hidano, A., & Gates, M. C. (2019b). Assessing biases in phylodynamic inferences in the presence of super-spreaders. *Veterinary Research*, *50*(1), 74. https://doi.org/10.1186/s13567-019-0692-5

Hill, W. G. (1988). Molecular Evolutionary Genetics. By Masatoshi Nei. New York: Columbia University Press. 1987. 512 pages. U.S. $50.00. ISBN 0 231 06320 2. *Genetical Research*, *52*(1), 74–75. https://doi.org/10.1017/S001667230002735X

Holland, J. J., De La Torre, J. C., & Steinhauer, D. A. (1992). RNA Virus Populations as Quasispecies. In *Current topics in microbiology and immunology* (Vol. 176, pp. 1–20). https://doi.org/10.1007/978-3-642-77011-1_1

Houldcroft, C. J., Beale, M. A., & Breuer, J. (2017). Clinical and biological insights from viral genome sequencing. In *Nature Reviews Microbiology* (Vol. 15, Issue 3, pp. 183–192). Nature Publishing Group. https://doi.org/10.1038/nrmicro.2016.182

Houlihan, C. F., Frampton, D., Ferns, R. B., Raffle, J., Grant, P., Reidy, M., Hail, L., Thomson, K., Mattes, F., Kozlakidis, Z., Pillay, D., Hayward, A., & Nastouli, E. (2018). Use of Whole-

Genome Sequencing in the Investigation of a Nosocomial Influenza Virus Outbreak. *The Journal of Infectious Diseases*, *218*(9), 1485–1489. https://doi.org/10.1093/infdis/jiy335

Hrecka, K., Hao, C., Gierszewska, M., Swanson, S. K., Kesik-Brodacka, M., Srivastava, S., Florens, L., Washburn, M. P., & Skowronski, J. (2011). Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. *Nature*, *474*(7353), 658–661. https://doi.org/10.1038/nature10195

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2011). ART: a next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594. https://doi.org/10.1093/bioinformatics/btr708

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of Levels of Gene Flow from DNA Sequence Data. *Genetics*, *132*(2), 583. https://doi.org/10.1093/GENETICS/132.2.583

Hughes, J. (2016). DiversiTools. In *GitHub repository*. GitHub.

Hughes, J., Allen, R. C., Baguelin, M., Hampson, K., Baillie, G. J., Elton, D., Newton, J. R., Kellam, P., Wood, J. L. N. N., Holmes, E. C., & Murcia, P. R. (2012a). Transmission of Equine Influenza Virus during an Outbreak Is Characterized by Frequent Mixed Infections and Loose Transmission Bottlenecks. *PLOS Pathogens*, *8*(12), 1–15. https://doi.org/10.1371/journal.ppat.1003081

Hughes, J., Allen, R. C., Baguelin, M., Hampson, K., Baillie, G. J., Elton, D., Newton, J. R., Kellam, P., Wood, J. L. N. N., Holmes, E. C., & Murcia, P. R. (2012b). Transmission of Equine Influenza Virus during an Outbreak Is Characterized by Frequent Mixed Infections and Loose Transmission Bottlenecks. *PLOS Pathogens*, *8*(12), 1–15. https://doi.org/10.1371/journal.ppat.1003081

Illingworth, C. J. R., & Mustonen, V. (2012). Components of Selection in the Evolution of the Influenza Virus: Linkage Effects Beat Inherent Selection. *PLoS Pathogens*, *8*(12). https://doi.org/10.1371/journal.ppat.1003091

Illingworth, C. J. R. R. (2016). SAMFIRE: Multi-locus variant calling for time-resolved sequence data. *Bioinformatics*, *32*(14), 2208–2209. https://doi.org/10.1093/bioinformatics/btw205

Illingworth, C. J. R., Raghwani, J., Serwadda, D., Sewankambo, N. K., Robb, M. L., Eller, M. A., Redd, A. R., Quinn, T. C., & Lythgoe, K. A. (2020). A de novo approach to inferring within-host fitness effects during untreated HIV-1 infection. *PLoS Pathogens*, *16*(6). https://doi.org/10.1371/journal.ppat.1008171

International Codes of Practice. (2023, May). Horserace Betting Levy Board,. https://codes.hblb.org.uk/index.php/page/168

Ip, D. K. M., Lau, L. L. H., Leung, N. H. L., Fang, V. J., Chan, K. H., Chu, D. K. W., Leung, G. M., Peiris, J. S. M., Uyeki, T. M., & Cowling, B. J. (2017). Viral Shedding and Transmission

Potential of Asymptomatic and Paucisymptomatic Influenza Virus Infections in the Community. *Clinical Infectious Diseases*, *64*(6), 736–742. https://doi.org/10.1093/CID/CIW841

Ito, T., Gorman, O. T., Kawaoka, Y., Bean, W. J., & Webster, R. G. (1991). Evolutionary analysis of the influenza A virus M gene with comparison of the M1 and M2 proteins. *Journal of Virology*, *65*(10), 5491–5498. https://doi.org/10.1128/jvi.65.10.5491-5498.1991

Ito, T., Kawaoka, Y., Ohira, M., Takakuwa, H., Yasuda, J., Kida, H., & Otsuki, K. (1999). Replacement of internal protein genes, with the exception of the matrix, in equine 1 viruses by equine 2 influenza virus genes during evolution in nature. *J Vet Med Sci*, *61*(8), 987–989. https://doi.org/10.1292/jvms.61.987

Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., Cohen, C., Gran, J. M., Schanzer, D., Cowling, B. J., Wu, P., Kyncl, J., Ang, L. W., Park, M., Redlberger-Fritz, M., Yu, H., Espenhain, L., Krishnan, A., Emukule, G., … Mustaquim, D. (2018). Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet*, *391*(10127), 1285–1300. https://doi.org/https://doi.org/10.1016/S0140-6736(17)33293-2

Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*, *45*(W1), W24–W29. https://doi.org/10.1093/nar/gkx346

Jiao J, Fefferman N. The dynamics of evolutionary rescue from a novel pathogen threat in a host metapopulation. Sci Rep. 2021 May 25;11(1):10932. doi: 10.1038/s41598-021-90407-z. PMID: 34035424; PMCID: PMC8149858.

Johnson, K. E. E., & Ghedin, E. (2020). Quantifying between-Host Transmission in Influenza Virus Infections. *Cold Spring Harb Perspect Med*, *10*(8). https://doi.org/10.1101/cshperspect.a038422

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., & Ferguson, N. (2014). Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, *10*(1), 1003457. https://doi.org/10.1371/journal.pcbi.1003457

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*(7181), 990–993. https://doi.org/10.1038/nature06536

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. In *Molecular Ecology* (Vol. 25, Issue 1, pp. 185–202). Blackwell Publishing Ltd. https://doi.org/10.1111/mec.13304

Jorba, N., Coloma, R., & Ortín, J. (2009). Genetic trans-Complementation Establishes a New Model for Influenza Virus RNA Transcription and Replication. *PLoS Pathogens*, *5*(5), e1000462. https://doi.org/10.1371/journal.ppat.1000462

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Karamendin, K., Kydyrmanov, A., Kasymbekov, Y., Khan, E., Daulbayeva, K., Asanova, S., Zhumatov, K., Seidalina, A., Sayatov, M., & Fereidouni, S. R. (2014). Continuing evolution of equine influenza virus in Central Asia, 2007-2012. *Arch Virol*, *159*(9), 2321–2327. https://doi.org/10.1007/s00705-014-2078-3

Kariuki, S. M., Selhorst, P., Ariën, K. K., & Dorfman, J. R. (2017a). *The HIV-1 transmission bottleneck*. *14*, 22. https://doi.org/10.1186/s12977-017-0343-8

Kariuki, S. M., Selhorst, P., Ariën, K. K., & Dorfman, J. R. (2017b). The {HIV}-1 transmission bottleneck. *Retrovirology*, *14*(1), 22. https://doi.org/10.1186/s12977-017-0343-8

Kawaoka, Y., Bean, W. J., & Webster, R. G. (1989). Evolution of the hemagglutinin of equine H3 influenza viruses. *Virology*, *169*(2), 283–292. https://doi.org/10.1016/0042-6822(89)90153-0

Kelvin, A. A., & Zambon, M. (2019). Influenza imprinting in childhood and the influence on vaccine response later in life. *Eurosurveillance*, *24*(48), 1–5. https://doi.org/10.2807/1560-7917.ES.2019.24.48.1900720

Kenah, E., Britton, T., Halloran, M. E., & Longini, I. M. (2016). Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. *PLoS Computational Biology*, *12*(4). https://doi.org/10.1371/journal.pcbi.1004869

Khan, A., Mushtaq, M. H., Muhammad, J., Ahmed, B., Khan, E. A., Khan, A., Zakki, S. A., Altaf, E., Saleem, A., Warraich, M. A., Ahmed, N., Rabaan, A. A., Haq, I., Saleem, A., Warraich, M. A., Ahmed, N., & Rabaan, A. A. (2021). Global epidemiology of Equine Influenza viruses; “A possible emerging zoonotic threat in future” an extensive systematic review with evidence. *Braz J Biol*, *83*, e246591. https://doi.org/10.1590/1519-6984.246591

Kim, J. H., Skountzou, I., Compans, R., & Jacob, J. (2009a). Original Antigenic Sin Responses to Influenza Viruses. *The Journal of Immunology*, *183*(5), 3294–3301. https://doi.org/10.4049/jimmunol.0900398

Kim, J. H., Skountzou, I., Compans, R., & Jacob, J. (2009b). Original Antigenic Sin Responses to Influenza Viruses. *The*

Journal of Immunology, *183*(5), 3294–3301. https://doi.org/10.4049/jimmunol.0900398

Kim K, Omori R, Ueno K, Iida S, Ito K (2016) Host-Specific and Segment-Specific Evolutionary Dynamics of Avian and Human Influenza A Viruses: A Systematic Review. PLoS ONE 11(1): e0147021. doi:10.1371/journal.pone.0147021

