Crossing host boundaries: the evolutionary drivers and correlates of viral host jumps

Cedric C.S. Tan[1,2], Lucy van Dorp[1,*], Francois Balloux[1,*]

[1]UCL Genetics Institute, University College London, Gower St, London WC1E 6BT, UK.
[2]The Francis Crick Institute, 1 Midland Rd, London NW1 1AT, UK.

* These authors contributed equally.
Correspondence: C.C.S.T (cedriccstan@gmail.com)

Abstract (224 words)

Most emerging and re-emerging infectious diseases stem from viruses that naturally circulate in non-human vertebrates. When these viruses cross over into humans, they can cause disease outbreaks and epidemics. While zoonotic host jumps have been extensively studied from an ecological perspective, little attention has gone into characterising the evolutionary drivers and correlates underlying these events. To address this gap, we harnessed the entirety of publicly available viral genomic data, employing a comprehensive suite of network and phylogenetic analyses. We address a series of questions concerning the evolutionary mechanisms underpinning viral host jumps. Notably, we challenge conventional assumptions about the directionality of host jumps, demonstrating that humans are as much a source as a sink for viral spillover events, insofar we infer more viruses to have jumped from humans to other animals, than from animals to humans. Moreover, we demonstrate heightened evolution in viral lineages that involve putative host jumps. We further observe that the mutational threshold associated with a host jump is lower for viruses with broad host ranges. Finally, we show that the genomic targets of natural selection upon a successful host jump vary across different viral families with either structural or auxiliary genes being the prime targets of selection. Collectively, our results illuminate some of the evolutionary drivers underlying viral host jumps that may contribute to mitigating viral threats across species boundaries.

Introduction

The majority of emerging and re-emerging infectious diseases in humans are caused by viruses that have jumped from wild and domestic animal populations into humans (i.e., zoonoses)[1]. Zoonotic viruses have caused countless disease outbreaks ranging from isolated cases to pandemics and have taken a major toll on human health throughout history. There is a pressing need to develop better approaches to pre-empt the emergence of viral infectious diseases and mitigate their effects. As such, there is an immense interest in understanding the correlates and mechanisms of zoonotic host jumps[1–10].

Most studies thus far have primarily investigated the ecological and phenotypic risk factors contributing to host-range based on host-virus association databases constructed mainly based on systematic literature reviews and online compendiums, including VIRION[11] and CLOVER[12]. For example, 'generalist' viruses that can infect a broader range of hosts have typically been shown to be associated with greater zoonotic potential[3,5,13]. Additionally, factors such as increasing human population density[1], alterations in human-related land use[4], ability to replicate in the cytoplasm or being vector-borne[3] tend to be positively associated with zoonotic risk. However, despite global efforts to understand the risk factors underlying the emergence of viral infectious diseases, our current understanding remains insufficient to effectively predict, prevent and manage imminent and future infectious disease threats. This may partly stem from the lack of integration of genomics into these ecological and phenotypic analyses.

One challenge to evolutionary studies predicting infectious diseases emergence is that only a small fraction of the viral diversity circulating in wild and domestic vertebrates has been characterised to date. Due to resource and logistical constraints, surveillance studies of novel pathogens in animals often have sparse geographical and/or temporal coverage[14,15], as they focus on selected host and pathogen taxa. Further, many of these studies do not perform downstream characterisation of the novel viruses recovered, and may lack sensitivity due to the use of PCR pre-screening to prioritise samples for

sequencing[16]. As such, our understanding of which viruses can, or are likely to emerge, and in which settings is poor. Additionally, while genomic analyses are important for exploring the drivers of viral host jumps[17], few studies thus far have incorporated genomic data into broad-scale analyses exploring the drivers and correlates of zoonotic risk. The studies that have considered genomic data have mostly focused on measures of host[13] or viral[3] diversity as predictors of zoonotic risk. As such, despite the limited characterisation of global viral diversity thus far, existing genomic databases remain a rich, largely untapped, resource to better understand the evolutionary processes surrounding viral host jumps.

Further, while research efforts have rightfully focused on zoonotic host jumps, humans are just one node in a large and complex network of host species in which viruses are endlessly exchanged, with viral zoonoses representing likely only rare outcomes of this wider ecological network. As such, viral host jumps between non-human animals remain relatively understudied. Another important process that has received less attention is human-to-animal (i.e., anthroponotic) spillover, which may impede biodiversity conservation efforts and could also negatively impact food security. For example, human-sourced metapneumovirus has caused fatal respiratory outbreaks in captive chimpanzees[18]. Anthroponotic events may also lead to the establishment of wild animal reservoirs that may re-seed infections in the human population, potentially following the acquisition of animal-specific adaptations that could increase transmissibility or pathogenicity of a virus in humans[19]. Uncovering the broader evolutionary processes surrounding host jumps across vertebrate species may therefore enhance our ability to pre-empt and mitigate the effects of infectious diseases on both human and animal health.

One major challenge for understanding macroevolutionary processes through large-scale genomic analyses is the traditional reliance on physical and biological properties of viruses to define viral taxa, which is largely a vestige of the pre-genomic era[20]. As a result, taxon names may not accurately reflect the evolutionary relatedness of viruses, precluding robust comparative analyses involving multiple viral taxa. Notably, the

International Committee on Taxonomy of Viruses (ICTV) has been strongly advocating for taxon names to also reflect the evolutionary history of viruses[20,21]. However, the increasing use of metagenomic sequencing technologies has resulted in a large influx of newly-discovered viruses that have not yet been incorporated in to the ICTV taxonomy. It also remains challenging to formally assess genetic relatedness through multi-sequence alignments of thousands of sequences comprising diverse viral taxa, particularly for highly recombining or poorly annotated taxa.

In this study, we leveraged the ~11 million viral sequences and associated host metadata hosted on NCBI to assess the current state of global viral genomic surveillance. We additionally analysed ~54k viral sequences isolated from various vertebrate hosts, using a bespoke approach that is agnostic to viral taxonomy, to understand the evolutionary processes surrounding host jumps. We ascertain overall trends in the directionality of viral host-jumps between human and non-human vertebrates and quantify the amount of detectable adaptation associated with putative host jumps. Finally, we quantify, for a subset of viruses, signatures of adaptive evolution detected in specific categories of viral proteins associated with facilitating or sustaining host-jumps. Together, we provide a comprehensive assessment of potential genomic correlates underpinning host-jumps in viruses across human and non-human vertebrates.

Results

An incomplete picture of global vertebrate viral diversity

Global genomic surveillance of viruses from different hosts is key to preparing for emerging and re-emerging infectious diseases in humans and animals[17,19]. To identify the scope of viral genomic data collected thus far, we downloaded the metadata of all viral sequences hosted on NCBI Virus (*n*=11,645,803; accessed 22 July 2023; **Supplementary Table 1**). Most (68%) of these sequences were associated with SARS-CoV-2, reflecting the intense sequencing efforts during the COVID-19 pandemic. Additionally, of these sequences, 93.6%, 3.3%, 1.5%, 1.1% and 0.6% were of viruses with single-stranded (ss) RNA, double-stranded (ds) DNA, dsRNA, ssDNA and unspecified genome compositions, respectively.

Genomic surveillance has thus far been highly human-centric, with 91% of publicly available viral sequences obtained from human-associated samples (**Fig. 1a**). Only a limited number of non-human vertebrate families have at least ten associated viral genome sequences deposited (**Fig. 1b**), reinforcing that a substantial proportion of viral diversity in vertebrates remains uncharacterised. Viral sequences collected thus far also display a strong geographic bias, with most virus samples collected from the United States of America (USA) and China, whereas countries in Africa, Central Asia, South America and Eastern Europe are highly underrepresented (**Fig. 1c**). Finally, the user-submitted host metadata associated with viral sequences, which is key for understanding global trends in the evolution and spread of viruses in wildlife, remains poor, with 45% of non-human viral sequences having no host information provided at the genus level. Overall, these results highlight the massive gaps in the genomic surveillance of viruses in wildlife globally and the need for more conscientious reporting of sample metadata. Publicly available viral genomic data from wild and domestic animals is at this stage far too sparse to be used for a comprehensive understanding of the full impact of vertebrate viruses on human and animal health.

