AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

| Ecology | Full-Length Text

Accumulated metagenomic studies reveal recent migration, whole genome evolution, and undiscovered diversity of orthomyxoviruses

Gytis Dudas,[1] **Joshua Batson**[2]

AUTHOR AFFILIATIONS See affiliation list on p. 12.

ABSTRACT Metagenomic studies have uncovered many novel viruses by looking beyond hosts of public health or economic interest. However, the resulting viral genomes are often incomplete, and analyses largely characterize the distribution of viruses over their dynamics. Here, we integrate accumulated data from metagenomic studies to reveal geographic and evolutionary dynamics in a case study of *Orthomyxoviridae*, the RNA virus family that includes influenza virus. First, we use sequences of the orthomyxovirid Wŭhàn mosquito virus 6 to track the migrations of its host. We then look at orthomyxovirus genome evolution, finding gene gain and loss across members of the family, especially in the surface proteins responsible for cell and host tropism. We find that the surface protein of Wŭhàn mosquito virus 6 exhibits accelerated non-synonymous evolution suggestive of antigenic evolution, i.e., vertebrate infection, and belongs to a wider quaranjavirid group bearing highly diverged surface proteins. Finally, we quantify the progress of orthomyxovirus discovery and forecast that many diverged *Orthomyxoviridae* members remain to be found. We argue that continued metagenomic studies will be fruitful for understanding the dynamics, evolution, ecology of viruses, and their hosts, regardless of whether novel viruses are identified or not, as long as study designs allowing for the resolution of complete viral genomes are employed.

IMPORTANCE The number of known virus species has increased dramatically through metagenomic studies, which search genetic material sampled from a host for non-host genes. Here, we focus on an important viral family that includes influenza viruses, the *Orthomyxoviridae*, with over 100 recently discovered viruses infecting hosts from humans to fish. We find that one virus called Wŭhàn mosquito virus 6, discovered in mosquitoes in China, has spread across the globe very recently. Surface proteins used to enter cells show signs of rapid evolution in Wŭhàn mosquito virus 6 and its relatives which suggests an ability to infect vertebrate animals. We compute the rate at which new orthomyxovirus species discovered add evolutionary history to the tree of life, predict that many viruses remain to be discovered, and discuss what appropriately designed future studies can teach us about how diseases cross between continents and species.

KEYWORDS *Orthomyxoviridae*, Wuhan mosquito virus 6, metagenomics, phylodynamics, segmentation, antigenic drift, natural selection, gp64, PB1, taxonomy

Metagenomic virus discovery efforts to date

Viruses that cause disease in humans and economically important organisms were the first to be isolated and characterized. Recently, cheap DNA sequencing has enabled a wave of metagenomic studies in a broader range of hosts, in which viruses are identified in a host sample by nucleic acid sequence alone and a new virus is said to be discovered if that sequence is sufficiently diverged. As a result, the number of known

Editor Colin R. Parrish, Cornell University Baker Institute for Animal Health, Ithaca, New York, USA

Address correspondence to Joshua Batson, joshua.batson@gmail.com, or Gytis Dudas, gytis.dudas@gmc.vu.lt.

The authors declare no conflict of interest.

See the funding table on p. 12.

Received 17 July 2023
Accepted 29 August 2023
Published 13 October 2023

viruses has increased by more than an order of magnitude in the decade since 2012 (1). While some entirely new viral families have been proposed (2), many of these new viruses are interleaved on the virus tree of life with viruses infecting hosts of economic importance. Studying their ecology (3) and host associations (4, 5) provides insight into the host-switching and genome evolution processes important for the evolution of pathogenicity.

This richer tree of viruses has provided some early success stories, such as jingmenvirids that are currently associated with human disease (6) but which were first discovered metagenomically in ticks (7), indicating a likely route of transmission. Surveillance in hosts known to pose disproportionate risk, such as bats (8), has similarly provided context for zoonotic pathogens like severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (9). Metagenomic studies carried out at scale can effectively multiplex other tasks previously addressed with targeted sampling, like understanding the evolutionary history of human pathogens (10) or using viruses that evolve faster than their hosts to track host movements (11).