Kim, W. K., Kim, J. A., Song, D. H., Lee, D., Kim, Y. C., Lee, S. Y., Lee, S. H., No, J. S., Kim, J. H., Kho, J. H., Gu, S. H., Jeong, S. T., Wiley, M., Kim, H. C., Klein, T. A., Palacios, G., & Song, J. W. (2016). Phylogeographic analysis of hemorrhagic fever with renal syndrome patients using multiplex PCR-based next generation sequencing. *Scientific Reports*, *6*(1), 1–8. https://doi.org/10.1038/srep26017

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, *19*(A), 27–43. https://doi.org/10.2307/3213548

Klenerman, P., & Zinkernagel, R. M. (1998). Original antigenic sin impairs cytotoxic T lymphocyte responses to viruses bearing variant epitopes. *Nature*, *394*(6692), 482–485. https://doi.org/10.1038/28860

Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., & Wallinga, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Computational Biology*, *13*(5). https://doi.org/10.1371/journal.pcbi.1005495

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

Koel, B. F., Vigeveno, R. M., Pater, M., Koekkoek, S. M., Han, A. X., Tuan, H. M., Anh, T. T. N., Hung, N. T., Thinh, L. Q., Hai, L. T., Ngoc, H. T. B., Chau, N. V. V., Ngoc, N. M., Chokephaibulkit, K., Puthavathana, P., Kinh, N. V., Trinh, T., Lee, R. T. C., Maurer-Stroh, S., … De Jong, M. D. (2020). Longitudinal sampling is required to maximize detection of intrahost A/H3N2 virus variants. *Virus Evolution*, *6*(2). https://doi.org/10.1093/ve/veaa088

Koelle, K., Khatri, P., Kamradt, M., & Kepler, T. B. (2010). A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. *J R Soc Interface*, *7*(50), 1257–1274. https://doi.org/10.1098/rsif.2010.0007

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., & Schlötterer, C. (2011). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLOS ONE*, *6*(1), 1–9. https://doi.org/10.1371/journal.pone.0015925

Kolaskar, A. S., & Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, *276*(1–2), 172–174. https://doi.org/10.1016/0014-5793(90)80535-q

Korneliussen, T.S., Moltke, I., Albrechtsen, A. *et al*. Calculation of Tajima's *D* and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* **14**, 289 (2013). https://doi.org/10.1186/1471-2105-14-289

Kratsch, C., Klingen, T. R., Mü Mken, L., Steinbrü Ck, L., & Mchardy, A. C. (2016). Determination of antigenicity-altering patches on the major surface protein of human influenza A/H3N2 viruses. *Virus Evolution*, *2*(1). https://doi.org/10.1093/ve/vev025

Kryazhimskiy, S., Dushoff, J., Bazykin, G. A., & Plotkin, J. B. (2011). Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PLOS Genetics*, *7*(2), e1001301. https://doi.org/10.1371/JOURNAL.PGEN.1001301

Kühnert, D., Wu, C. H., & Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. In *Infection, Genetics and Evolution* (Vol. 11, Issue 8, pp. 1825–1841). https://doi.org/10.1016/j.meegid.2011.08.005

Kuiken, T., Holmes, E. C., McCauley, J., Rimmelzwaan, G. F., Williams, C. S., & Grenfell, B. T. (2006). Host Species Barriers to Influenza Virus Infections. *Science*, *312*(5772), 394–397. https://doi.org/10.1126/science.1122818

Kwong, J. C., Mccallum, N., Sintchenko, V., & Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology*, *47*(3), 199–210. https://doi.org/10.1097/PAT.0000000000000235

Laabassi, F., Lecouturier, F., Amelot, G., Gaudaire, D., Mamache, B., Laugier, C., Legrand, L., Zientara, S., & Hans, A. (2015). Epidemiology and Genetic Characterization of H3N8 Equine Influenza Virus Responsible for Clinical Disease in Algeria in 2011. *Transbound Emerg Dis*, *62*(6), 623–631. https://doi.org/10.1111/tbed.12209

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Ségéral, E., Yatim, A., Emiliani, S., Schwartz, O., & Benkirane, M. (2011). SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature*, *474*(7353), 654–657. https://doi.org/10.1038/nature10117

Lai, A. C. K., Chambers, T. M., Holland Jr., R. E., Morley, P. S., Haines, D. M., Townsend, H. G. G., & Barrandeguy, M. (2001). Diverged evolution of recent equine-2 influenza (H3N8) viruses in the Western Hemisphere. *Archives of Virology*, *146*(6), 1063–1074. https://doi.org/10.1007/s007050170106

Lai, A. C. K., Chambers, T. M., Holland, R. E., & Morley, P. S. (2001). *Diverged evolution of recent equine-2 influenza (H3N8) viruses in the Western Hemisphere*. 1063–1074.

Lai, A. C. K., Rogers, K. M., Glaser, A., Tudor, L., & Chambers, T. (2004). Alternate circulation of recent equine-2 influenza viruses (H3N8) from two distinct lineages in the United States. *Virus Res*, *100*(2), 159–164. https://doi.org/10.1016/j.virusres.2003.11.019

Landolt, G. A. (2014). Equine influenza virus. *Veterinary Clinics of North America - Equine Practice*, *30*(3), 507–522. https://doi.org/10.1016/j.cveq.2014.08.003

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lau, L. L. H., Cowling, B. J., Fang, V. J., Chan, K. H., Lau, E. H. Y., Lipsitch, M., Cheng, C. K. Y., Houck, P. M., Uyeki, T. M., Malik Peiris, J. S., & Leung, G. M. (2010). Viral Shedding and Clinical Illness in Naturally Acquired Influenza Virus Infections. *The Journal of Infectious Diseases*, *201*(10), 1509–1516. https://doi.org/10.1086/652241

Lauring, A. S. (2020a). Within-Host Viral Diversity: A Window into Viral Evolution. *Annual Review of Virology*, *7*(1), 63–81. https://doi.org/10.1146/annurev-virology-010320-061642

Lauring, A. S. (2020b). Within-Host Viral Diversity: A Window into Viral Evolution. *Annual Review of Virology*, *7*(1), 63–81. https://doi.org/10.1146/annurev-virology-010320

Lazarus, L., Patel, S., Shaw, A., Leblanc, S., Lalonde, C., Hladio, M., Mandryk, K., Horvath, C., Petrcich, W., Kendall, C., & Tyndall, M. W. (2016). Uptake of community-based peer administered HIV point-of-care testing: Findings from the PROUD study. *PLoS ONE*, *11*(12), e0166942. https://doi.org/10.1371/journal.pone.0166942

Lazniewski, M., Dawson, W. K., Szczepińska, T., & Plewczynski, D. (2018). The structural variability of the influenza A hemagglutinin receptor-binding site. *Briefings in Functional Genomics*, *17*(6), 415–427. https://doi.org/10.1093/bfgp/elx042

LeClair, J. S., & Wahl, L. M. (2018). The Impact of Population Bottlenecks on Microbial Adaptation. *Journal of Statistical Physics*, *172*(1), 114–125. https://doi.org/10.1007/s10955-017-1924-6

Lee, C. Y., Raghunathan, V., Caceres, C. J., Geiger, G., Seibert, B., Faccin, F. C., Gay, L. C., Ferreri, L. M., Kaul, D., Wrammert, J., Tan, G. S., Perez, D. R., & Lowen, A. C. (2023). Epistasis reduces fitness costs of influenza A virus escape from stem-binding antibodies. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(17), e2208718120. https://doi.org/10.1073/PNAS.2208718120/SUPPL_FILE/PNAS.2208718120.SD18.XLSX

Lee, C.-Y., An, S.-H., Choi, J.-G., Lee, Y.-J., Kim, J.-H., & Kwon, H.-J. (2020). Rank orders of mammalian pathogenicity-related PB2 mutations of avian influenza A viruses. *Scientific Reports*, *10*(1), 5359. https://doi.org/10.1038/s41598-020-62036-5

Lee, H. K., Lee, C. K., Tang, J. W. T., Loh, T. P., & Koay, E. S. C. (2016). Contamination-controlled high-throughput whole genome sequencing for influenza A viruses using the MiSeq sequencer. *Scientific Reports*, *6*(1), 1–11. https://doi.org/10.1038/srep33318

Lee, K., Pusterla, N., Barnum, S. M., Lee, D.-H., & Martínez-López, B. (2021). Genome-informed characterisation of antigenic drift in the haemagglutinin gene of equine influenza strains circulating in the United States from 2012 to 2017. *Transbound Emerg Dis*, *69*(4), 1–12. https://doi.org/10.1111/tbed.14262

Leeks, A., Segredo-Otero, E. A., Sanjuán, R., & West, S. A. (2018). Beneficial coinfection can promote within-host viral diversity. *Virus Evolution*, *4*(2), 1–12. https://doi.org/10.1093/ve/vey028

Lefeuvre, P., Lett, J.-M., Varsani, A., & Martin, D. P. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of Virology*, *83*(6), 2697 – 2707. https://doi.org/10.1128/JVI.02152-08

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, *46*(D1), D708–D717. https://doi.org/10.1093/nar/gkx932

Legrand, L. J., Pitel, P.-H. Y., Cullinane, A. A., Fortier, G. D., Pronost, S. L., & Universite, N. (2015). Genetic evolution of equine influenza strains isolated in France from 2005 to 2010. *Equine Vet J*, *47*(2), 207–211. https://doi.org/10.1111/evj.12244

Lennox, K., Dahl, D., Vannucci, M., & Tsai, J. (2009). Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics. *Journal of the American Statistical Association*, *104*, 586–596. https://doi.org/10.1198/jasa.2009.0024

Leonard, A. S., Weissman, D. B., Greenbaum, B. C., Ghedin, E., Koelle, K., Elodie Ghedin, D., Koellea, K., Ghedin, E., Koelle, K., Elodie Ghedin, D., Koellea, K., Ghedin, E., & Koelle, K. (2017). Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza a virus. In *bioRxiv* (Vol. 91, Issue 14). https://doi.org/10.1101/101790

Leonard, A. S., Weissman, D. B., Greenbaum, B., Ghedin, E., & Koelle, K. (2019). Correction for Sobel Leonard et al., “Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human

Influenza A Virus." *Journal of Virology*, *93*(17), 10.1128/jvi.00936-19. https://doi.org/10.1128/jvi.00936-19