Humans give more viruses to animals than they do to us

To investigate the relative frequency of anthroponotic and zoonotic host jumps, we retrieved 53,631 quality-controlled viral genomes spanning 30 viral families and associated to 61 vertebrate host orders. We found that the user-submitted species identifiers of viruses in public sequence databases are poorly ascribed, with only 37% of species names consistent with those in the ICTV viral taxonomy[22]. Additionally, the genetic diversity represented by different viral species is highly variable, since they are conventionally defined based on the genetic, phenotypic and ecological attributes of viruses[20]. Thus, we implemented a species-agnostic approach based on network theory to define 'viral cliques' that represent discrete taxonomic units with similar degrees of genetic diversity (see Methods). A similar approach was previously shown to effectively partition the genomic diversity of plasmids in a biologically-relevant manner[23]. Using this approach, we identified 4,937 viral cliques across 30 viral families. Some clique assignments aggregated multiple viral species identifiers, while others disaggregated species into multiple cliques (**Fig. 2**; clique assignments for *Coronaviridae* illustrated in **Extended Data Fig. 1**).

We then identified putative host jumps within these viral cliques by producing curated whole genome alignments (or single-gene alignments for segmented viruses) to which we applied maximum-likelihood phylogenetic reconstruction. Phylogenetic trees were rooted with suitable outgroups identified using metrics of alignment-free distances (see Methods). We subsequently reconstructed the host states of all ancestral nodes in each phylogenetic tree, allowing us to determine the most probable direction of a host jump for each viral sequence (approach illustrated in **Fig. 3a**). In total, we identified 2,722 putative host jumps between vertebrates in 163 of these viral cliques.

Amongst the putative host jumps inferred to involve human hosts (514/2,722; 19%), we found a significantly higher frequency of anthroponotic compared to zoonotic host jumps (67% vs. 33%, respectively; paired t-test, t=256, d.f.=999, p<0.0001; **Fig. 3b**). The highest number of anthroponotic jumps were contributed by the cliques representing SARS-CoV-2 (132/343; 38%), MERS-CoV (39/343; 11%) and Influenza A (36/343;

10%). This is concordant with the repeated independent anthroponotic spillovers into farmed, captive and wild animals, described for SARS-CoV-2[19,24–27] and influenza A[28,29]. Meanwhile, there has only been circumstantial evidence for human-to-camel transmission of MERS-CoV[30–32]. Noting the disproportionate number of anthroponotic jumps contributed by the SARS-CoV-2 clique, we re-performed the analysis without this clique though found similar relative frequencies of inferred anthroponotic versus zoonotic jumps (64% vs. 36%; paired t-test, t=132, d.f.=999, p<0.0001). Overall, these results highlight the high, but largely underappreciated frequency of anthroponotic jumps amongst vertebrate viruses.

The degree of adaptation required for a host jump is less for multi-host viruses

Before jumping to a new host, a virus in its natural reservoir may fortuitously acquire pre-adaptive mutations that facilitate its transition to a new host. This may be followed by the further acquisition of adaptive mutations as the virus adapts to its new host environment[17]. In practice, the degree of adaptation inferred may vary based on different factors, including sampling intensity and the time interval between when the host jump occurred and when the virus was isolated from its new host. As such, for each viral clique, we considered only the minimum extent of adaptation associated with a host jump. We first examined whether the minimum mutational distance associated with a host jump for each viral clique was higher than that for a random selection of viral lineages not involved in host jumps (**Fig. 3a**; see Methods). Indeed, the minimum mutational distance for a putative host jump within each clique was significantly higher than that for non-host jumps (**Fig. 4a**; one-tailed Mann-Whitney U test, U=6767, p=6e-12). Noting that the different mutation rates of viral families may confound these results, we corrected for the effects of viral family affiliation using a logistic regression model but found a similar effect ($OR_{host\ jump}$=1.39; two-tailed Z-test for slope=0 Z=7.37, d.f.=283, p=1e-13). We then considered the commonly used measure of directional selection acting on genomes, the ratio of non-synonymous mutations per non-synonymous site (dN) to the number of synonymous mutations per synonymous site (dS). Comparing the minimum dN/dS for host jumps within each clique, we observed that minimum dN/dS was also significantly higher for host jumps compared to non-host jumps (**Fig. 4b**;

$OR_{host\ jump}$=2.29; *Z*=4.70, d.f.=267, *p*=3e-6). These results are consistent with the hypothesised need for heightened selection following a change in host environment and additionally provide confidence in our ancestral state reconstruction method for assigning host-jump status.

However, the extent of adaptive change required for a viral host-jump may vary. For instance, some zoonotic viruses may require minimal adaptation to infect new hosts while in other cases, more substantial genetic changes might be necessary for the virus to overcome barriers that prevent efficient infection or transmission in the new host. We therefore tested the hypothesis that the strength of selection associated with a host jump decreases for viruses which tend to have broader host ranges. To do so, we compared the minimum mutational distance between ancestral and observed host states to the number of host genera sampled for each viral clique. We found that the observed host range for each viral clique is positively associated with more intense sequencing effort (i.e., number of viral genomes in each clique; Pearson's *r*=0.505; two-tailed t-test for *r*=0, *t*=10.3, d.f.=310, *p*<0.0001), similar to the strong correlation between per-host viral diversity and surveillance effort reported in previous studies[3,8,13]. After correcting for both sequencing effort and viral family membership, we found that the mutational threshold for host jumps tends to decrease with broader host ranges (Poisson regression, slope=-0.110; two-tailed Z-test for slope=0, *Z*=-9.31, d.f.=127, *p*<0.0001). In contrast, the relationship between mutational threshold and host range for viral lineages which have not experienced host jumps is only weakly positive (slope=0.0843; *Z*=7.16, d.f.=127, *p*=8e-13) (**Fig. 4c**). Similarly, the minimum dN/dS for a host jump decreases significantly for viral cliques with broader host ranges (slope=-0.415; *Z*=-9.56, d.f.=120, *p*<0.0001), which was not the case for the non-host jump controls (slope=0.252; *Z*=5.49, d.f.=118, *p*=4e-8) (**Fig. 4d**). The effects of mutational and dN/dS thresholds were consistent when the same analysis was performed for ssDNA, dsDNA, +ssRNA and -ssRNA viruses separately (**Extended Data Fig. 2**). These results indicate that, on average, 'generalist' multi-host viruses have lower adaptation barriers for jumping into new vertebrate hosts.

Host jump-associated adaptations are gene and region specific and vary by viral family

We next examined whether genes with different established functions displayed distinctive patterns of adaptive evolution linked to host jump events. Since gene function remains poorly characterised in the large and complex genomes of double-stranded DNA viruses, we focused on the shorter single-stranded (ss) RNA and ssDNA viral families. We selected for analysis the four viral families with the greatest number of putative host jumps in our dataset: *Coronaviridae* (+ssRNA; n=2,229), *Rhabdoviridae* (-ssRNA; *n*=1,096), *Paramyxoviridae* (-ssRNA; *n*=762), and *Circoviridae* (ssDNA; *n*=686). For these viral families, we extracted all annotated protein-coding regions from their genomes and categorised them as either being associated with cell entry (termed 'entry'), viral replication ('replication-associated'), or virion formation ('structural'), and classifying the remaining genes as 'auxiliary' genes. To estimate putative signatures of adaptation in relation to lineages which have experienced host jumps for the different gene categories, we computed the change in log10(dN/dS) in host jumps versus non-host jumps using a linear model, while correcting for the effects of clique membership (see Methods). Contrary to our expectation, that entry genes would be under the strongest adaptive pressures during a host jump, we found that the degree of selection inferred was strongest for structural genes in coronaviruses and paramyxoviruses, and for auxiliary genes in rhabdoviruses and circoviruses (**Fig. 5**). Additionally, the inferred strength of selection acting on the different gene categories varied by viral family, with structural genes showing the strongest signatures of adaptation associated to host jumps for *Coronaviridae* and *Paramyxoviridae* but not for *Rhabdoviridae* and *Circoviridae* (**Fig. 5**).