Going beyond mere virus discovery

Here, we seek to show how accumulated data from metagenomic studies can provide deep insights into viral evolution and dispersion across a moderately diverged viral clade through a case study of *Orthomyxoviridae*. *Orthomyxoviridae* is a family of enveloped segmented negative-sense single-stranded RNA viruses that infect vertebrates and arthropods. Traditionally, it is split into genera *Thogotovirus*, *Quaranjavirus* (both mostly tick-borne and occasionally infecting vertebrates), *Isavirus* (found only in fish so far), and a genus for each of influenza A, B, C, and D viruses (found in various vertebrates, including humans). Orthomyxovirus discovery has historically been driven by impact on human health (e.g., influenza virus) and livelihood (e.g., salmon infectious anemia virus), or association with known disease vectors (e.g., the tick-borne Johnston Atoll quaranja- and Thogoto thogotovirus). The metagenomic revolution has resulted in 10 times more orthomyxovirids being discovered over the last decade than in the previous 79 years since the first orthomyxovirus discovery, of influenza A virus, in 1933. The vast majority of known orthomyxoviruses use one of two surface protein classes, with vertebrate-infecting-only members (influenza and isavirids) using one or more class I membrane fusion proteins derived from hemagglutinin-esterase-fusion (HEF) (12), sometimes delegating the receptor cleavage function upon exit from the host cell to a separate protein neuraminidase (NA). Meanwhile, arthropod-infecting orthomyxovirids (quaranja- and thogotovirids, which sometimes spill over into vertebrates) use a class III membrane fusion protein called gp64 (13). Orthomyxovirus genomes are known to have six to eight segments, but many metagenomically discovered viruses in this group have incomplete genomes. To our knowledge, an inventory of surface protein class use and segment content of *Orthomyxoviridae* members is not yet available.

We start by showing how closely related virus sequences observed across numerous studies can reveal host spatial dynamics and virus microevolution, using the orthomyxovirid Wǔhàn mosquito virus 6 (WuMV-6). WuMV-6 was discovered in China in 2013 from a single polymerase basic 1 (PB1) sequence (4). WuMV-6 belongs to a diverse clade of arthropod orthomyxovirids somewhat distantly related to members of the genus *Quaranjavirus*. WuMV-6 has since turned up in *Culex* mosquitoes across numerous metagenomic studies all around the world (14–16). The abundance of WuMV-6 genome data and its amenability to molecular clock analyses are the main reasons why we chose to focus on it. To our knowledge, WuMV-6 has not been isolated. Looking beyond WuMV-6, we map out known genome composition across members of *Orthomyxoviridae*, highlighting parts of the tree where changes to segment numbers are likely to have taken place. In this task, we focus on the PB1 protein of the heterotrimeric orthomyxovirid RNA-directed RNA polymerase (RdRp) complex for most of our analyses, since PB1 encodes the RdRp motif responsible for replicase activity (the two other proteins of the RdRp heterotrimer are PB2 and polymerase acidic (PA)). In looking at genome

composition, we pay close attention to surface protein use across the PB1 tree, and focus particularly on gp64 proteins used by thogoto- and quaranjaviruses. We find surface proteins to be quite mobile across members of *Orthomyxoviridae* over evolutionary timescales and identify a clade of quaranjaviruses known to have acquired new segments using distinctly diverged gp64 proteins. Finally, we borrow methods from macroevolutionary research to quantitatively assess the pace at which orthomyxovirus evolutionary history is being uncovered, finding that despite their already transformative effect, metagenomic discovery efforts are likely to continue to find substantially diverged members of *Orthomyxoviridae* for some time.

RESULTS

WuMV-6 exhibits rapid population dynamics

WuMV-6, a mosquito orthomyxovirus seen frequently across much of the world (4, 14, 15), belongs to a clade related to the genus *Quaranjavirus* and has two extra segments compared to other quaranjavirids (16). WuMV-6 is also distinct from many other arthropod RNA viruses in being found very often yet exhibiting limited genetic diversity and a strong molecular clock signal (Fig. S1 to S3), allowing the use of phylogenetic methods like the reassortment network (17). Figure 1A (also Fig. S4) shows a reassortment network analysis of currently available complete WuMV-6 genomes (seven from Australia, 13 from California, three from Cambodia, one from China, three from Sweden) collected between 2006 and 2020, depicting relationships between segments, their reassortments with respect to each other, and timings of both. We find that all WuMV-6 segments share a common ancestor within the last 60-odd years, which is not unusual for insect viruses (18) (Fig. 1A; Fig. S5 and S6), and that a more recent, potentially global, sweep is underway, with segments from six continents sharing a common ancestor in the last 20 years (Fig. 1B).

Although the geographic population structure of WuMV-6 is appreciable, with samples from the same country often close on the tree, reassortment events indicate contact between genomic lineages across vast distances. For example, reassortment events #2 and #3 in Fig. 1A indicate contact as recently as 2010–2015 between WuMV-6 lineages eventually found in Australia and California. Similarly, some lineages found in China are related to recent (*circa* 2017) Californian lineages (reassortment event #1 in Fig. 1A). Even lineages not represented in the reassortment network due to incomplete genomes show evidence of gene flow, like Chinese and Greek PB1 sequences in Fig. 1B. These results indicate that WuMV-6 populations are very mobile.