Lewis, N. S., Daly, J. M., Russell, C. A., Horton, D. L., Skepner, E., Bryant, N. A., Burke, D. F., Rash, A. S., Wood, J. L. N., Chambers, T. M., Fouchier, R. A. M., Mumford, J. A., Elton, D. M., & Smith, D. J. (2011). Antigenic and Genetic Evolution of Equine Influenza A (H3N8) Virus from 1968 to 2007. *Journal of Virology*, *85*(23), 12742–12749. https://doi.org/10.1128/jvi.05319-11

Lewis, N. S., Russell, C. A., Langat, P., Anderson, T. K., Berger, K., Bielejec, F., Burke, D. F., Dudas, G., Fonville, J. M., Fouchier, R. A., Kellam, P., Koel, B. F., Lemey, P., Nguyen, T., Nuansrichy, B., Peiris, J. M., Saito, T., Simon, G., Skepner, E., … Vincent, A. L. (2016). The global antigenic diversity of swine influenza A viruses. *ELife*, *5*. https://doi.org/10.7554/eLife.12217

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, Volume 25, Issue 14, Pages 1754-1760 https://doi.org/10.1093/bioinformatics/btp324

Li, J., Ishaq, M., Prudence, M., Xi, X., Hu, T., Liu, Q., & Guo, D. (2009). Single mutation at the amino acid position 627 of PB2 that leads to increased virulence of an H5N1 avian influenza virus during adaptation in mice can be compensated by multiple mutations at other sites of PB2. *Virus Research*, *144*(1), 123–129. https://doi.org/https://doi.org/10.1016/j.virusres.2009.04.008

Li, W., Lee, H. H. Y., Li, R. F., Zhu, H. M., Yi, G., Peiris, J. S. M., Yang, Z. F., & Mok, C. K. P. (2017). The PB2 mutation with lysine at 627 enhances the pathogenicity of avian influenza (H7N9) virus which belongs to a non-zoonotic lineage. *Scientific Reports*, *7*(1), 2352. https://doi.org/10.1038/s41598-017-02598-z

Liao, C. M., Yang, S. C., Chio, C. P., & Chen, S. C. (2010). Understanding influenza virus-specific epidemiological properties by analysis of experimental human infections. *Epidemiology & Infection*, *138*(6), 825–835. https://doi.org/10.1017/S0950268809991178

Lindstrom, S., Endo, A., Sugita, S., Pecoraro, M., Hiromoto, Y., Kamada, M., Takahashi, T., & Nerome, K. (1998). Phylogenetic analyses of the matrix and non-structural genes of equine influenza viruses. *Arch Virol*, *143*(8), 1585–1598. https://doi.org/10.1007/s007050050400

Lipsitch, M., Barclay, W., Raman, R., Russell, C. J., Belser, J. A., Cobey, S., Kasson, P. M., Lloyd-Smith, J. O., Maurer-Stroh, S., Riley, S., Beauchemin, C. A., Bedford, T., Friedrich, T. C., Handel, A., Herfst, S., Murcia, P. R., Roche, B., Wilke, C. O., & Russell, C. A. (2016). Viral factors in influenza

pandemic risk assessment. *ELife*, *5*, 38. https://doi.org/10.7554/eLife.18491

Liu, T., Wang, Y., Tan, T. J. C., Wu, N. C., & Brooke, C. B. (2022). The evolutionary potential of influenza A virus hemagglutinin is highly constrained by epistatic interactions with neuraminidase. *Cell Host & Microbe*, *30*(10), 1363-1369.e4. https://doi.org/10.1016/j.chom.2022.09.003

Lloyd, L. E., Jonczyk, M., Jervis, C. M., Flack, D. J., Lyall, J., Foote, A., Mumford, J. A., Brown, I. H., Wood, J. L., & Elton, D. M. (2011). Experimental transmission of avian-like swine H1N1 influenza virus between immunologically naïve and vaccinated pigs. *Influenza and Other Respiratory Viruses*, *5*(5), 357–364. https://doi.org/https://doi.org/10.1111/j.1750-2659.2011.00233.x

Lopes, A. M., Domingues, P., Zell, R., & Hale, B. G. (2017). Structure-Guided Functional Annotation of the Influenza A Virus NS1 Protein Reveals Dynamic Evolution of the p85ß-Binding Site during Circulation in Humans. *Journal of Virology*, *91*(21), 1–16. https://doi.org/10.1128/JVI.01081-17

Lumby, C. K., Nene, N. R., & Illingworth, C. J. R. (2018). A novel framework for inferring parameters of transmission from viral sequence data. In *PLoS Genetics* (Vol. 14, Issue 10). Public Library of Science. https://doi.org/10.1371/journal.pgen.1007718

Lumby, C. K., Zhao, L., Breuer, J., & Illingworth, C. J. R. J. (2020). A large effective population size for established within-host influenza virus infection. *ELife*, *9*, 1–17. https://doi.org/10.7554/eLife.56915

Lyons, D. M., & Lauring, A. S. (2018). Mutation and Epistasis in Influenza Virus Evolution. *Viruses 2018, Vol. 10, Page 407*, *10*(8), 407. https://doi.org/10.3390/V10080407

Mackenzie, J. S., Childs, J. E., & Richt, J. A. (2007). The Biology, circumstances and consequences of cross-species transmission. In *Current topics in microbiology and immunology*.

Maeda, Y., Takemura, T., Chikata, T., Kuwata, T., Terasawa, H., Fujimoto, R., Kuse, N., Akahoshi, T., Murakoshi, H., Tran, G. Van, Zhang, Y., Pham, C. H., Pham, A. H. Q., Monde, K., Sawa, T., Matsushita, S., Nguyen, T. V., Nguyen, K. Van, Hasebe, F., … Takiguchi, M. (2020). Existence of replication-competent minor variants with different coreceptor usage in plasma from hiv-1-infected individuals. *Journal of Virology*, *94*(12), 1–17. https://doi.org/10.1128/JVI.00193-20

Magori, K., & Park, A. W. (2014). The evolutionary consequences of alternative types of imperfect vaccines. *J Math Biol*, *68*(4), 969–987. https://doi.org/10.1007/s00285-013-0654-x

Mak, L., Perera, D., Lang, R., Kossinna, P., He, J., Gill, M. J., Long, Q., & van Marle, G. (2020). Evaluation of a phylogenetic pipeline to examine transmission networks in a

canadian HIV cohort. *Microorganisms*, *8*(2). https://doi.org/10.3390/microorganisms8020196

Manuguerra, J. C., Zientara, S., Sailleau, C., Rousseaux, C., Gicquel, B., Rijks, I., & van der Werf, S. (2000). Evidence for evolutionary stasis and genetic drift by genetic analysis of two equine influenza H3 viruses isolated in France. *Vet Microbiol*, *74*(1–2), 59–70. https://doi.org/10.1016/s0378-1135(00)00166-8

Marjanovic, S., Romanelli, R. J., Ali, G.-C., Leach, B., Bonsu, M., Rodriguez-Rincon, D., & Ling, T. (2022). COVID-19 Genomics UK (COG-UK) Consortium: Final Report. *Rand Health Quarterly*, *9*(4), 24.

Marshall, N., Priyamvada, L., Ende, Z., Steel, J., & Lowen, A. C. (2013). Influenza Virus Reassortment Occurs with High Frequency in the Absence of Segment Mismatch. *PLOS Pathogens*, *9*(6), 1–11. https://doi.org/10.1371/journal.ppat.1003421

Mather, A. E., Matthews, L., Mellor, D. J., Reeve, R., Denwood, M. J., Boerlin, P., Reid-Smith, R. J., Brown, D. J., Coia, J. E., Browning, L. M., Haydon, D. T., & Reid, S. W. J. (2012). An ecological approach to assessing the epidemiology of antimicrobial resistance in animal and human populations. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1733), 1630–1639. https://doi.org/10.1098/rspb.2011.1975

Matthews, L., & Woolhouse, M. (2005). New approaches to quantifying the spread of infection. In *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro1178

McCrone, J. T., & Lauring, A. S. (2016). Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. *Journal of Virology*, *90*(15), 6884–6895. https://doi.org/10.1128/jvi.00667-16

McCrone, J. T., & Lauring, A. S. (2018a). Genetic bottlenecks in intraspecies virus transmission. In *Current Opinion in Virology* (Vol. 28, pp. 20–25). Elsevier B.V. https://doi.org/10.1016/j.coviro.2017.10.008

McCrone, J. T., & Lauring, A. S. (2018b). Genetic bottlenecks in intraspecies virus transmission. *Current Opinion in Virology*, *28*, 20–25. https://doi.org/10.1016/j.coviro.2017.10.008

McCrone, J. T., Woods, R. J., Martin, E. T., Malosh, R. E., Monto, A. S., & Lauring, A. S. (2018a). Stochastic processes constrain the within and between host evolution of influenza virus. *ELife*, *7*. https://doi.org/10.7554/eLife.35962

McCrone, J. T., Woods, R. J., Martin, E. T., Malosh, R. E., Monto, A. S., & Lauring, A. S. (2018b). Stochastic processes constrain the within and between host evolution of influenza virus. *ELife*, *7*, e35962. https://doi.org/10.7554/eLife.35962

McCrone, J. T., Woods, R. J., Monto, A. S., Martin, E. T., & Lauring, A. S. (2020a). *The effective population size and*

mutation rate of influenza {A} virus in acutely infected individuals. https://doi.org/10.1101/2020.10.24.353748

McCrone, J. T., Woods, R. J., Monto, A. S., Martin, E. T., & Lauring, A. S. (2020b). The effective population size and mutation rate of influenza A virus in acutely infected individuals. In *bioRxiv*. https://doi.org/10.1101/2020.10.24.353748

McCrone, J. T., Woods, R. J., Monto, A. S., Martin, E. T., & Lauring, A. S. (2020c). The effective population size and mutation rate of influenza {A} virus in acutely infected individuals. In *bioRxiv*. https://doi.org/10.1101/2020.10.24.353748

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. https://doi.org/10.1093/ofid/ofaa494