We further hypothesised that within each gene, adaptative changes are likely to be localised to regions of functional importance and/or that are under relatively stronger selective pressures exerted by host immunity. To test this, we focused on the spike gene (entry) of the viral clique that includes SARS-CoV-2-related viruses since the key region involved in viral entry is well-characterised (i.e., the receptor-binding domain (RBD))[33]. We found that dN/dS estimates consistent with adaptive evolution were indeed localised to the first half of the spike gene containing the RBD (**Extended Data**

Fig. 3a), consistent with the strong immune pressures exerted on this region of the spike entry protein[34]. Further, the dN/dS estimates were significantly higher in the spike protein RBD compared to the other spike domains (two-sided Mann-Whitney U test, U=429,827, $p<0.0001$; **Extended Data Fig. 3b**), likely reflecting its central role in host-cell recognition and entry. Overall, our results indicate that adaptive changes associated to a switch to a new host differ by gene function and region, and may vary by viral family.

Discussion

The post-genomic era has opened opportunities to advance our understanding of the diversity of viruses in circulation, and the macroevolutionary principles of viral host range. Leveraging ~56k publicly available viral sequences isolated from vertebrate hosts, we inferred that humans give more viruses to other vertebrates than they give to us across the 30 viral families we considered. We further demonstrated that host jumps are associated with heightened signals of adaptive evolution that tends to decrease in viruses with broader host ranges. This indicates that there may be a minimum mutational threshold necessary for viruses to expand their host range. Finally, we showed that adaptive evolution linked to host jumps may vary by gene function and may be localised to specific gene regions of functional importance.

To bypass the limitations of existing viral taxonomies, we used a taxonomy-agnostic approach to define roughly equivalent units of viral diversity, which formed the basis for most of the analyses presented in this study. The use of quantitative taxonomic units rather than traditional taxonomic species names further allowed us to perform like-for-like analyses across the entire diversity of viruses. Our approach identified cliques that were largely concordant with traditional viral species nomenclature, but also highlighted inconsistencies, where in some cases, single viral species appear to form distinct taxonomic groups while other groups of species seem to form a single group solely based on their genetic relatedness (**Fig. 2**; **Extended Data Fig. 1**). However, we do not claim that our approach should supersede existing taxonomic classification systems, especially since a robust and meaningful species definition requires the integration of viral properties with finer-scale evolutionary analyses that was not necessary for our purposes. Nevertheless, we anticipate that the use of similar network-based approaches may pave the way for the development of efficient classification frameworks that can rapidly incorporate novel, metagenomically-derived viruses into existing taxonomies.

Harnessing cliques as a mechanism of identifying clusters of related viruses for phylogenetic inspection allowed us to quantify the number and sources of recent host

jump events. Of the 514 putative host jumps identified, 67% were inferred as anthroponotic (**Fig. 3b**). While the relative importance of anthroponotic versus zoonotic events has been speculated[19,29,35,36], we provide a formal evaluation of the zoonotic-to-anthroponotic ratio in vertebrates, showing that anthroponoses are equally, if not more, critical to consider than zoonoses when assessing viral spillover dynamics. It stands to reason that the substantial global human population size and ubiquitous spatial distribution position them as a major source for viral exchange. However, it is also likely that behavioural factors might amplify the risk of anthroponotic transmission, for example through changes in land use, agricultural methods, or heightened interactions between humans and wildlife[4]. Overall, our results highlight the importance of surveying and monitoring the human-to-animal transmission of viruses.

We observed heightened evolution and adaptive signals in association with host jumps (**Fig. 4**). This result is largely intuitive, since a virus jumping into a new host is likely to be under different selective pressures exerted directly by the novel host environment and indirectly by changes in host-to-host transmission dynamics. The evolutionary signals we captured may include pre-requisite adaptations that enable a virus to infect the new host. Additionally, they likely also represent the burst of adaptive mutations which may be acquired following a host jump, which has been demonstrated for multiple viral systems[24,37–39]. Further, these signals could potentially reflect a relaxation of prior selective pressures no longer present in the novel host. We note that these signals of heightened evolution could also, in principle, be inflated by sampling bias, where two viruses circulating in the same host are more often drawn from the same population. However, this was largely controlled for in our analysis through comparisons to representative non-host jump lineages that are expected to be affected by the same sampling bias.

We observed a lower mutational and adaptive threshold associated with host jumps for viruses known to infect a broader range of hosts (**Fig. 4c** and **d**). The most likely explanation for this pattern is that some viruses are intrinsically more capable of infecting a diverse range of hosts, possibly by exploiting host-cell machinery that are

conserved across different hosts. For example, sarbecoviruses (the subgenus comprising SARS-CoV-2) target the ACE2 host-cell receptor, which is conserved across vertebrates[40,41], and the high structural conservation of the sarbecovirus spike protein[42], may explain the observation that single mutations can enable sarbecoviruses to expand their host tropism[43]. In other words, some viruses possess 'generalist' infection mechanisms that reduces the mutational barrier for them to productively infect new hosts. This may provide a mechanistic explanation for previous observations that viruses with broad host range have a higher risk of emerging as zoonotic diseases[3,5,13].

Our approach to identifying putative host jumps hinges on ancestral state reconstruction (**Fig. 3a**), which has been shown to be affected by sampling biases[44,45]. However, we accounted for this, at least in part, by including sequencing effort as a measure of sampling bias in our statistical models, allowing us to draw inferences that were robust to disproportionate sampling of viruses in different hosts. Our approach also does not consider the epidemiology or ecology of viral transmission, as this is largely dependent on host features such as population size, social structure and behaviour for which comprehensive datasets at this scale are not currently available. We anticipate that future datasets that integrate ecology, epidemiology and genomics may allow finer-scale investigation of these patterns in specific host and viral systems. Additionally, the patterns we describe are broad and do not capture the idiosyncrasies of individual host-pathogen associations. Those include a variety of biological features, intrinsic ones, such as the molecular adaptations required for receptor binding, as well as more complex ones including cross-immunity and interference with other viral pathogens circulating in a host population.

Overall, our work highlights the large scope of genomic data in the public domain, and its utility in exploring the evolutionary mechanisms of viral host jumps. However, the large gaps in the genomic surveillance of viruses thus far suggests that we have only just scratched the surface of the true viral diversity in nature. Additionally, despite the strong anthropocentric bias in viral surveillance, 81% of the putative host jumps identified in this study do not involve humans, emphasising the large under-appreciated

scale of the global viral-sharing network. Widening our field of view beyond zoonoses and investigating the flow of viruses within this larger network could yield valuable insights that may help us prepare for and manage infectious disease emergence at the human-animal interface.

Methods

Data acquisition, curation and quality control

The metadata of all partial and complete viral genomes was downloaded from NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) on 22 July 2023, with filters excluding sequences isolated from environmental sources, lab hosts, or associated with vaccine strains or proviruses (*n*=11,645,803). Where possible, host taxa names in the metadata were resolved in accordance with the NCBI taxonomy using the *taxizedb* v0.3.1 package in R. User-submitted viral species names were compared to the ICTV master species list version 'MSL38.V2', dated 6 July 2023.