Surface protein gp64 of WuMV-6 and its relatives evolves rapidly

By estimating the rates of synonymous and non-synonymous evolution, we find that gp64, the surface protein of WuMV-6, is evolving faster in terms of non-synonymous substitutions per codon per year than the rest of the known WuMV-6 proteome, save for the smallest segment, which is expected to be spliced [hypothetical 3 (16)] and therefore likely to contain overlapping reading frames (Fig. 2A; Fig. S7). The highest dN/dS values in gp64 appear to be concentrated around its fusion loops (13) (Fig. S8). We see elevated rates of amino acid evolution in gp64 across the wider clade defined by Astopletus and Ūsinis viruses, to which WuMV-6 belongs. Members have PB1 proteins (encoding RdRp) closely related (Fig. 2B) but gp64 proteins substantially diverged from other quaranjaviruses and each other (Fig. 2C; Fig. S9). The pronounced non-synonymous divergence in gp64 at the WuMV-6 population level and the wider Asto-Ūsinis clade level indicates some evolutionary pressure on this surface protein, such as diversifying selection pressure from repeat infections of hosts with humoral immune systems.

Comparing the numbers of PB1 and gp64 protein sequences discovered so far indicates a clear paucity of the latter, especially in the Asto-Ūsinis clade, which highlights the poor general state of knowledge of genome composition across members of *Orthomyxoviridae*. The closest relative with reliable segment information [based on