McKellar, J., Rebendenne, A., Wencker, M., Moncorge, O., & Goujon, C. (2021). Mammalian and Avian Host Cell Influenza A Restriction Factors. *VIRUSES-BASEL*, *13*(3). https://doi.org/10.3390/v13030522

Meinel, D. M., Heinzinger, S., Eberle, U., Ackermann, N., Schönberger, K., & Sing, A. (2018). Whole genome sequencing identifies influenza A H3N2 transmission and offers superior resolution to classical typing methods. *Infection*, *46*(1), 69–76. https://doi.org/10.1007/s15010-017-1091-3

Mendenhall, I. H., Wen, D. L. H., Jayakumar, J., Gunalan, V., Wang, L., Mauer-Stroh, S., Su, Y. C. F., & Smith, G. J. D. (2019). Diversity and evolution of viral pathogen community in cave nectar bats (eonycteris spelaea). *Viruses*, *11*(3). https://doi.org/10.3390/v11030250

Min, J.-Y., Santos, C., Fitch, A., Twaddle, A., Toyoda, Y., DePasse, J. V, Ghedin, E., & Subbarao, K. (2013). Mammalian adaptation in the PB2 gene of avian H5N1 influenza virus. *Journal of Virology*, *87*(19), 10884–10888. https://doi.org/10.1128/JVI.01016-13

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Monne, I., Fusaro, A., Nelson, M. I., Bonfanti, L., Mulatti, P., Hughes, J., Murcia, P. R., Schivo, A., Valastro, V., Moreno, A., Holmes, E. C., & Cattoli, G. (2014). Emergence of a

Highly Pathogenic Avian Influenza Virus from a Low-Pathogenic Progenitor. *Journal of Virology*, *88*(8), 4375–4388. https://doi.org/10.1128/jvi.03181-13

Monto, A. S., Malosh, R. E., Petrie, J. G., & Martin, E. T. (2017). The Doctrine of Original Antigenic Sin: Separating Good From Evil. *The Journal of Infectious Diseases*, *215*(12), 1782–1788. https://doi.org/10.1093/infdis/jix173

Morelli, M. J., Wright, C. F., Knowles, N. J., Juleff, N., Paton, D. J., King, D. P., & Haydon, D. T. (2013). Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Veterinary Research*, *44*(1), 12. https://doi.org/10.1186/1297-9716-44-12

Morens, D. M., & Taubenberger, J. K. (2010). Historical thoughts on influenza viral ecosystems, or behold a pale horse, dead dogs, failing fowl, and sick swine. *Influenza and Other Respiratory Viruses*, *4*(6), 327–337. https://doi.org/https://doi.org/10.1111/j.1750-2659.2010.00148.x

Morse, S. S., Mazet, J. A. K., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., Zambrana-Torrelio, C., Lipkin, W. I., & Daszak, P. (2012). Prediction and prevention of the next pandemic zoonosis. In *The Lancet* (Vol. 380, Issue 9857, pp. 1956–1965). https://doi.org/10.1016/S0140-6736(12)61684-5

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K. E., Venter, J. C., & Telenti, A. (2017). The blood DNA virome in 8,000 humans. *PLoS Pathogens*, *13*(3), e1006292. https://doi.org/10.1371/journal.ppat.1006292

Muller, H. J. (1932). Some Genetic Aspects of Sex. *The American Naturalist*, *66*(703), 118–138. https://doi.org/10.1086/280418

Müller, I., Pinto, E., Santibáñez, M. C., Celedón, M. O., Valenzuela, P. D. T., Santiba, C., Mu, I., & Valenzuela, P. D. T. (2009). Isolation and characterization of the equine influenza virus causing the 2006 outbreak in Chile. *Vet Microbiol*, *137*(1–2), 172–177. https://doi.org/10.1016/j.vetmic.2008.12.011

Mumford, J. A. (2007). Vaccines and viral antigenic diversity. *Rev Sci Tech*, *26*(1), 69–90.

Murcia, P. R., Baillie, G. J., Daly, J., Elton, D., Jervis, C., Mumford, J. A., Newton, R., Parrish, C. R., Hoelzer, K., Dougan, G., Parkhill, J., Lennard, N., Ormond, D., Moule, S., Whitwham, A., McCauley, J. W., McKinley, T. J., Holmes, E. C., Grenfell, B. T., & Wood, J. L. N. (2010a). Intra- and Interhost Evolutionary Dynamics of Equine Influenza Virus. *Journal of Virology*, *84*(14), 6943–6954. https://doi.org/10.1128/jvi.00112-10

Murcia, P. R., Baillie, G. J., Daly, J., Elton, D., Jervis, C., Mumford, J. A., Newton, R., Parrish, C. R., Hoelzer, K., Dougan, G., Parkhill, J., Lennard, N., Ormond, D., Moule, S., Whitwham, A., McCauley, J. W., McKinley, T. J., Holmes, E. C., Grenfell, B. T., & Wood, J. L. N. (2010b). Intra- and Interhost Evolutionary Dynamics of Equine Influenza Virus. *Journal of Virology*, *84*(14), 6943–6954. https://doi.org/10.1128/jvi.00112-10

Murcia, P. R., Baillie, G. J., Daly, J., Elton, D., Jervis, C., Mumford, J. A., Newton, R., Parrish, C. R., Hoelzer, K., Dougan, G., Parkhill, J., Lennard, N., Ormond, D., Moule, S., Whitwham, A., McCauley, J. W., McKinley, T. J., Holmes, E. C., Grenfell, B. T., & Wood, J. L. N. (2010c). Intra- and Interhost Evolutionary Dynamics of Equine Influenza Virus. *Journal of Virology*, *84*(14), 6943–6954. https://doi.org/10.1128/jvi.00112-10

Murcia, P. R., Baillie, G. J., Stack, J. C., Jervis, C., Elton, D., Mumford, J. A., Daly, J., Kellam, P., Grenfell, B. T., Holmes, E. C., & Wood, J. L. N. (2013a). Evolution of Equine Influenza Virus in Vaccinated Horses. *Journal of Virology*, *87*(8), 4768–4771. https://doi.org/10.1128/jvi.03379-12

Murcia, P. R., Baillie, G. J., Stack, J. C., Jervis, C., Elton, D., Mumford, J. A., Daly, J., Kellam, P., Grenfell, B. T., Holmes, E. C., & Wood, J. L. N. (2013b). Evolution of Equine Influenza Virus in Vaccinated Horses. *Journal of Virology*, *87*(8), 4768–4771. https://doi.org/10.1128/jvi.03379-12

Murcia, P. R., Baillie, G. J., Stack, J. C., Jervis, C., Elton, D., Mumford, J. A., Daly, J., Kellam, P., Grenfell, B. T., Holmes, E. C., & Wood, J. L. N. (2013c). Evolution of Equine Influenza Virus in Vaccinated Horses. *Journal of Virology*, *87*(8), 4768–4771. https://doi.org/10.1128/jvi.03379-12

Murcia, P. R., Hughes, J., Battista, P., Lloyd, L., Baillie, G. J., Ramirez-Gonzalez, R. H., Ormond, D., Oliver, K., Elton, D., Mumford, J. A., Caccamo, M., Kellam, P., Grenfell, B. T., Holmes, E. C., & Wood, J. L. N. (2012). Evolution of an Eurasian Avian-like Influenza Virus in Naïve and Vaccinated Pigs. *PLOS Pathogens*, *8*(5), 1–12. https://doi.org/10.1371/journal.ppat.1002730

Murcia, P. R., Wood, J. L. N., & Holmes, E. C. (2011). Genome-Scale Evolution and Phylodynamics of Equine H3N8 Influenza A Virus. *Journal of Virology*, *85*(11), 5312–5322. https://doi.org/10.1128/JVI.02619-10

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Neil, S. J. D., Zang, T., & Bieniasz, P. D. (2008). Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature*, *451*(7177), 425–430. https://doi.org/10.1038/nature06553

Neira, V., Rabinowitz, P., Rendahl, A., Paccha, B., Gibbs, S. G., & Torremorell, M. (2016). Characterization of Viral Load, Viability and Persistence of Influenza A Virus in Air and on Surfaces of Swine Production Facilities. *PLOS ONE*, *11*(1), 1–11. https://doi.org/10.1371/journal.pone.0146616

Nelson, C. W., & Hughes, A. L. (2015). Within-host nucleotide diversity of virus populations: Insights from next-generation sequencing. *Infection, Genetics and Evolution*, *30*, 1–7. https://doi.org/10.1016/j.meegid.2014.11.026

Nemoto, M., Ohta, M., Yamanaka, T., Kambayashi, Y., Bannai, H., Tsujimura, K., Yamayoshi, S., Kawaoka, Y., & Cullinane, A. (2021). Antigenic differences between equine influenza virus vaccine strains and Florida sublineage clade 1 strains isolated in Europe in 2019. *Vet J*, *272*, 105674. https://doi.org/10.1016/j.tvjl.2021.105674

Neverov, A. D., Kryazhimskiy, S., Plotkin, J. B., & Bazykin, G. A. (2015). Coordinated Evolution of Influenza A Surface Proteins. *PLoS Genetics*, *11*(8), e1005404. https://doi.org/10.1371/journal.pgen.1005404

Newton, J. R., Daly, J. M., Spencer, L., & Mumford, J. A. (2006). Description of the outbreak of equine influenza (H3N8) in the United Kingdom in 2003, during which recently vaccinated horses in Newmarket developed respiratory disease. *Veterinary Record*, *158*(6), 185–192. https://doi.org/10.1136/vr.158.6.185

Nicholls, S. M., Poplawski, R., Bull, M. J., Underwood, A., Chapman, M., Abu-Dahab, K., Taylor, B., Colquhoun, R. M., Rowe, W. P. M., Jackson, B., Hill, V., O'Toole, Á., Rey, S., Southgate, J., Amato, R., Livett, R., Gonçalves, S., Harrison, E. M., Peacock, S. J., … Loman, N. J. (2021). CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. In *Genome biology* (Vol. 22, Issue 1, p. 196). https://doi.org/10.1186/s13059-021-02395-y

Nicholson, K. G. (1992). Clinical features of influenza. *Seminars in Respiratory Infections*, *7*(1), 26—37. http://europepmc.org/abstract/MED/1609165