To generate a candidate list of viral sequences for further genomic analysis, the metadata was filtered to include 53 viral families known to infect vertebrate hosts based on information provided in the 2022 release of the ICTV taxonomy (https://ictv.global/taxonomy), and with reference to that provided by ViralZone (https://viralzone.expasy.org/). We then retained only sequences from viral families comprising at least 100 sequences of greater than 1000nt in length. Since the sequences of segmented viral families are rarely deposited as whole genomes, we identified sequences for genes conserved within each of these families for further analysis (*Arenaviridae*: L segment; *Birnaviridae*: ORF1/RdRP/VP1/Segment B; *Peribunyaviridae*: L segment; *Orthomyxoviridae*: PB1; *Picobirnaviridae*: RdRP). For non-segmented genomes, we retained all non-human-associated sequences, and subsampled the human-associated sequences as follows: we selected a random subsample of 1000 SARS-CoV-2 genomes of greater than 28000nt from distinct countries, isolation sources, and with distinct collection dates. For other human-associated sequences, we retained viruses with distinct species, country, isolation source and collection date information. We then downloaded the final candidate list of viral sequences (*n*=88,161) using the *ncbi-acc-download* v0.2.8 (https://github.com/kblin/ncbi-acc-download). Further quality control of the genomes downloaded was performed using *CheckV* v1.0.1[46], retaining sequences with more than 95% completeness (for non-segmented viruses) and less than 5% contamination (for all

sequences). This resulted in a final genomic dataset comprising 53,631 observations (**Supplementary Table 2**), comprising gene sequences for segmented viruses and complete genomes for non-segmented viruses. For simplicity, we will henceforth refer to the gene sequences and complete genomes as 'genomes'.

Taxonomy-agnostic identification of viral cliques

To identify viral cliques, we calculated the pairwise alignment-free Mash distances of genomes within each viral family via *Mash* v1.1[47] with a *k*-mer size of 13. This *k*-mer size ensures that the probability of observing a *k*-mer by chance given the median genome length for each clique is less than 0.01. Given a genome length, *l*, alphabet, Σ={A, T, G, C}, and the desired probability of observing a *k*-mer by chance, *q*=0.01, this was computed using the formula described previously[47]:

$$k = \left\lceil log_{|\Sigma|}\left(\frac{l(1-q)}{q}\right)\right\rceil$$

We then constructed undirected, weighted bipartite graphs for each viral family with nodes and edges representing genomes and Mash distances, respectively. We removed edges from these networks with Mash distance values greater than a certain threshold, *t*, before we applied the community-detection algorithm, Infomap[48]. This community-detection algorithm performs well in both large (>1000 nodes) and small (≤1000 nodes) bipartite graphs[49] and seeks to identify subgraphs within these weighted bipartite graphs that minimise the information required to constrain the movement of a random walker[48]. We refer to the subgraphs identified through this algorithm as 'viral cliques'. Here, we force the community-detection algorithm to identify taxonomically-relevant cliques by removing edges with Mash distance values greater than *t*, which results in sparser graphs with closely-related genomes (e.g. from the same species) being more densely connected than more distantly-related genomes (e.g. different species). The value of *t* was selected by maximising the concordance of the viral cliques identified with the viral species names from the NCBI taxonomy, based on the commonly used clustering performance metrics, adjusted mutual information (AMI) and adjusted Rand index (ARI). These metrics were computed using the *AMI* and *ARI* functions in *Aricode* v1.0.2. To further assess whether the viral cliques identified fulfil

the species definition criterion of being monophyletic[20], we reconstructed the phylogenies of each viral family by applying the neighbour-joining algorithm[50] implemented in the *Ape* v5.7.1 R package on their pairwise Mash distance matrices. We then computed the proportion of monophyletic viral cliques using the *is.monophyletic* function in *Ape* v5.7.1 across the various values of *t*. Given the discordance between the NCBI and ICTV taxonomies, we applied the above optimisation protocol to *t* using the viral species names in the ICTV taxonomy. Using the NCBI viral species names, *t*=0.15 maximised both the median AMI and ARI across all families (**Extended Data Fig. 4a**), with 94.3% of the cliques identified being monophyletic (**Extended Data Fig. 4b**). Using the ICTV viral species names, *t*=0.2 and *t*=0.25 maximised the median AMI and median ARI across families (**Extended Data Fig. 4c**), with 93.7% and 87.8% cliques being monophyletic (**Extended Data Fig. 4b**), respectively. Since *t*=0.15 produced the highest proportion of monophyletic clades that were approximately concordant with existing viral taxonomies, we used this threshold to generate the final viral clique assignments for downstream analyses (**Supplementary Table 2**)

Identification of putative host jumps

We retrieved all viral cliques that were associated with at least two distinct hosts and comprise at least 10 genomes (*n*=204). We then generated clique-level genome alignments using the *FFT-NS-2* algorithm in *MAFFT* v7.490[51,52]. We masked regions of the alignments that were poorly aligned or prone to sequencing-error by replacing alignment sites that had more than 10% of gaps or ambiguous nucleotides with N's. Clique-level genome alignments that had more than 20% of the median genome length masked were considered to be poorly aligned and removed from further analysis (*n*=6; **Extended Data Fig. 5**). Following this procedure, we reconstructed maximum-likelihood phylogenies for each viral clique with *IQ-Tree* v2.1.4-beta[53], using 1000 ultrafast bootstrap (UFBoot)[54] replicates. The optimal substitution model for each tree was automatically determined using the *ModelFinder*[55] utility native to *IQ-Tree*. To estimate the root position for each clique tree, we reconstructed neighbour-joining Mash trees for each viral clique, including ten additional genomes, whose minimum pairwise Mash distance to the genomes in each tree was 0.3-0.5, as potential outgroups. The most

basal tip in these neighbour-joining Mash trees were identified and used to root the maximum-likelihood clique trees.

To identify putative host jumps, we performed ancestral state reconstruction on the resultant rooted maximum-likelihood phylogenies with host as a discrete trait using the *ace* function in *Ape* v5.7.1. Traversing from a tip to the root node, a putative host jump is identified if the reconstructed host state of an ancestral node is different from the observed tip state and has a two-fold greater likelihood compared to alternative states, and is different from the host state of the sampled tip. Where the tip and ancestral host states were of different taxonomic ranks, we excluded putative host jumps where the ancestral host state is nested within the tip host state, or vice versa (e.g. '*Homo*' and '*Hominidae*'). Missing host metadata were encoded as 'unknown' and included in the ancestral state reconstruction analysis. Host jumps involving unknown or non-vertebrate host states were excluded from further analysis. Separately, we extracted non-host jump lineages to control for any biases in our analysis approach. To do so, we randomly selected an ancestral node – where the reconstructed host state is the same as the observed tip state and has a two-fold greater likelihood than alternative host states – for each viral genome that is not involved in any putative host-jumps. For the mutational distance and dN/dS analyses, we retained only viral cliques where non-host jump lineages could be identified (*n*=157).

Calculating mutational distances and dN/dS

For host jump lineages, mutational distances were calculated as the sum of the branch lengths between the tip sequence and the ancestral node for which the first host state transition occurred (in substitutions per site) using the *get_pairwise_distances* function in the *Castor* v1.7.10[56] R package, which was then multiplied by the alignment length to obtain the estimated number of substitutions (**Fig. 3a**). To calculate the dN/dS estimates, we extracted coding regions from the clique-level masked alignments based on the user-submitted gene annotations (in 'gff' format) of each viral genome. We then computed the dN/dS estimates using Li's[57] method implemented in the *dnastring2kaks* function from the *MSA2dist* v1.4.0 R package (https://github.com/kullrich/MSA2dist). We

then calculated the minimum mutational distance and dN/dS across all host jump events in each clique for our downstream statistical analyses, which, in principle, represents the minimum evolutionary signal associated with a host jump in each viral clique.