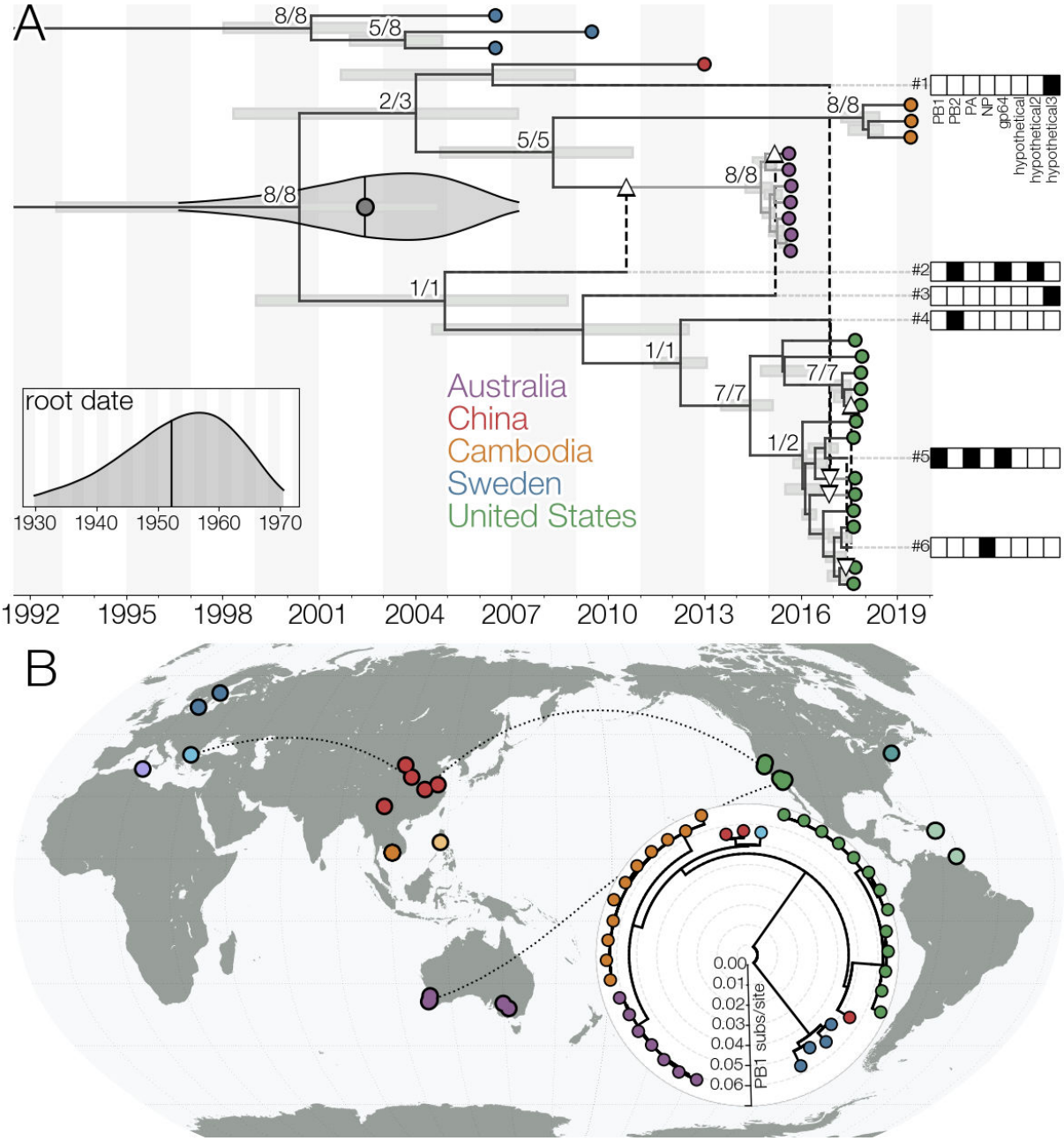

FIG 1 (A) Reassortment network of complete WuMV-6 genomes. Tips are indicated with circles and colored based on location. Reassortant edges are indicated with dashed lines terminating in white arrowheads and numbered with segments carried along the edge indicated with filled-in rectangles to the right of the plot (for clearer segment embeddings, see Fig. S4). The network is truncated to 1995 with a violin plot indicating the 95% highest posterior density (HPD) interval for the date of the common ancestor of non-Swedish WuMV-6 samples estimated from the posterior distribution of reassortment networks. Horizontal gray bars indicate node height 95% HPD intervals in the summary network. Black vertical line with the gray dot within the violin plot indicates the mean estimate. The inset plot indicates same for the root date of the network with black vertical line indicating the mean. Since the summary procedure for the posterior distribution of networks is overly conservative (see Fig. S5), node supports are expressed as number of times a given node is seen with ≥0.95 probability in segment embedding summary trees, after carrying out the subtree prune-regraft procedures for any given embedding indicated by reassortant edges, out of all such nodes. (B) A maximum likelihood (ML) tree of WuMV-6 PB1 sequences, showing additional samples for which only incomplete genomes are available (Greece and China). Dots on the map indicate all locations where the presence of WuMV-6 has been detected, including complete (Australia, Cambodia, China, Sweden, USA) and incomplete genomes (China and Greece), and detections at the read level (Connecticut, Philippines, Puerto Rico, Trinidad and Tobago, and Tunisia). Dotted lines connect locations that have experienced recent WuMV-6 gene flow based on reassortment patterns. For all available WuMV-6 segment data, see Fig. S6.