NOAH. (2016). NOAH Compendium. *National Organisation of Animal Health*. https://www.noahcompendium.co.uk/?id=-457321

Oakeson, K. F., Wagner, J. M., Mendenhall, M., Rohrwasser, A., & Atkinson-Dunn, R. (2017). Bioinformatic analyses of whole-genome sequence data in a public health laboratory. *Emerging Infectious Diseases*, *23*(9), 1441–1445. https://doi.org/10.3201/eid2309.170416

O'Carroll, I. P., & Rein, A. (2016). Viral Nucleic Acids. In R. A. Bradshaw & P. D. Stahl (Eds.), *Encyclopedia of Cell Biology* (pp. 517–524). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-394447-4.10061-6

Ohmit, S. E., Petrie, J. G., Malosh, R. E., Johnson, E., Truscon, R., Aaron, B., Martens, C., Cheng, C., Fry, A. M., & Monto, A. S. (2015). Substantial Influenza Vaccine Effectiveness in Households With Children During the 2013--2014 Influenza Season, When 2009 Pandemic Influenza A(H1N1) Virus Predominated. *The Journal of Infectious Diseases*, *213*(8), 1229–1236. https://doi.org/10.1093/infdis/jiv563

Oladunni, F. S., Oseni, S. O., Martinez-Sobrido, L., & Chambers, T. M. (2021). Equine Influenza Virus and Vaccines. *Viruses*, *13*(8), 1657. https://doi.org/10.3390/v13081657

Olguin-Perglione, C., & Barrandeguy, M. E. (2021). An Overview of Equine Influenza in South America. *Viruses*, *13*(5), 888. https://doi.org/10.3390/v13050888

Ørsted, M. I., Anthony Hoffmann, A. I., Sverrisdó ttir, E. I., Lehmann Nielsen, K., Nygaard Kristensen, T., Hoffmann, A. A., Sverrisdóttir, E., Nielsen, K. L., & Kristensen, T. N. (2019a). Genomic variation predicts adaptive evolutionary responses better than population bottleneck history. *PLOS Genetics*, *15*(6), e1008205. https://doi.org/10.1371/journal.pgen.1008205

Ørsted, M. I., Anthony Hoffmann, A. I., Sverrisdó ttir, E. I., Lehmann Nielsen, K., Nygaard Kristensen, T., Hoffmann, A. A., Sverrisdóttir, E., Nielsen, K. L., & Kristensen, T. N. (2019b). Genomic variation predicts adaptive evolutionary responses better than population bottleneck history. *PLOS Genetics*, *15*(6), e1008205. https://doi.org/10.1371/journal.pgen.1008205

Orton, R. (2022). VSensus. In *GitHub repository*. GitHub.

Orton, R. J., Wright, C. F., King, D. P., & Haydon, D. T. (2020). Estimating viral bottleneck sizes for FMDV transmission within and between hosts and implications for the rate of viral evolution. *Interface Focus*, *10*(1). https://doi.org/10.1098/rsfs.2019.0066

Ostfeld, R. S., Glass, G. E., & Keesing, F. (2005). Spatial epidemiology: An emerging (or re-emerging) discipline. In *Trends in Ecology and Evolution* (Vol. 20, Issue 6 SPEC. ISS., pp. 328–336). https://doi.org/10.1016/j.tree.2005.03.009

Oxburgh, L., & Klingeborn, B. (1999). Cocirculation of two distinct lineages of equine influenza virus subtype H3N8. *J Clin Microbiol*, *37*(9), 3005–3009. https://doi.org/10.1128/JCM.37.9.3005-3009.1999

Paillot, R., Rash, N. L., Garrett, D., Prowse-Davis, L., Montesso, F., Cullinane, A., Lemaitre, L., Thibault, J.-C., Wittreck, S., & Dancer, A. (2016). How to Meet the Last OIE Expert Surveillance Panel Recommendations on Equine Influenza (EI) Vaccine Composition: A Review of the Process Required for the Recombinant Canarypox-Based EI Vaccine. *Pathogens*, *5*(4), 1–13. https://doi.org/10.3390/pathogens5040064

Pan, K. (2011). Understanding Original Antigenic Sin in Influenza with a Dynamical System. *PLoS ONE*, *6*(8), e23910. https://doi.org/10.1371/journal.pone.0023910

Paradis, E., & Schliep, K. (2018). 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, A. W., Daly, J. M., Lewis, N. S., Smith, D. J., Wood, J. L. N. N., & Grenfell, B. T. (2009). Quantifying the Impact of Immune Escape on Transmission Dynamics of Influenza. *Science*, *326*(5953), 726–728. https://doi.org/10.1126/science.1175980

Parker, I. M., Saunders, M., Bontrager, M., Weitz, A. P., Hendricks, R., Magarey, R., Suiter, K., & Gilbert, G. S. (2015). Phylogenetic structure and host abundance drive disease pressure in communities. *Nature*, *520*(7548), 542–544. https://doi.org/10.1038/nature14372

Parrish, C. R., Holmes, E. C., Morens, D. M., Park, E.-C., Burke, D. S., Calisher, C. H., Laughlin, C. A., Saif, L. J., & Daszak, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiology and Molecular Biology Reviews : MMBR*, *72*(3), 457–470. https://doi.org/10.1128/MMBR.00004-08

Parsons TL, Bolker BM, Dushoff J, Earn DJD. The probability of epidemic burnout in the stochastic SIR model with vital dynamics. Proc Natl Acad Sci U S A. 2024 Jan 30;121(5):e2313708120. doi: 10.1073/pnas.2313708120. Epub 2024 Jan 26. PMID: 38277438; PMCID: PMC10835029.

Pauly, M. D., Procario, M. C., & Lauring, A. S. (2017). A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. *ELife*, *6*, e26437. https://doi.org/10.7554/eLife.26437

Pedruzzi, G., & Rouzine, I. M. (2021). An evolution-based high-fidelity method of epistasis measurement: Theory and application to influenza. *PLOS Pathogens*, *17*(6), e1009669. https://doi.org/10.1371/JOURNAL.PPAT.1009669

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 : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *30*, 296–307. https://doi.org/10.1016/j.meegid.2014.12.022

Perglione, C. O., Golemba, M. D., Torres, C., & Barrandeguy, M. (2016). Molecular epidemiology and spatio-temporal dynamics of the H3N8 equine influenza virus in South America. *Pathogens*, *5*(4). https://doi.org/10.3390/pathogens5040061

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E.

(2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*, *30*(1), 70–82. https://doi.org/10.1002/pro.3943

Pevsner, J. (2009). *Bioinformatics and Functional Genomics*. Wiley. https://books.google.co.uk/books?id=awjHNAEACAAJ

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

Pietrasik, T. (2023). *Pandemic Influenza Preparedness (PIP) Framework*.

Poelvoorde, L. A. E. Van, Van Poelvoorde, L. A. E., Delcourt, T., Vuylsteke, M., De Keersmaecker, S. C. J., Thomas, I., Van Gucht, S., Saelens, X., Roosens, N., & Vanneste, K. (2022). A general approach to identify low-frequency variants within influenza samples collected during routine surveillance. *Microbial Genomics*, *8*(9), 1–13. https://doi.org/10.1099/mgen.0.000867

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

Poon, L. L. M. M., Song, T., Rosenfeld, R., Lin, X., Rogers, M. B., Zhou, B., Sebra, R., Halpin, R. A., Guan, Y., Twaddle, A., DePasse, J. V., Stockwell, T. B., Wentworth, D. E., Holmes, E. C., Greenbaum, B., Peiris, J. S. M., Cowling, B. J., & Ghedin, E. (2016). Quantifying influenza virus diversity and transmission in humans. *Nature Genetics*, *48*(2), 195–200. https://doi.org/10.1038/ng.3479

Power, R. A., Davaniah, S., Derache, A., Wilkinson, E., Tanser, F., Gupta, R. K., Pillay, D., & de Oliveira, T. (2016). Genome-Wide Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. *PLOS ONE*, *11*(9), e0163746. https://doi.org/10.1371/journal.pone.0163746

Rabadan, R., Levine, A. J., & Krasnitz, M. (2008). Non-random reassortment in human influenza A viruses. *Influenza and Other Respiratory Viruses*, *2*(1), 9–22. https://doi.org/https://doi.org/10.1111/j.1750-2659.2007.00030.x

Rambaut, A. (2018, November). *FigTree v1.4.4*. https://github.com/rambaut/figtree

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

Rash, A., Woodward, A., Bryant, N., McCauley, J., & Elton, D. (2014). An efficient genome sequencing method for equine influenza [H3N8] virus reveals a new polymorphism in the PA-

X protein. *Virol J*, *11*, 159. https://doi.org/10.1186/1743-422X-11-159

Rees, E. E., Pond, B. A., Cullingham, C. I., Tinline, R. R., Ball, D., Kyle, C. J., & White, B. N. (2009). Landscape modelling spatial bottlenecks: implications for raccoon rabies disease spread. *Biology Letters*. https://doi.org/10.1098/rsbl.2009.0094

Reeve, R., Leinster, T., Cobbold, C. A., Thompson, J., Brummitt, N., Mitchell, S. N., & Matthews, L. (2014). How to partition diversity. *Q-Bio.QM*, *September 2015*. http://arxiv.org/abs/1404.6520

Righetto, I., & Filippini, F. (2018). Pandemic Avian Influenza and Intra/Interhaemagglutinin Subtype Electrostatic Variation among Viruses Isolated from Avian, Mammalian, and Human Hosts. *BioMed Research International*, *2018*, 1–10. https://doi.org/10.1155/2018/3870508

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L. M., Lam, T. H., Thach, T. Q., Chau, P., Chan, K. P., Lo, S. V., Leung, P. Y., Tsang, T., Ho, W., Lee, K. H., Lau, E. M. C., Ferguson, N. M., & Anderson, R. M. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science*, *300*(5627), 1961–1966. https://doi.org/10.1126/science.1086478

Rioux, M., McNeil, M., Francis, M. E., Dawe, N., Foley, M., Langley, J. M., & Kelvin, A. A. (2020). The power of first impressions: Can influenza imprinting during infancy inform vaccine design? *Vaccines*, *8*(3), 1–21. https://doi.org/10.3390/vaccines8030546