Statistical analyses

All statistical analyses were performed using the *stats* package native to R v4.3.1. To generate the bootstrapped distributions shown in **Fig. 3b**, we randomly resampled the host jumps within each clique with replacement (1000 iterations) and performed two-tailed paired t-tests using the *t.test* function. Mann-Whitney U tests, linear regressions, and Poisson and logistic regressions were implemented using *wilcox.test*, *lm* and *glm* functions, respectively.

To assess the relationship between host range and adaptative signals (**Fig. 4**), we used Poisson regressions to model the expected number of host genera observed in each viral clique, $\lambda_{host\ range}$. We corrected for the number of genomes in each clique, *g*, as a measure of sampling effort, and viral family membership, *v*, by including them as fixed effects in these models. These models can be formalised for mutational distance or dN/dS, *d*, with some *p* number of viral families, and residual error, ε, as:

$$ln\left(\lambda_{host\,range}\right) = \beta_0 + \beta_1(\ln(g)) + \sum_{i=1}^{p} \beta_{i+1} v_i + \beta_{p+2}(\ln(d)) + \varepsilon$$

We tested if the parameter estimates were non-zero by performing two-tailed *Z*-tests implemented within the *summary* function in R.

To estimate the strength of adaptive signals for corona-, paramyxo-, rhabdo- and circoviruses (**Fig. 5**) by gene type, we implemented two linear regression models for each viral family. Since the overall adaptive signal may differ for each viral clique, we corrected for this effect using an initial linear model – where the number of viral cliques, viral clique membership, and residual, are given by *q*, *c*, and ε, respectively – as follows:

$$ln(dN/dS) = \beta_0 + \sum_{i=1}^{q} \beta_i c_i + \beta_{p+2}(\ln(d)) + \varepsilon + \varepsilon_{model\,1}$$

Subsequently, we used the corrected log-dN/dS estimates represented by the residuals of model 1, $\varepsilon_{model\ 1}$, in a second linear model partitioning the effects of gene type by host jump status, *j*. Given *r* number of gene types, this model can be formalised as follows:

$$\varepsilon_{model\,1} = \beta_0 + \sum_{i=1}^{r}\sum_{j=1}^{2}\beta_{i,j}c_{i,j} + \varepsilon_{model\,2}$$

The estimated effects shown in **Fig. 5**, representative of the difference in adaptive signals associated to jump and non-host jump lineages for each gene type, was then computed as:

$$Effect = \beta_{r,jump} - \beta_{r,nonjump}$$

To test whether this effect is statistically significant, we used a one tailed *t*-test, with a *t* statistic computed using the standard error of the parameter estimates in model 2:

$$t_r = \frac{Effect}{\sqrt{SE_{\beta_{r,\,jump}}{}^2 + SE_{\beta_{r,\,nonjump}}{}^2}}$$

The residuals of model 2 were confirmed to be approximately normal by visual inspection (**Extended Data Fig. 6**).

Data analysis and visualisation

All data analyses were performed using R v4.3.1. All visualisations were performed using ggplot v3.4.2[58] or *ggtree* v3.8.2[59].

References

1. Jones, K. E. *et al.* Global trends in emerging infectious diseases. *Nature* **451**, 990–993 (2008).
2. Shaw, L. P. *et al.* The phylogenetic range of bacterial and viral pathogens of vertebrates. *Mol. Ecol.* **n/a**, (2020).
3. Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nat. 2017 5467660* **546**, 646–650 (2017).
4. Gibb, R. *et al.* Zoonotic host diversity increases in human-dominated ecosystems. *Nature* **584**, 398–402 (2020).
5. Woolhouse, M. E. J. & Gowtage-Sequeria, S. Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.* **11**, 1842 (2005).
6. Albery, G. F., Eskew, E. A., Ross, N. & Olival, K. J. Predicting the global mammalian viral sharing network using phylogeography. *Nat. Commun.* **11**, 2260 (2020).
7. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. J. Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **356**, 983–989 (2001).
8. Albery, G. F. *et al.* Urban-adapted mammal species have more known pathogens. *Nat. Ecol. Evol.* **6**, 794–801 (2022).
9. Karesh, W. B. *et al.* Ecology of zoonoses: natural and unnatural histories. *The Lancet* **380**, 1936–1945 (2012).
10. Cleaveland, S., Laurenson, M. K. & Taylor, L. H. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **356**, 991–999 (2001).

11. Carlson, C. J. *et al.* The Global Virome in One Network (VIRION): an atlas of vertebrate-virus associations. *MBio* **13**, e02985-21 (2022).
12. Gibb, R. *et al.* Data proliferation, reconciliation, and synthesis in viral ecology. *BioScience* **71**, 1148–1156 (2021).
13. Shaw, L. P. *et al.* The phylogenetic range of bacterial and viral pathogens of vertebrates. *Mol. Ecol.* **29**, 3361–3379 (2020).
14. Kuchipudi, S. V. *et al.* Coordinated surveillance is essential to monitor and mitigate the evolutionary impacts SARS-CoV-2 spillover and circulation in animal hosts. *Nat. Ecol. Evol*.
15. Watsa, M. & Wildlife Disease Surveillance Focus Group. Rigorous wildlife disease surveillance. *Science* **369**, 145–147 (2020).
16. Cedric Tan *et al.* Genomic screening of 16 UK native bat species through conservationist networks uncovers coronaviruses with zoonotic potential. *bioRxiv* 2023.01.17.524183 (2023) doi:10.1101/2023.01.17.524183.
17. Pepin, K. M., Lass, S., Pulliam, J. R. C., Read, A. F. & Lloyd-Smith, J. O. Identifying genetic markers of adaptation for surveillance of viral host jumps. *Nat. Rev. Microbiol.* **8**, 802–813 (2010).
18. Kaur, T. *et al.* Descriptive epidemiology of fatal respiratory outbreaks and detection of a human-related metapneumovirus in wild chimpanzees (Pan troglodytes) at Mahale Mountains National Park, Western Tanzania. *Am. J. Primatol. Off. J. Am. Soc. Primatol.* **70**, 755–765 (2008).

19. Kuchipudi, S. V. *et al.* Coordinated surveillance is essential to monitor and mitigate the evolutionary impacts of SARS-CoV-2 spillover and circulation in animal hosts. *Nat. Ecol. Evol.* (2023) doi:10.1038/s41559-023-02082-0.
20. Simmonds, P. *et al.* Four principles to establish a universal virus taxonomy. *PLoS Biol.* **21**, e3001922 (2023).
21. Adams, M. J. *et al.* 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. *Arch. Virol.* **162**, 1441–1446 (2017).
22. Walker, P. J. *et al.* Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch. Virol.* **167**, 2429–2440 (2022).
23. Acman, M., van Dorp, L., Santini, J. M. & Balloux, F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 2452 (2020).
24. Tan, C. C. S. *et al.* Transmission of SARS-CoV-2 from humans to animals and potential host adaptation. *Nat. Commun.* **13**, 2988 (2022).
25. Kuchipudi, S. V. *et al.* Multiple spillovers from humans and onward transmission of SARS-CoV-2 in white-tailed deer. *Proc. Natl. Acad. Sci.* **119**, e2121644119 (2022).
26. Munnink, B. B. O. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**, 172–177 (2021).
27. McAloose, D. *et al.* From people to Panthera: Natural SARS-CoV-2 infection in tigers and lions at the Bronx Zoo. *MBio* **11**, e02220-20 (2020).
28. Short, K. R. *et al.* One health, multiple challenges: The inter-species transmission of influenza A virus. *One Health* **1**, 1–13 (2015).
29. Nelson, M. I. & Vincent, A. L. Reverse zoonosis of influenza to swine: new perspectives on the human–animal interface. *Trends Microbiol.* **23**, 142–153 (2015).