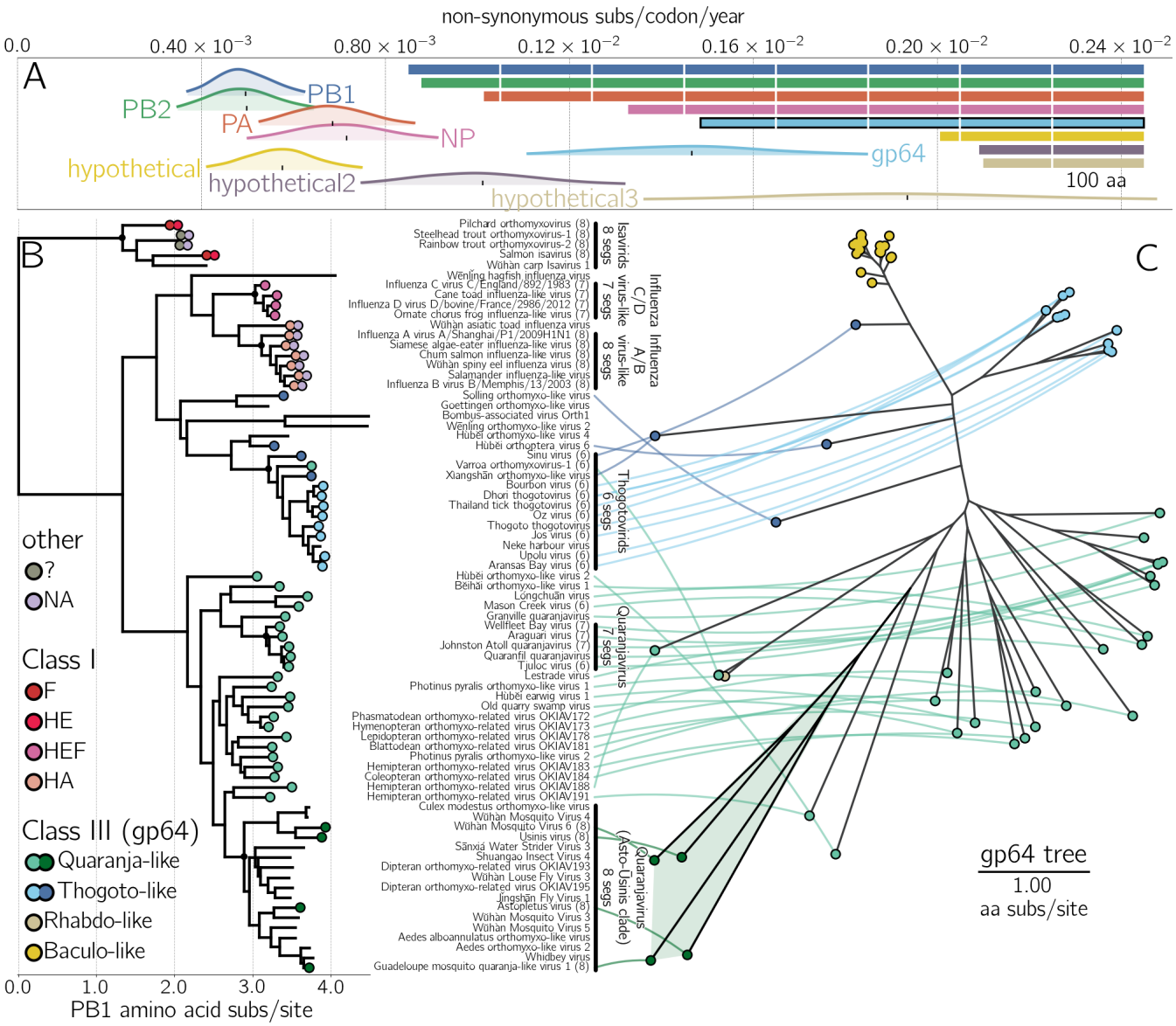

FIG 2 (A) 95% highest posterior densities for the rate of non-synonymous mutations per codon per year for each known WuMV-6 gene (indicated by color). Black vertical ticks indicate the mean estimate. Lengths of open reading frames are shown to the right, with white lines denoting 100 amino acid increments and gp64 outlined in black. Note that putative hypothetical 3 protein is expected to be spliced (16) and may contain overlapping coding regions where synonymous changes in one frame may be non-synonymous in another. (B) Rooted phylogenetic tree depicts the relationships between PB1 proteins of orthomyxoviruses. Surface proteins are marked as colored circles at the tips: red hues for class I membrane fusion proteins, green/blue and yellow/brown for class III proteins (gp64) of *Orthomyxoviridae* and non-*Orthomyxoviridae*, respectively, and lilac for neuraminidases. Likely genome composition for certain groups are highlighted with black vertical lines to the right of the tree, with corresponding implied common ancestor with such organization marked with a black circle in the tree. (C) An unrooted phylogeny to the right of the PB1 tree shows the relationships between gp64 proteins found in thogoto- or thogoto-like- (blue), quaranja- or quaranja-like- (green), baculo- (yellow), and some rhabdovirids (tan). Where available, each orthomyxovirid gp64 protein is connected to its corresponding PB1 sequence. Black branches with a faded green area in the gp64 tree to the right indicate the position of Asto-Ūsinis gp64 proteins.

electron micrographs (EM) (19)] is the genus *Quaranjavirus* with seven segments (Fig. 2B), indicating that since the common ancestor of the *Quaranjavirus* genus and the Asto-Ūsinis clade, segments were either lost in the former or gained in the latter. Such gaps (Fig. 2B) in understanding seem to increase with phylogenetic distance from vertebrate-pathogenic viruses, a hallmark of retrospective research into outbreaks rather than

prospective efforts aimed at understanding the correlates of pathogenicity across members of this family.

Surface proteins of *Orthomyxoviridae* are prone to horizontal gene transfer

The phylogenetic tree of PB1 with surface protein classes indicated also demonstrates the plasticity in orthomyxovirus genome composition—regardless of rooting, the PB1 tree requires at least two switches in viral membrane fusion protein class to explain the current distribution of HEF-like (class I) and gp64-like (class III) proteins. Even within gp64-using orthomyxoviruses, changes between different gp64 lineages are apparent, e.g., Hùběi orthomyxo-like virus 2 carries gp64 related to the Asto-Ūsinis clade yet does not belong to it in PB1 (Fig. 2B; Fig. S9). Almost all orthomyxoviruses use either HEF-like or gp64-like proteins, with Rainbow/Steelhead trout orthomyxoviruses [and one additional relative (20), not shown] being the only exceptions. Both clearly possess an influenza A/B virus-like NA but the protein termed "hemagglutinin" (20) does not resemble any known protein (21). While *Orthomyxoviridae* are a moderately sized virus family, their members make use of a diverse and evolving set of surface proteins.

Rate of *Orthomyxoviridae* PB1 diversity discovery remains high

We now analyze the progress made by virus discovery studies on *Orthomyxoviridae* members by quantifying the evolutionary contribution (branch length in amino acid substitutions per site) contributed by each new orthomyxovirid taxon. There are two clear phases of discovery: before 2015, public health investigations of pathogens infecting humans and farmed animals or vectored by ticks led to the discovery of 14 viruses. Since 2015, when lower costs of sequencing enabled large metagenomic surveys in arthopods and vertebrates of little immediate economic value (4, 5), 115 additional viruses have been discovered (Fig. S10A) with current trends, when focusing on smaller clades, being consistent with arthropod discovery efforts being the most fruitful (Fig. S10B).

To quantify how each discovery contributed to our knowledge of the family's evolutionary history, we take a phylogenetic approach, building a maximum-likelihood tree of the sole protein shared by all RNA viruses, RdRp (22), encoded here in the PB1 gene (23) (Fig. 3A). We scan through the tree based on the chronology of discovery, attributing to each taxon the sum of the lengths of the branches ancestral to that taxon but not to earlier taxa. This quantity is called the phylogenetic diversity (PD), a metric commonly used in ecology and macroevolution (24), and represents the amount of independent evolution (25) contributed by a sequence to a tree.