Rivailler, P., Perry, I. A., Jang, Y., Davis, C. T., Chen, L.-M., Dubovi, E. J., & Donis, R. O. (2010). Evolution of canine and equine influenza (H3N8) viruses co-circulating between 2005 and 2008. *Virology*, *408*(1), 71–79. https://doi.org/10.1016/j.virol.2010.08.022

Rodríguez-Nevado, C., Lam, T. T. T.-Y., Holmes, E. C., Pagán, I., Rodr\'\iguez-Nevado, C., Lam, T. T. T.-Y., Holmes, E. C., & Pagán, I. (2018). The impact of host genetic diversity on virus evolution and emergence. *Ecology Letters*, *21*(2), 253–263. https://doi.org/10.1111/ele.12890

Rodríguez-Nevado, C., Lam, T. T. Y., Holmes, E. C., & Pagán, I. (2018). The impact of host genetic diversity on virus evolution and emergence. In *Ecology Letters* (Vol. 21, Issue 2, pp. 253–263). Blackwell Publishing Ltd. https://doi.org/10.1111/ele.12890

Rosenberg, C. E. (1992). Framing disease: studies in cultural history. Introduction. Framing disease: illness, society, and history. *Hospital Practice (Office Ed.)*, *27*(7), 179-182,185-186,191-192. https://doi.org/10.1080/21548331.1992.11705460

Rouli, L., Merhej, V., Fournier, P.-E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, *7*, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005

Rouzine, I. M., & Rozhnova, G. (2018). Antigenic evolution of viruses in host populations. *PLoS Pathog*, *14*(9), e1007291. https://doi.org/10.1371/journal.ppat.1007291

Rozek, W., Kwasnik, M., Socha, W., Sztromwasser, P., & Rola, J. (2021). Analysis of Single Nucleotide Variants (SNVs) Induced by Passages of Equine Influenza Virus H3N8 in Embryonated Chicken Eggs. *Viruses*, *13*(8). https://doi.org/10.3390/v13081551

Russell, C. J. (2021). Hemagglutinin Stability and Its Impact on Influenza A Virus Infectivity, Pathogenicity, and Transmissibility in Avians, Mice, Swine, Seals, Ferrets, and Humans. *Viruses*, *13*(5), 746. https://doi.org/10.3390/v13050746

Rutty, J. (1770). *A chronological history of the weather and seasons, and of the prevailing diseases in Dublin: With the various periods, successions, and revolutions, during the space of forty years: With a comparative view of the difference of the Irish Climate and Disea*. Robinson and Roberts.

Ryu, S., & Cowling, B. J. (2021). Human Influenza Epidemiology. *Cold Spring Harbor Perspectives in Medicine*, *11*(12), a038356. https://doi.org/10.1101/cshperspect.a038356

Saito, T., Kawaoka, Y., & Webster, R. G. (1993). Phylogenetic analysis of the N8 neuraminidase gene of influenza A viruses. *Virology*, *193*(2), 868–876. https://doi.org/10.1006/viro.1993.1196

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687–695. https://doi.org/10.1038/265687a0

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

Schönherz, A. A., Lorenzen, N., Guldbrandtsen, B., Buitenhuis, B., & Einer-Jensen, K. (2016). Ultra-deep sequencing of VHSV isolates contributes to understanding the role of viral quasispecies. *Veterinary Research*, *47*(1). https://doi.org/10.1186/s13567-015-0298-5

Schotsaert, M., & García-Sastre, A. (2014). Influenza vaccines: A moving interdisciplinary field. In *Viruses* (Vol. 6, Issue 10, pp. 3809–3826). https://doi.org/10.3390/v6103809

Schürch, A. C., & van Schaik, W. (2017). Challenges and opportunities for whole-genome sequencing–based surveillance of antibiotic resistance. *Annals of the New York*

Academy of Sciences, *1388*(1), 108–120. https://doi.org/10.1111/nyas.13310

Seladi-Schulman, J., Campbell, P. J., Suppiah, S., Steel, J., & Lowen, A. C. (2014). Filament-producing mutants of influenza A/Puerto Rico/8/1934 (H1N1) virus have higher neuraminidase activities than the spherical wild-type. *PloS One*, *9*(11), e112462. https://doi.org/10.1371/journal.pone.0112462

Selzer, L., Su, Z., Pintilie, G. D., Chiu, W., & Kirkegaard, K. (2020). Full-length three-dimensional structure of the influenza a virus M1 protein and its organization into a matrix layer. *PLoS Biology*, *18*(9), 1–26. https://doi.org/10.1371/journal.pbio.3000827

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst Techn J*, *27*, 379–423.

Shrestha, S., Foxman, B., Weinberger, D. M., Steiner, C., Viboud, C., & Rohani, P. (2013). Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. *Science Translational Medicine*, *5*(191), 191ra84. https://doi.org/10.1126/scitranslmed.3005982

Sievers, F., Geoffrey, J. B., & Higgins, D. G. (2020). Multiple Sequence Alignments. In A. D. Baxevanis, G. D. Bader, & D. S. Wishart (Eds.), *Bioinformatics* (4th ed., pp. 227–250).

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145. https://doi.org/https://doi.org/10.1002/pro.3290

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539. https://doi.org/10.1038/msb.2011.75

Sigal, D., Reid, J. N. S., & Wahl, L. M. (2018). Effects of Transmission Bottlenecks on the Diversity of Influenza A Virus. *Genetics*, *210*(3), 1075–1088. https://doi.org/10.1534/genetics.118.301510

Simmonds, P., Aiewsakun, P., & Katzourakis, A. (2019a). Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*, *17*, 321–328. https://doi.org/10.1038/s41579-018-0120-2

Simmonds, P., Aiewsakun, P., & Katzourakis, A. (2019b). Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*, *17*(5), 321–328. https://doi.org/10.1038/s41579-018-0120-2

Simonsen, L., Reichert, T. A., & Miller, M. A. (2004). The virtues of antigenic sin: consequences of pandemic recycling on influenza-associated mortality. *International Congress Series*, *1263*, 791–794. https://doi.org/https://doi.org/10.1016/j.ics.2004.01.029

Skehel, J. J., Stevens, D. J., Daniels, R. S., Douglas, A. R., Knossow, M., Wilson, I. A., & Wiley, D. C. (1984). A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(6), 1779–1783. https://doi.org/10.1073/pnas.81.6.1779

Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D., Bunimovich, L., Costenbader, E., Sexton, C., O'Connor, S., Xia, G. L., & Khudyakov, Y. (2018). QUENTIN: Reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, *34*(1), 163–170. https://doi.org/10.1093/bioinformatics/btx402

Smith, B. P. (2004). Evolution of equine infection control programs. *Vet Clin North Am Equine Pract*, *20*(3), 521–530, v. https://doi.org/10.1016/j.cveq.2004.07.002

Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S. M., Guan, Y., & Rambaut, A. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, *459*(7250), 1122–1125. https://doi.org/10.1038/nature08182

Sobel Leonard, A., Weissman, D. B., Greenbaum, B., Ghedin, E., & Koelle, K. (2017a). Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *Journal of Virology*, *91*(14), 1–19. https://doi.org/10.1128/JVI.00171-17

Sobel Leonard, A., Weissman, D. B., Greenbaum, B., Ghedin, E., & Koelle, K. (2017b). Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *Journal of Virology*, *91*(14), 1–19. https://doi.org/10.1128/JVI.00171-17

Sobolev, O. V., Afonine, P. V., Moriarty, N. W., Hekkelman, M. L., Joosten, R. P., Perrakis, A., & Adams, P. D. (2020). A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry. *Structure*, *28*(11), 1249-1258.e2. https://doi.org/10.1016/j.str.2020.08.005

Sovinova, O., Tumova, B., Pouska, F., & Nemec, J. (1957). [Isolation of virus responsible for respiratory diseases in horses]. *Ceskoslovenska epidemiologie, mikrobiologie, imunologie*, *6*(4), 213–220. http://www.ncbi.nlm.nih.gov/pubmed/13472742

Sovinova, O., Tumova, B., Pouska, F., & Nemec, J. (1958). Isolation of a virus causing respiratory disease in horses. *Acta Virologica*, *2*(1), 52–61.

Spielman, S. J., Weaver, S., Shank, S. D., Magalis, B. R., Li, M., & Kosakovsky Pond, S. L. (2019). Chapter 14 Evolution of Viral Genomes: Interplay Between Selection, Recombination, and

Other Forces. *Methods in Molecular Biology*, *1910*. https://doi.org/10.1007/978-1-4939-9074-0_14

Stack, J. C., Murcia, P. R., Grenfell, B. T., Wood, J. L. N. N., Holmes, E. C., Conrad Stack, J., Murcia, P. R., Grenfell, B. T., Wood, J. L. N. N., Holmes, E. C., Stack, J. C., Murcia, P. R., Grenfell, B. T., Wood, J. L. N. N., Holmes, E. C., & Holmes, E. C. (2013). Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1750), 20122173. https://doi.org/10.1098/rspb.2012.2173

Stremlau, M., Owens, C. M., Perron, M. J., Kiessling, M., Autissier, P., & Sodroski, J. (2004). The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature*, *427*(6977), 848–853. https://doi.org/10.1038/nature02343

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vey016

Sunayana, S. J. (2019). *Investigation of Influenza B Virus Replication Potential in Swine Primary Respiratory Epithelial Cells and Phylodynamic Analysis of Equine Influenza A H3N8 Viruses*. South Dakota State University.

Taddese, B., Garnier, A., Deniaud, M., Pele, J., Bellenger, L., Becu, J.-M., & Chabbert, M. (2022). *Bios2cor: From Biological Sequences and Simulations to Correlation Analysis* (R package version 2.2.1).

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/10.1093/genetics/123.3.585

Team, R. C. (2022). *R: A Language and Environment for Statistical Computing*.