30. Samara, E. M. & Abdoun, K. A. Concerns about misinterpretation of recent scientific data implicating dromedary camels in epidemiology of Middle East respiratory syndrome (MERS). *MBio* **5**, 10–1128 (2014).
31. Du, L. & Han, G.-Z. Deciphering MERS-CoV evolution in dromedary camels. *Trends Microbiol.* **24**, 87–89 (2016).
32. Zhang, Z., Shen, L. & Gu, X. Evolutionary dynamics of MERS-CoV: potential recombination, positive selection and transmission. *Sci. Rep.* **6**, 25049 (2016).
33. Huang, Y., Yang, C., Xu, X., Xu, W. & Liu, S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).
34. Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* **21**, 162–177 (2023).
35. Messenger, A. M., Barnes, A. N. & Gray, G. C. Reverse zoonotic disease transmission (zooanthroponosis): a systematic review of seldom-documented human biological threats to animals. *PloS One* **9**, e89055 (2014).
36. Edwards, S. J., Chatterjee, H. J. & Santini, J. M. Anthroponosis and risk management: a time for ethical vaccination of wildlife? *Lancet Microbe* **2**, e230–e231 (2021).
37. Schrauwen, E. J. & Fouchier, R. A. Host adaptation and transmission of influenza A viruses in mammals. *Emerg. Microbes Infect.* **3**, 1–10 (2014).
38. Villordo, S. M., Carballeda, J. M., Filomatori, C. V. & Gamarnik, A. V. RNA structure duplications and flavivirus host adaptation. *Trends Microbiol.* **24**, 270–283 (2016).

39. Urbanowicz, R. A. *et al.* Human adaptation of Ebola virus during the West African outbreak. *Cell* **167**, 1079–1087 (2016).
40. Damas, J. *et al.* Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci.* **117**, 22311–22322 (2020).
41. Lam, S. D. *et al.* SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. *Sci. Rep.* **10**, 16471 (2020).
42. Tan, C. C. *et al.* Genomic screening of 16 UK native bat species through conservationist networks uncovers coronaviruses with zoonotic potential. *Nat. Commun.* **14**, 3322 (2023).
43. Starr, T. N. *et al.* ACE2 binding is an ancestral and evolvable trait of sarbecoviruses. *Nature* **603**, 913–918 (2022).
44. Wright, A. M., Lyons, K. M., Brandley, M. C. & Hillis, D. M. Which came first: the lizard or the egg? Robustness in phylogenetic reconstruction of ancestral states. *J. Exp. Zoolog. B Mol. Dev. Evol.* **324**, 504–516 (2015).
45. Liu, P., Song, Y., Colijn, C. & MacPherson, A. The impact of sampling bias on viral phylogeographic reconstruction. *PLoS Glob. Public Health* **2**, e0000577 (2022).
46. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
47. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

48. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci.* **105**, 1118–1123 (2008).

49. Yang, Z., Algesheimer, R. & Tessone, C. J. A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.* **6**, 30750 (2016).

50. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

51. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

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

53. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

54. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).

55. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

56. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).

57. Li, W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**, 96–99 (1993).

58. Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).

59. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

Acknowledgments

The authors would like to thank Rory J. Gibbs, Gemma Murray and Liam P. Shaw for their helpful feedback and discussions. C.C.S.T. is funded by the National Science Scholarship from the Agency for Science, Technology and Research (A*STAR), Singapore. F.B. and L.v.D. are funded by the European Commission (Horizon 2021-2024, END-VOC Project). L.v.D. is also funded by the UCL Excellence Fellowship. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. For the purpose of open access, the corresponding author has applied a 'Creative Commons Attribution' (CC BY) licence to any Author Accepted Manuscript version arising. The authors acknowledge the use of the UCL Myriad High Performance Computing Facility (Myriad@UCL), and associated support services, in the completion of this work.

Author contributions

C.C.S.T performed all analyses. L.v.D. and F.B. jointly supervised the study. C.C.S.T, L.v.D. and F.B wrote the manuscript.

Declaration of competing interest

The authors declare that to current knowledge, there are no legal, financial or personal competing interests.

Data and code availability

All custom code used to perform the analyses reported here are hosted on GitHub (https://github.com/cednotsed/vertebrate_host_jumps). The full list of accessions considered in this study are provided in **Supplementary Table 1**.

Main figures

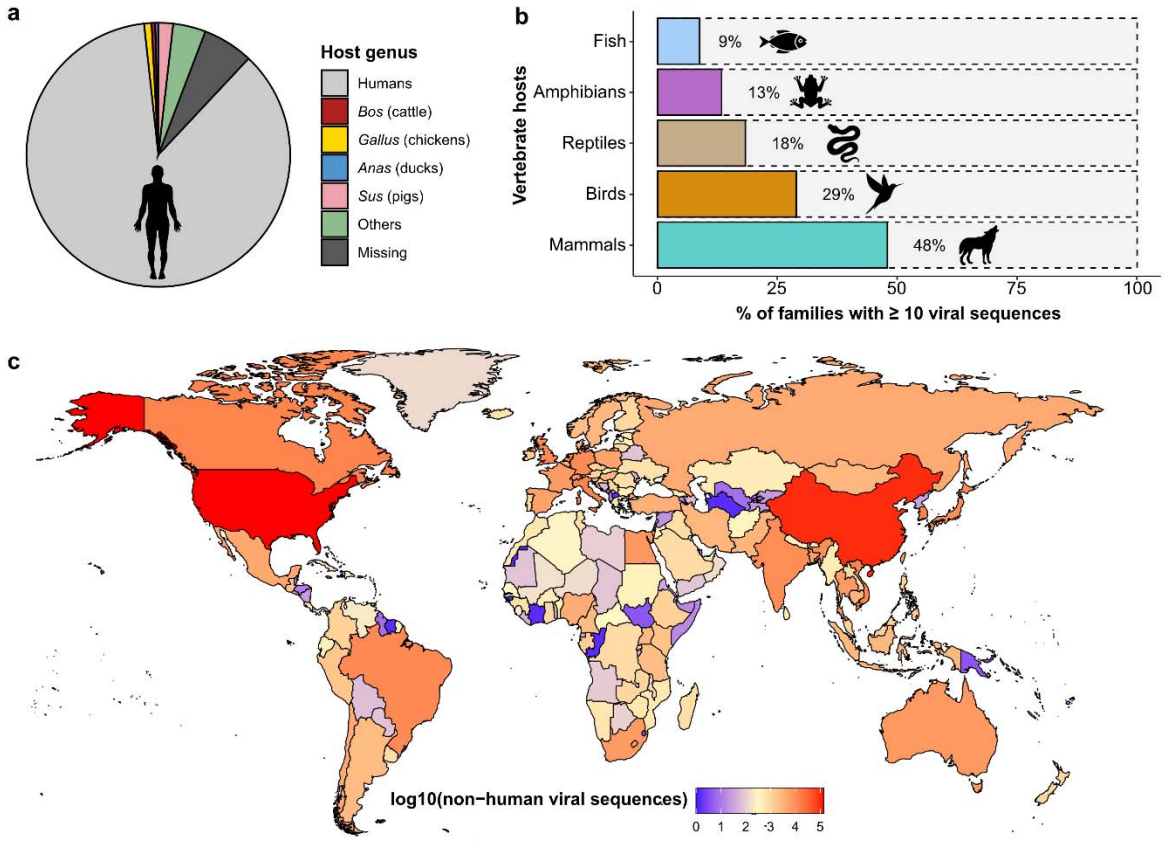

Figure 1. Current state of the global genomic surveillance of vertebrate viruses. *(a) Proportion of all viral sequences deposited on public sequence databases (n=11,645,803) stratified by host. Viral sequences associated with humans and the next four most-sampled hosts are shown. Sequences with no host metadata resolved at the genus level are denoted as 'missing'. (b) Proportion of host families represented by at least 10 associated viral sequences for the five major vertebrate host groups. (c) Global heatmap of sequencing effort, generated from all viral sequences deposited on public sequence databases (n=11,645,803).*