We find that distinctive viruses, those contributing significant PD, have continued to be discovered each year (Fig. 3B). For example, the Wēnlǐng orthomyxo-like virus 2 found in 2018 is nearly as distinctive relative to the viruses discovered before it as the infectious salmon anemia virus found in 1984 was. There is no correlation between the year of discovery and the maximum PD contributed by an orthomyxovirid (Spearman's $r = 0.02$, P-value = 0.95). In contrast, the average PD contributed per virus does decrease with time (Spearman's $r = -0.17$, one-tailed P-value = 0.054/2 = 0.027), as shared evolutionary history is attributed to earlier discoveries. While the orthomyxoviruses discovered each year are, on average, less distinctive, the increased host breadth and rapid pace of current studies result in evolutionarily highly distinctive viruses.

Projecting future PB1 phylogenetic diversity discovery

Figure 3C shows the cumulative PD of PB1 after each new orthomyxovirid discovery. Early orthomyxovirid discovery efforts do show some bias, finding viruses more related to one another than by chance: until 2018, the empirical accumulation of PD (black dots) is mostly below the 95 percentile envelope of 10,000 random permutations of discovery order (gray hatched area). We fit the empirical data with a logarithmic function [$f(x) = A \times log_2(1 + x/B)$, where $A = 46.3$ and $B = 62.4$], indicated with a red line in