Team, S. D. (2022). *RStan: the R Interface to Stan* (R package version 2.21.7). https://mc-stan.org/

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V, Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

Theys, K., Libin, P., Pineda-Peña, A.-C. C., Nowé, A., Vandamme, A.-M. M., & Abecasis, A. B. (2018a). The impact of HIV-1 within-host evolution on transmission dynamics. *Current Opinion in Virology*, *28*, 92–101. https://doi.org/10.1016/j.coviro.2017.12.001

Theys, K., Libin, P., Pineda-Peña, A.-C. C., Nowé, A., Vandamme, A.-M. M., & Abecasis, A. B. (2018b). The impact of HIV-1 within-host evolution on transmission dynamics. *Current Opinion in Virology*, *28*, 92–101. https://doi.org/10.1016/j.coviro.2017.12.001

Theys, K., Libin, P., Pineda-Peña, A.-C. C., Nowé, A., Vandamme, A.-M. M., & Abecasis, A. B. (2018c). The impact of HIV-1 within-host evolution on transmission dynamics. *Current Opinion in Virology*, *28*, 92–101. https://doi.org/10.1016/j.coviro.2017.12.001

Thompson, K.-A., & Bennett, A. M. (2017). Persistence of influenza on surfaces. *J Hosp Infect*, *95*(2), 194–199. https://doi.org/10.1016/j.jhin.2016.12.003

To, K. K. W., Chan, K.-H., Li, I. W. S., Tsang, T.-Y., Tse, H., Chan, J. F. W., Hung, I. F. N., Lai, S.-T., Leung, C.-W., Kwan, Y.-W., Lau, Y.-L., Ng, T.-K., Cheng, V. C. C., Peiris, J. S. M., & Yuen, K.-Y. (2010). Viral load in patients infected with pandemic H1N1 2009 influenza A virus. *Journal of Medical Virology*, *82*(1), 1–7. https://doi.org/https://doi.org/10.1002/jmv.21664

Toh, X., Soh, M. L., Ng, M. K., Yap, S. C., Harith, N., Fernandez, C. J., Huangfu, T., Lien, M., Mee, S., Ng, K., Choo, S., Nurshilla, Y., Judith, C., & Taoqi, F. (2019). Isolation and characterization of equine influenza virus (H3N8) from an equine influenza outbreak in Malaysia in 2015. *Transbound Emerg Dis*, *66*(5), 1884–1893. https://doi.org/10.1111/tbed.13218

Tran, T. T., Phung, T. T. B., Tran, D. M., Bui, H. T., Nguyen, P. T. T., Vu, T. T., Ngo, N. T. P., Nguyen, M. T., Nguyen, A. T. V. H., & Nguyen, A. T. V. H. (2023). Efficient symptomatic treatment and viral load reduction for children with influenza virus infection by nasal-spraying Bacillus spore probiotics. *Scientific Reports*, *13*(1), 14789. https://doi.org/10.1038/s41598-023-41763-5

Tusche, C., Steinbrück, L., & McHardy, A. C. (2012). Detecting patches of protein sites of influenza a viruses under positive selection. *Molecular Biology and Evolution*, *29*(8), 2063–2071. https://doi.org/10.1093/molbev/mss095

Vahey, M. D., & Fletcher, D. A. (2019a). Influenza A virus surface proteins are organized to help penetrate host mucus. *ELife*, *8*. https://doi.org/10.7554/eLife.43764

Vahey, M. D., & Fletcher, D. A. (2019b). Low-Fidelity Assembly of Influenza A Virus Promotes Escape from Host Cells. *Cell*, *176*(1–2), 281-294.e19. https://doi.org/10.1016/j.cell.2018.10.056

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Ž\'\idek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database:

massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Varble, A., Albrecht, R. A., Backes, S., Crumiller, M., Bouvier, N. M., Sachs, D., García-Sastre, A., & Tenoever, B. R. (2014). Influenza a virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host and Microbe*, *16*(5), 691–700. https://doi.org/10.1016/j.chom.2014.09.020

Varble, A., Benitez, A. A., Schmid, S., Sachs, D., Shim, J. V., Rodriguez-Barrueco, R., Panis, M., Crumiller, M., Silva, J. M., Sachidanandam, R., & Tenoever, B. R. (2013). An in vivo RNAi screening approach to identify host determinants of virus Replication. *Cell Host and Microbe*. https://doi.org/10.1016/j.chom.2013.08.007

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2022). *Pareto Smoothed Importance Sampling*.

Verbist, B. M. P. P., Thys, K., Reumers, J., Wetzels, Y., Van Der Borght, K., Talloen, W., Aerssens, J., Clement, L., Thas, O., der Borght, K., Talloen, W., Aerssens, J., Clement, L., & Thas, O. (2014). VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics*, *31*(1), 94–101. https://doi.org/10.1093/bioinformatics/btu587

Vihinen, M., Torkkila, E., & Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics*, *19*(2), 141–149. https://doi.org/https://doi.org/10.1002/prot.340190207

Vijaykrishna, D., Mukerji, R., & Smith, G. J. D. (2015). RNA Virus Reassortment: An Evolutionary Mechanism for Host Jumps and Immune Evasion. *PLOS Pathogens*, *11*(7), 1–6. https://doi.org/10.1371/journal.ppat.1004902

Virmani, N., Bera, B. C., Shanumugasundaram, K., Singh, B. K., Gulati, B. R., Singh, R. K., & Vaid, R. K. (2011). Genetic analysis of the matrix and non-structural genes of equine influenza virus (H3N8) from epizootic of 2008-2009 in India. *Vet Microbiol*, *152*(1–2), 169–175. https://doi.org/10.1016/j.vetmic.2011.04.011

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339–D343. https://doi.org/10.1093/nar/gky1006

von Itzstein, M. (2007). The war against influenza: discovery and development of sialidase inhibitors. *Nature Reviews Drug Discovery*, *6*(12), 967–974. https://doi.org/10.1038/nrd2400

Voorhees, I. E. H., Lee, H., Allison, A. B., Lopez-Astacio, R., Goodman, L. B., Oyesola, O. O., Omobowale, O., Fagbohun, O., Dubovi, E. J., Hafenstein, S. L., Holmes, E. C., & Parrish, C. R. (2019). Limited Intrahost Diversity and Background

Evolution Accompany 40 Years of Canine Parvovirus Host Adaptation and Spread. *Journal of Virology*, *94*(1), 1162–1181. https://doi.org/10.1128/JVI.01162-19

Wakeley, J. (1996). The variance of pairwise nucleotide differences in two populations with migration. *Theoretical Population Biology*, *49*(1), 39–57. https://doi.org/10.1006/tpbi.1996.0002

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., García, M. L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., Oksanen, H. M., Orton, R. J., Robertson, D. L., Rubino, L., … Zerbini, F. M. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Archives of Virology*, *167*(11), 2429–2440. https://doi.org/10.1007/s00705-022-05516-5

Ward, C. L., Dempsey, M. H., Ring, C. J. A., Kempson, R. E., Zhang, L., Gor, D., Snowden, B. W., & Tisdale, M. (2004). Design and performance testing of quantitative real time PCR assays for influenza A and B viral load measurement. *Journal of Clinical Virology*, *29*(3), 179–188. https://doi.org/https://doi.org/10.1016/S1386-6532(03)00122-7

Wargo, A. R., & Kurath, G. (2012). Viral fitness: definitions, measurement, and current insights. *Current Opinion in Virology*, *2*(5), 538–545. https://doi.org/10.1016/j.coviro.2012.07.007

Wasik, B. R., Voorhees, I. E. H., Barnard, K. N., Alford-Lawrence, B. K., Weichert, W. S., Hood, G., Nogales, A., Martínez-Sobrido, L., Holmes, E. C., & Parrish, C. R. (2019). Influenza Viruses in Mice: Deep Sequencing Analysis of Serial Passage and Effects of Sialic Acid Structural Variation. *Journal of Virology*, *93*(23). https://doi.org/10.1128/JVI.01039-19

Waters, L., Ahmed, N., Angus, B., Boffito, M., Bower, M., Churchill, D., Dunn, D., Edwards, S., Emerson, C., & Fidler, S. (2016). BHIVA guidelines for the treatment of HIV-1-positive adults with antiretroviral therapy 2015 (2016 interim update). *BHIVA) BHA, Ed. London, UK: BHIVA*. https://www.bhiva.org/file/RVYKzFwyxpgIl/treatment-guidelines-2016-interim-update.pdf

Watson, J., Halpin, K., Selleck, P., Axell, A., Bruce, K., Hansson, E., Hammond, J., Daniels, P., & Jeggo, M. (2011). Isolation and characterisation of an H3N8 equine influenza virus in Australia, 2007. *Aust Vet J*, *89 Suppl 1*, 35–37. https://doi.org/10.1111/j.1751-0813.2011.00738.x

Webb, G. F., Hsieh, Y.-H., Wu, J., & Blaser, M. J. (2010). Pre-symptomatic Influenza Transmission, Surveillance, and School Closings: Implications for Novel Influenza A (H1N1) Pre-symptomatic influenza transmission. *Math. Model. Nat.*

Phenom, *5*(3), 191–205. https://doi.org/10.1051/mmnp/20105312

Webster, R. G., & Laver, W. G. (1980). Determination of the number of nonoverlapping antigenic areas on Hong Kong (H3N2) influenza virus hemagglutinin with monoclonal antibodies and the selection of variants with potential epidemiological significance. *Virology*, *104*(1), 139–148. https://doi.org/10.1016/0042-6822(80)90372-4

Weng, X., Heiden, J. Vander, Xing, Y., Liu, J., & Vissa, V. (2011). Transmission of leprosy in Qiubei County, Yunnan, China: Insights from an 8-year molecular epidemiology investigation. *Infection, Genetics and Evolution*, *11*(2), 363–374. https://doi.org/10.1016/J.MEEGID.2010.11.014

Whitlock, F., Rash, A., & Elton, D. (2018a). Equine influenza: evolution of a highly infectious virus. *Veterinary Record*, *182*(25), 710–711. https://doi.org/https://doi.org/10.1136/vr.k2727

Whitlock, F., Rash, A., & Elton, D. (2018b). Equine influenza: evolution of a highly infectious virus. *Veterinary Record*, *182*(25), 710–711. https://doi.org/https://doi.org/10.1136/vr.k2727