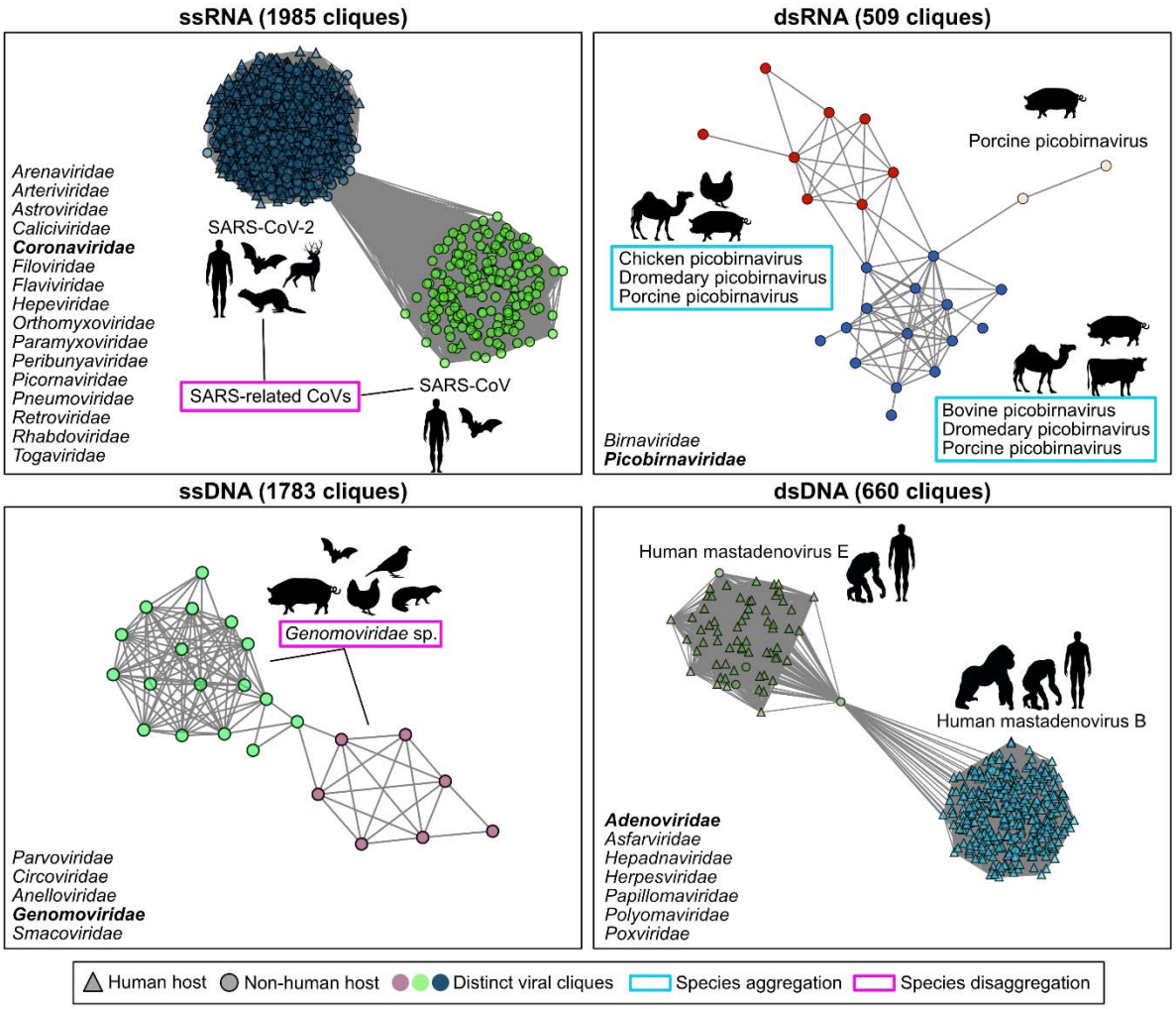

Figure 2. Taxonomy-agnostic approach for identifying equivalent units of viral diversity. *Networks of representative viral cliques identified within the Coronaviridae (ssRNA), Picobirnaviridae (dsRNA), Genomoviridae (ssDNA) and Adenoviridae (dsDNA). Some viral clique assignments aggregated multiple viral species, while others disaggregated species into multiple cliques. Nodes, node shapes, and edges represent individual genomes, their associated host and their pairwise Mash distances, respectively. Edges representing >0.15 Mash distance were removed (see Methods). The list of viral families considered in our analysis are shown on the bottom-left corner of each panel. Silhouettes were sourced from Flaticon.com and Adobe Stock Images (https://stock.adobe.com) with a standard license.*

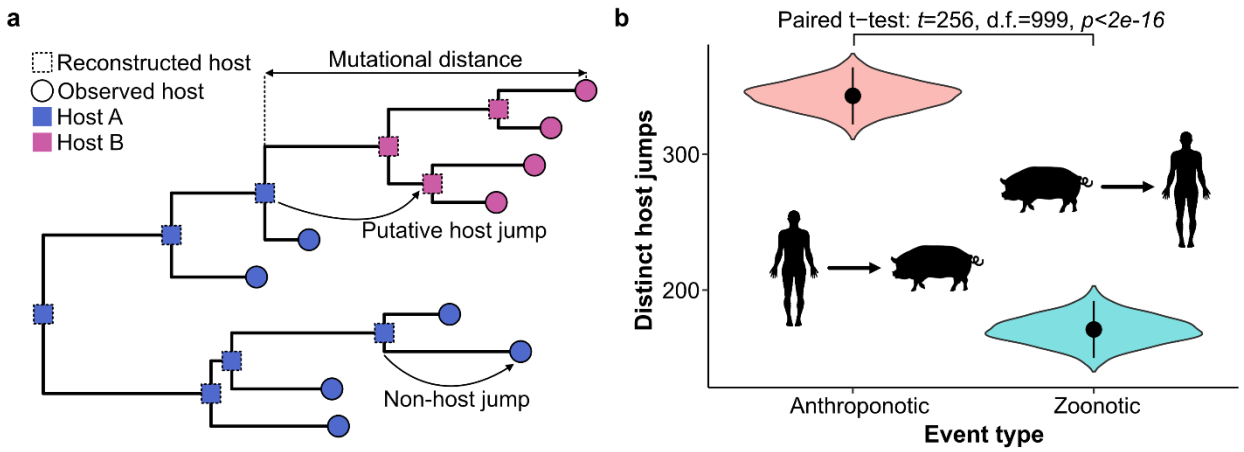

Figure 3. Humans give more viruses to animals than they give to us. *(a) Illustration of ancestral host reconstruction approach used to identify the directionality of putative host jumps. The mutational distance represents the sum of the branch lengths between the tip sequence and the ancestral node for which the first host state transition occurred (in substitutions per site) multiplied by the alignment length. (b) Number of distinct putative host jumps involving humans across all viral families considered (n=30). Black dots represent the observed point estimates for each type of host jump. The violin plots show the bootstrap distributions of these estimates, where the host jumps in each viral clique were resampled with replacement for 1000 iterations. Black lines show the 95% confidence intervals associated with these bootstrap distributions. Silhouettes were sourced from Flaticon.com and Adobe Stock Images (https://stock.adobe.com) with a standard license. A two-tailed paired t-test was performed to test for a difference in the zoonotic and anthroponotic bootstrap distributions.*