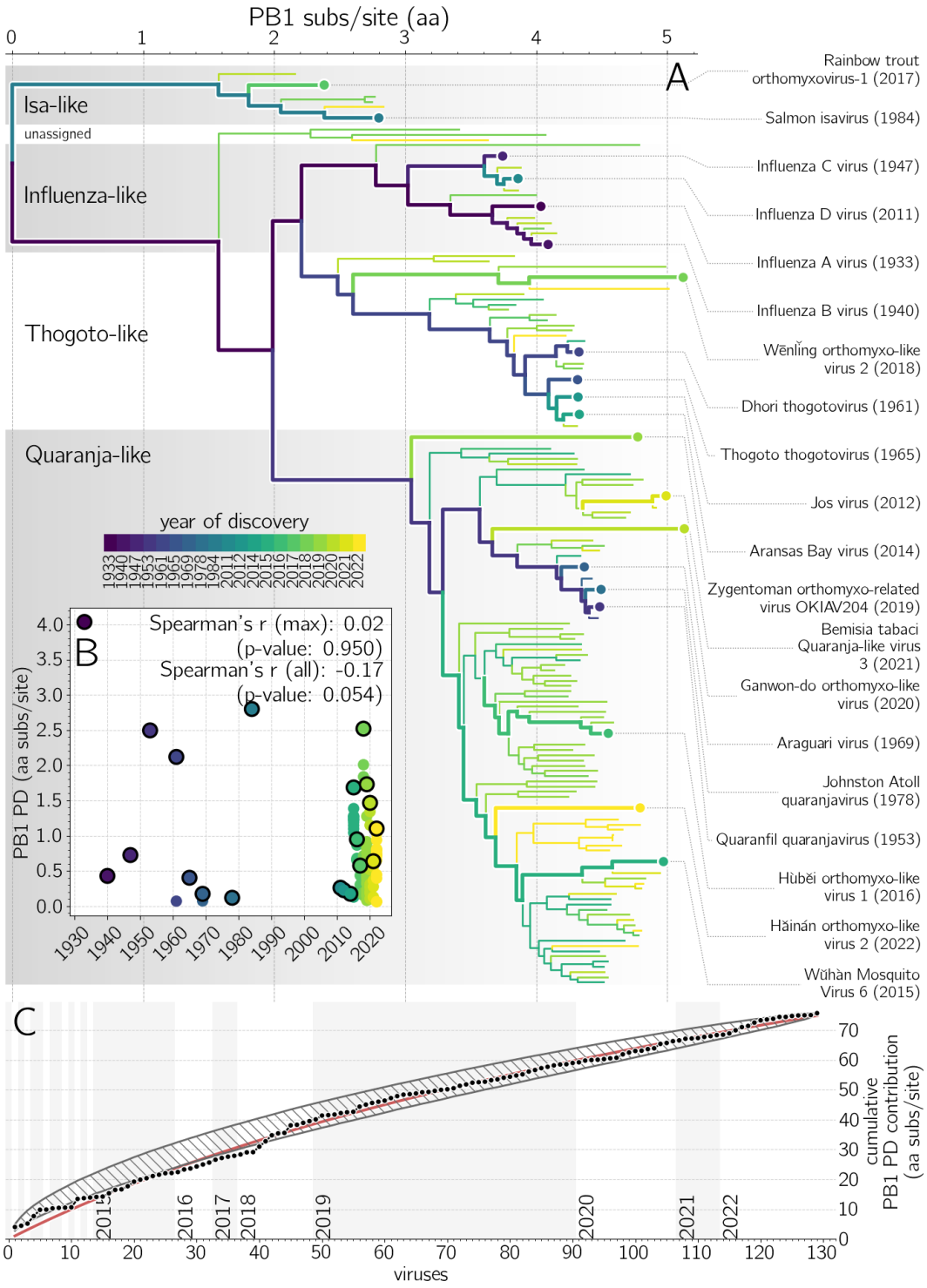

FIG 3 Discovery of orthomyxovirid PB1 phylogenetic diversity (PD). (A) Maximum likelihood (ML) tree of orthomyxovirid PB1 proteins. Branch color indicates earliest discovery year of the lineage. Evolutionary history in purple branches was discovered earlier than yellow branches. Viruses contributing the most PD in their year of discovery are labeled on the right, and indicated with thicker paths in the phylogeny. (B) Inset scatter plot shows PD contributions of each virus versus its year of discovery, with black outlines indicating the maximal PD contributor each year. While the average PD contribution of a newly discovered orthomyxovirus is decreasing with time (Spearman's $r = -0.17$, one-tailed P-value: 0.027), the PD contribution of the most novel virus discovered each year has held steady (Spearman's $r = 0.02$, one-tailed P-value: 0.950/2 = 0.425). (C) Cumulative PB1 PD contribution from successive orthomyxovirid discoveries (black dots) with logarithmic least-squares fit (red line). Gray hatched area indicates the 95 percentile range of cumulative PD contributions under 10,000 random permutations of taxa discovery order.

Fig. 3C. We can extrapolate this curve into the future, e.g., 200th orthomyxovirid is expected to contribute ≈0.26 amino acid substitutions per site to PB1 PD (bringing total PD to 95.9), and the 500th ≈0.12 (total PD 146.8). Note that any remaining bias in the current viral discovery paradigm, leaving some parts of the *Orthomyxoviridae* tree of life undersampled, would manifest in a future PD curve higher than the extrapolation of the one observed so far. There is, regardless, an eventual limit, in which the PD gain of a new discovery is less than the threshold of difference used to define a new taxon (e.g., 0.1 aa sub/site would currently be reached around the 600th taxon). If current trends continue, there would remain at least hundreds of additional taxa to be discovered.

DISCUSSION

Summary

In this work, we endeavored to show how the accumulation of metagenomic data can lead to a new stage in viral studies. We focused on the family *Orthomyxoviridae*, synthesizing data across numerous studies to analyze geographic, evolutionary, and diversity discovery trends.

We first focus on a single recently discovered mosquito RNA virus—Wǔhàn mosquito virus 6 (4)—whose frequency and fast evolution uniquely enable the tracking of mosquito populations. In our experience, metagenomically discovered RNA viruses can be rare or, when encountered often, do not always contain sufficient signal to calibrate molecular clocks (18). WuMV-6 has rapidly disseminated across vast distances, and while anthropogenic (shipping, air travel) (26–28) or abiotic (windborne migration) (29) mechanisms may contribute, the virus' extremely diverged and actively diversifying gp64 surface proteins suggest a potential vertebrate host. Indeed, the rapid sweep of the USA by West Nile virus was accelerated by the movement of both its mosquito vector and its diverse avian hosts (30). While alternation between vertebrate host species could theoretically produce diversifying selection on the WuMV-6 surface protein, gp64 uses Niemann Pick disease type C1 (NPC1) (31), a highly conserved metazoan protein, as its receptor. We thus believe it more likely that WuMV-6 gp64 diversity is selected for by repeat exposure to vertebrate hosts (32), which help disperse WuMV-6 (33) and reduce its effective population sizes (34). Previously, this sort of phylodynamic analysis was limited to known human and animal pathogens (11, 35). As metagenomic discovery efforts continue, more systems like WuMV-6 will undoubtedly be found, contributing to research areas outside of virus evolution, like disease vector dispersal and shifting host distributions under climate change.

Molecular clock analyses applied to arthropod RNA viruses

Arthropod RNA virus population dynamics in the wild are not well understood at this time and, as such, the application of sophisticated phylogenetic analyses to these systems are only getting started. For example, *Drosophila* research, to date, has uncovered the full spectrum of dynamics—a suspected sweep of a Sigma virus in British *Drosophila obscura* over roughly a decade (36), the P element invasion of *Drosophila melanogaster* that took place over a few decades (37), and viruses of *D. melanogaster* and *Drosophila simulans* whose common ancestry is on the order of hundreds of years (18). Similarly, vertical transmission [described in at least one member of the eight-segmented Asto-Ūsinis clade (38)] and diapausing could also complicate molecular clock analyses of arthropod RNA viruses by affecting viral population dynamics and/or replication rates. As such, the identification of more arthropod RNA virus systems that are amenable to molecular clock analyses will be hugely important in illuminating the timescales of RNA virus population turnover in arthropods, identifying departures from neutral evolution, and providing much-needed context.