WHO. (2023). *Pandemic Influenza Preparedness Framework: Partnership Contribution High-Level Implementation Plan III 2024-2030*. https://creativecommons.org/licenses/by-nc-sa/3.0/igo/

Wiley, D. C., & Skehel, J. J. (1987). The structure and function of the haemagglutinin membrane glycoprotein of influenza virus. *Structure*. https://doi.org/10.1146/annurev.bi.56.070187.002053

Wiley, D. C., Wilson, I. A., & Skehel, J. J. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, *289*(5796), 373–378. https://doi.org/10.1038/289373a0

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L., & Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. https://doi.org/10.1093/nar/gks918

Wilson, I. A., Skehel, J. J., & Wiley, D. C. (1981). Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. *Nature*, *289*(5796), 366–373. https://doi.org/10.1038/289366a0

Wißmann, J. E., Kirchhoff, L., Brüggemann, Y., Todt, D., Steinmann, J., & Steinmann, E. (2021). Persistence of Pathogens on Inanimate Surfaces: A Narrative Review. *Microorganisms*, *9*(2). https://doi.org/10.3390/microorganisms9020343

Wood, J. L. N., & Grenfell, B. T. (2009). *Dynamics of Influenza*. *21645*(October), 726–728. https://doi.org/10.1126/science.1175980.Quantifying

Wood J. M., Mumford J., Folkers C., Scott A.M., Schild G.C. Studies with inactivated equine influenza vaccine. J Hyg (Lond). 1983 Jun;90(3):371-84. doi: 10.1017/s0022172400029004

Wood, J. M. (1993). "Frozen" evolution of equine influenza viruses? *Equine Vet J*, *25*(2), 87. https://doi.org/10.1111/j.2042-3306.1993.tb02912.x

Woodward, A., Rash, A. S., Medcalf, E., Bryant, N. A., & Elton, D. M. (2015). Using epidemics to map H3 equine influenza virus determinants of antigenicity. *Virology*, *481*, 187–198. https://doi.org/10.1016/j.virol.2015.02.027

Woolhouse, M., Chase-Topping, M., Haydon, D., Friar, J., Matthews, L., Hughes, G., Shaw, D., Wilesmith, J., Donaldson, A., Cornell, S., Keeling, M., & Grenfell, B. (2001). Foot-and-mouth disease under control in the UK. *Nature*, *411*(6835), 258–259. https://doi.org/10.1038/35077149

Woolhouse, M. E. J. J., Haydon, D. T., & Antia, R. (2005). Emerging pathogens: The epidemiology and evolution of species jumps. *Trends in Ecology and Evolution*, *20*(5), 238–244. https://doi.org/10.1016/j.tree.2005.02.009

Wu, N. C., & Wilson, I. A. (2020). Influenza hemagglutinin structures and antibody recognition. *Cold Spring Harbor Perspectives in Medicine*, *10*(8), 1–20. https://doi.org/10.1101/cshperspect.a038778

Wu, Y., Wu, Y., Tefsen, B., Shi, Y., & Gao, G. F. (2014). Bat-derived influenza-like viruses H17N10 and H18N11. *Trends in Microbiology*, *22*(4), 183–191. https://doi.org/10.1016/j.tim.2014.01.010

Wyper, G. M. A., Fletcher, E., Grant, I., McCartney, G., Fischbacher, C., Harding, O., Jones, H., de Haro Moro, M. T., Speybroeck, N., Devleesschauwer, B., & Stockton, D. L. (2022). Measuring disability-adjusted life years (DALYs) due to COVID-19 in Scotland, 2020. *Archives of Public Health*, *80*(1), 105. https://doi.org/10.1186/s13690-022-00862-x

Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., Dai, Z., Lam, T. T., Guan, Y., & Yu, G. (2022). Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *IMeta*, *1*(4), e56. https://doi.org/10.1002/imt2.56

Xue, K. S., & Bloom, J. D. (2019). Reconciling disparate estimates of viral genetic diversity during human influenza infections. In *Nature Genetics* (Vol. 51, Issue 9, pp. 1298–1301). Nature Publishing Group. https://doi.org/10.1038/s41588-019-0349-3

Xue, K. S., & Bloom, J. D. (2020). Linking influenza virus evolution within and between human hosts. *Virus Evolution*, *6*(1). https://doi.org/10.1093/ve/veaa010

Xue, K. S., Moncla, L. H., Bedford, T., & Bloom, J. D. (2018). Within-Host Evolution of Human Influenza Virus. *Trends in Microbiology*, *26*(9), 781–793. https://doi.org/10.1016/j.tim.2018.02.007

Xue, K. S., Stevens-Ayers, T., Campbell, A. P., Englund, J. A., Pergam, S. A., Boeckh, M., & Bloom, J. D. (2017). Parallel evolution of influenza across multiple spatiotemporal scales. *ELife*, *6*. https://doi.org/10.7554/eLife.26875

Yan, N., & Chen, Z. Z. J. (2012). Intrinsic antiviral immunity. *Nature Immunology*, *13*(3), 214–222. https://doi.org/10.1038/ni.2229

Yang, R., Sun, H., Gao, F., Luo, K., Huang, Z., Tong, Q., Song, H., Han, Q., Liu, J. J. J., Lan, Y., Qi, J., Li, H., Chen, S., Xu, M., Qiu, J., Zeng, G., Zhang, X., Huang, C., Pei, R., … Liu, J. J. J. (2022). Human infection of avian influenza A H3N8 virus and the viral origins: a descriptive study. *The Lancet Microbe*, *3*(11), e824–e834. https://doi.org/10.1016/S2666-5247(22)00192-6

Yang, X., Charlebois, P., Macalalad, A., Henn, M. R., & Zody, M. C. (2013). V-Phaser 2: Variant inference for viral populations. *BMC Genomics*, *14*(1), 674. https://doi.org/10.1186/1471-2164-14-674

Yewdell, J. W., & Santos, J. J. S. (2021). Original Antigenic Sin: How Original? How Sinful? *Cold Spring Harbor Perspectives in Medicine*, *11*(5), a038786. https://doi.org/10.1101/cshperspect.a038786

Yondon, M., Heil, G. L., Burks, J. P., Zayat, B., Waltzek, T. B., Jamiyan, B.-O., Mckenzie, P. P., Krueger, W. S., Friary, J. A., Gray, G. C., & Gray, C. (2013). Isolation and characterization of H3N8 equine influenza A virus associated with the 2011 epizootic in Mongolia. *Influenza Other Respir Viruses*, *7*(5), 659–665. https://doi.org/10.1111/irv.12069

Yongfeng, Y., Xiaobo, S., Nan, X., Jingwen, Z., Wenqiang, L., Xiaobo, S., & Nan, X. (2020). Detection of the epidemic of the H3N8 subtype of the equine influenza virus in large-scale donkey farms. *Int J Vet Sci Med*, *8*(1), 26–30. https://doi.org/10.1080/23144599.2020.1739844

Yoon, S.-W., Webby, R. J., & Webster, R. G. (2014). Evolution and Ecology of Influenza A Viruses. In *Curr Top Microbiol Immunol* (Vol. 385, pp. 359–375). https://doi.org/10.1007/82_2014_396

Ypma, R. J. F., Bataille, A. M. A., Stegeman, A., Koch, G., Wallinga, J., & van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1728), 444–450. https://doi.org/10.1098/rspb.2011.0913

Yu, F., Wen, Y., Wang, J., Gong, Y., Feng, K., Ye, R., Jiang, Y., Zhao, Q., Pan, P., Wu, H., Duan, S., Su, B., & Qiu, M. (2018). The Transmission and Evolution of HIV-1 Quasispecies within One Couple: a Follow-up Study based on Next-Generation Sequencing OPEN. *SCIENTIfIC REPoRTs |*, *8*, 1404. https://doi.org/10.1038/s41598-018-19783-3

Zarnitsyna, V. I., Ellebedy, A. H., Davis, C., Jacob, J., Ahmed, R., & Antia, R. (2015). Masking of antigenic epitopes by antibodies shapes the humoral immune response to influenza. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1676). https://doi.org/10.1098/RSTB.2014.0248

Zhang, T., Wu, Q., & Zhang, Z. (2020). Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology*, *30*(7), 1346-1351.e2. https://doi.org/10.1016/j.cub.2020.03.022

Zhang, Y. Z., & Holmes, E. C. (2020). A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell*, *181*(2), 223–227. https://doi.org/10.1016/j.cell.2020.03.035

Zhao, L., Abbasi, A. B., & Illingworth, C. J. R. R. (2019). Mutational load causes stochastic evolutionary outcomes in acute RNA viral infection. *Virus Evolution*, *5*(1), 1–12. https://doi.org/10.1093/ve/vez008

Zhao, L., & Illingworth, C. J. R. R. (2019a). Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evolution*, *5*(1), 1–7. https://doi.org/10.1093/ve/vey041

Zhao, L., & Illingworth, C. J. R. R. (2019b). Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evolution*, *5*(1), 1–7. https://doi.org/10.1093/ve/vey041

Zhou, B., Donnelly, M. E., Scholes, D. T., St. George, K., Hatta, M., Kawaoka, Y., & Wentworth, D. E. (2009). Single-Reaction Genomic Amplification Accelerates Sequencing and Vaccine Production for Classical and Swine Origin Human Influenza A Viruses. *Journal of Virology*, *83*(19), 10309–10313. https://doi.org/10.1128/JVI.01109-09

Zhu, H., Damdinjav, B., Gonzalez, G., Patrono, L. V., Ramirez-Mendoza, H., Amat, J. A. R. R., Crispell, J., Parr, Y. A., Hammond, T.-A. A., Shiilegdamba, E., Leung, Y. H. C. C., Peiris, M., Marshall, J. F., Hughes, J., Gilbert, M., & Murcia, P. R. (2019). Absence of adaptive evolution is the main barrier against influenza emergence in horses in Asia despite frequent virus interspecies transmission from wild birds. *PLoS Pathogens*, *15*(2), 1–23. https://doi.org/10.1371/journal.ppat.1007531

Zwart, M. P., & Elena, S. F. (2015). Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. *Annual Review of Virology*, *2*, 161–179. https://doi.org/10.1146/annurev-virology-100114-055135