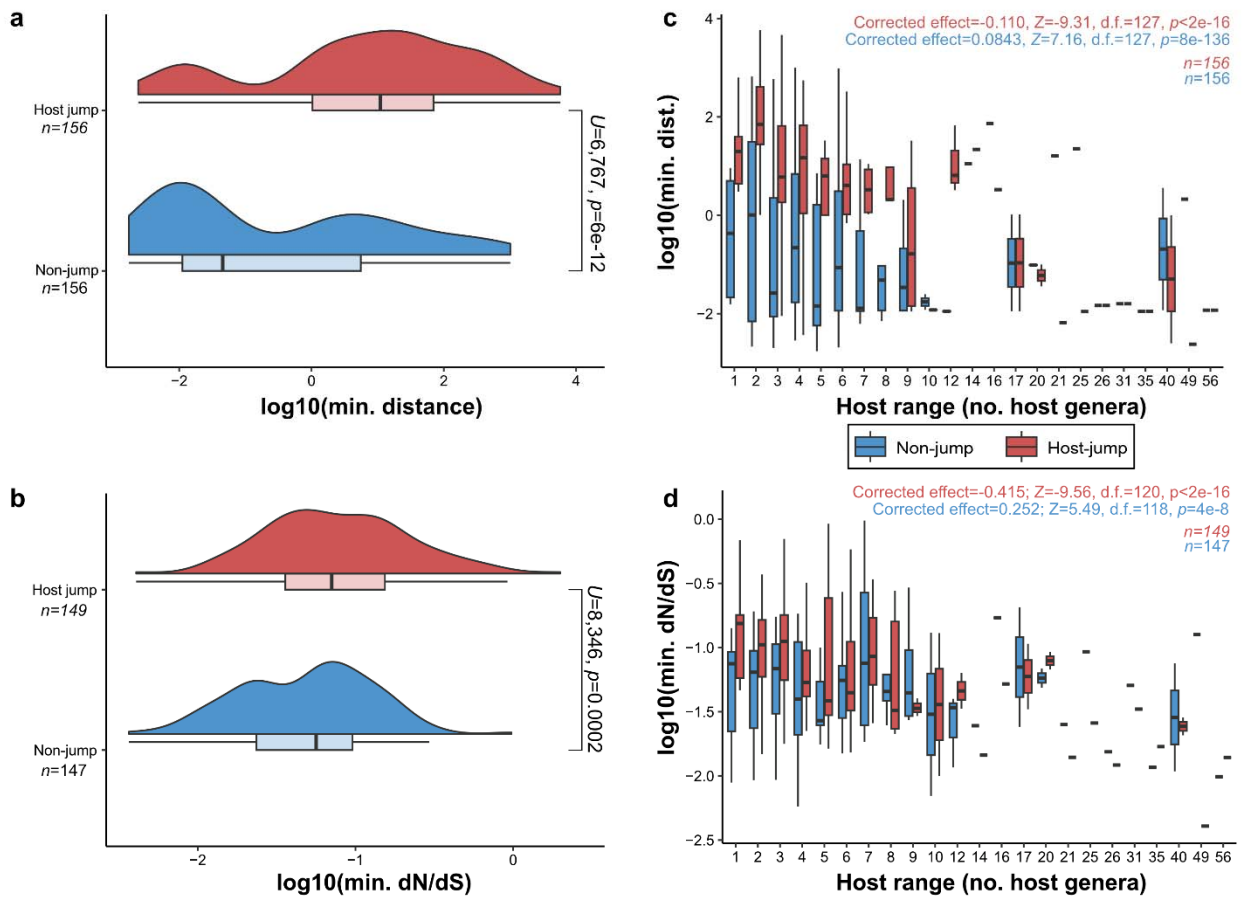

Figure 4. The degree of adaptation required for host jumps is less for multi-host viruses. *Distributions of minimum (a) mutational distance and (b) dN/dS for inferred host jump events and non-host jump controls on the logarithmic scale. Boxplots of (c) mutational distance and (d) dN/dS for host jump and non-host jumps on the natural-logarithmic scale. We corrected for the effects of sequencing effort and viral family membership using Poisson regression models. The estimated effects of patristic distance on host range after these corrections are annotated. We tested whether the estimated effects were non-zero using two-tailed Z-test. For all panels, each data point represents the minimum distance or dN/dS across all host jump or randomly selected non-host jump lineages in a single clique. Boxplot elements are defined as follows: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.*

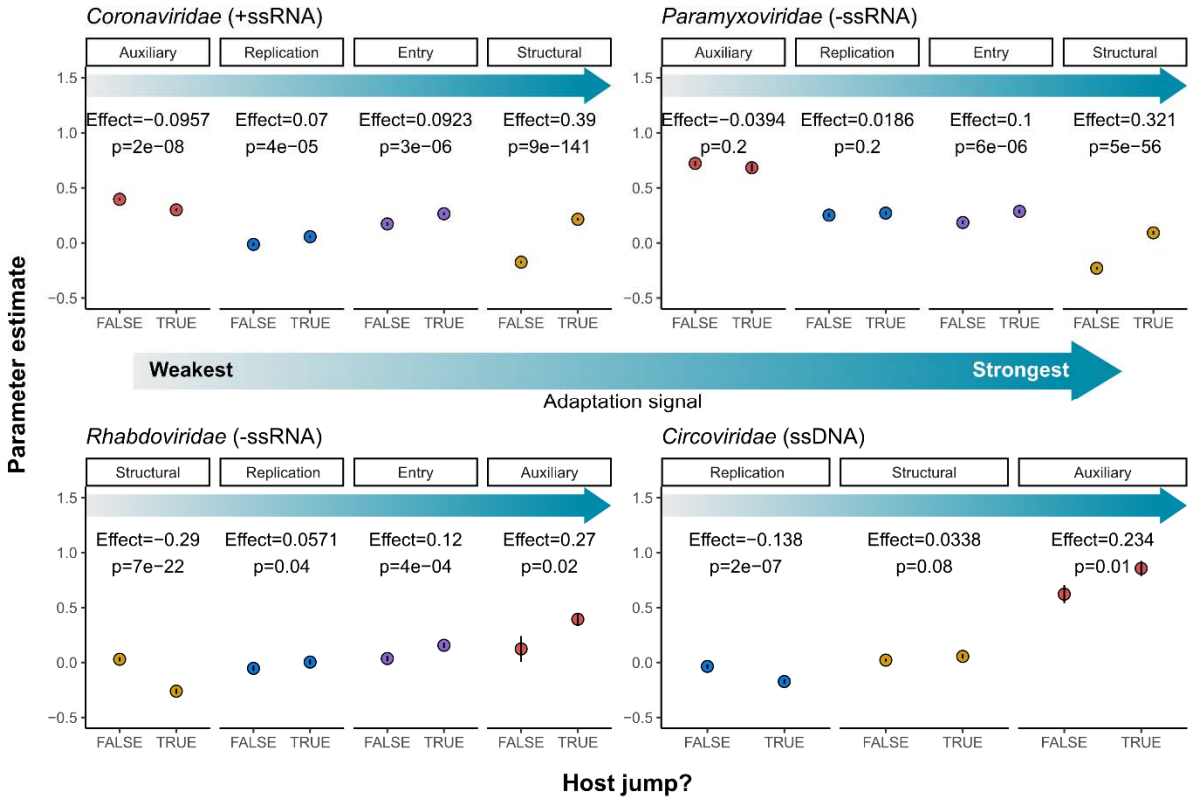

Figure 5. Adaptation signals are gene specific. Strength of adaptation signals per gene type were estimated using linear models. We modelled the effects of gene type and host jump status on log(dN/dS) while correcting for viral clique membership and, for each gene type, inferred the strength of adaptive signal (denoted 'effect') as the difference in parameter estimates for host jumps versus non-host jumps. Points and lines represent the parameter estimates and their standard errors respectively. Differences in parameter estimates were tested against zero using a one-tailed t-test. Sub-panels for each gene type were ordered from left to right with increasing effect estimates.