Long-term evolution of viral receptor binding and membrane fusion proteins

Zooming out, we find a highly modular and fluid genomic organization across members of the family *Orthomyxoviridae*. This presents an interesting conundrum—how and why are novel segments acquired so readily by orthomyxovirids, given that packaging signals encompass multiple sites (39) and obtaining segments via recombination is hard (40)? [Splitting genomes via segmentation is better documented (7, 41) and easier to explain conceptually (42).] Observed frequent switches in surface proteins may be selected for because of their importance in determining host and tissue tropism. We may also be missing additional classes of surface protein, beyond HEF-like and gp64, because of a reliance on sequence homology for protein identification. Indeed, many pooled studies produce incomplete viral genomes, with too few segments relative to their clade, so there is a possibility that significant undetected gain and loss of segments have occurred even within already-discovered viruses. To assess such evolutionary questions will require metagenomic studies to look beyond discovering conserved genes in increasing numbers of host species to completing viral genomes by sequencing individuals across geographic transects (16). Laboratory studies will be necessary to identify the functions of these novel segments and to confirm/determine tropism of discovered surface proteins (43).

Predicting discovery of *Orthomyxoviridae* diversity

Finally, we assess the overall progress of orthomyxovirus discovery from the perspective of PD. We find that the many new orthomyxovirids being discovered every year are adding significant evolutionary history to their family tree. We may contrast this to the situation for birds, which have been studied and characterized for centuries and for which the discovery of new and distinctive species is now rare (24). Where the PD contribution of the most distinctive avian species discovered in each year exhibits a strong downward trend [Fig. S4 from reference (24)], the PD of the most distinctive orthomyxovirus discovered each year remains high (our Fig. 3B). The aggregate trend also indicates that significant PD remains to be discovered: if logarithmic trend continues, known *Orthomyxoviridae* member diversity would double on the discovery of the 531st member. [While there is a risk that some metagenomic sequence represents endogenized viral genes, this is extremely rare for the RdRp gene we use to calculate PD (44).] This complements the argument made by some groups (12) for the existence of many more influenza-like viruses based on virus-host codivergence and the existence of many unsampled host species. We believe that phylogenetic diversity measures, already in widespread use in ecology and macroevolution, will prove useful to the metagenomic virus discovery community as it seeks to assess ongoing progress and predict future payoff. One roadblock to the systematic application of these methods is the lack of consensus on how to define lineages as belonging to the same or different taxa. The International Committee for the Taxonomy of Viruses has created a number of higher taxonomic degrees of organization [some of which are in dispute (45)], while the growth of sequence databases has left many new viruses without an official taxonomic designation. Leadership in curating sequence databases can therefore have a disproportionately bigger impact on the direction the field of long-term virus evolution will take.

Keep going

This work was made possible by the public sharing of annotated genomes, raw sequencing data, and sampling metadata from groups across the world. As metagenomic surveys expand across diverse hosts and geographies, the accumulation of sequence data allows a depth of analysis that moves beyond virus discovery and into ecological and evolutionary dynamics; encountering new samples of previously seen viruses, instead of being seen as a disappointment, can be viewed as opportunity for more granular phylodynamic analysis. The evolutionary interdependence of sequence within and between organisms generates increasing returns on additional surveys. With

appropriate study designs, good data organization, and public sharing strategies, the community's search into the shape of the "virosphere" will offer large dividends for many fields of research.

MATERIALS AND METHODS

Use of viruses for host tracking

Most of WuMV-6 virus genome data (Chinese, Californian, and Australian genomes) were derived from a previous publication (16). Assembled contigs from the Swedish study (15) were provided by John Pettersson while the Cambodian sequences were kindly provided by Jessica Manning, Jennifer Bohl, Dara Kong, and Sreyngim Lay, where WuMV-6 segments described later (16) were identified by similarity.

Puerto Rican segments were recovered by mapping reads from Sequence Read Archive (SRA) entries SRR3168916, SRR3168920, SRR3168922, and SRR3168925 (46) to segments of Californian strain CMS001_038_Ra_S22 using bwa v0.7.17 (47) but most segments except for nucleoprotein (NP) did not have good coverage to be assembled with certainty. Greek segments were recovered by mapping reads from SRA entry SRR13450231 (48) using the same approach as described earlier. New Chinese segments from 2018 (49) were similarly recovered by mapping reads from China National GeneBank Sequence Archive accessions CNS0267022 and CNS0267023 using the same approach as described earlier. New Chinese and Greek segments tended to have acceptable coverage except for segments hypothetical 2 and hypothetical 3 where only individual reads could be detected. The presence of WuMV-6 in more locations around the world was determined based on the presence of reads with ≥90% identity to WuMV-6 PB1 amino acid sequence AJG39094 in Serratus (50). In this way, the presence of WuMV-6 was detected in Connecticut, Trinidad and Tobago, Tunisia, and Philippines.

All successfully assembled segments (from China, Australia, Cambodia, California, Greece, and Sweden) were aligned using multiple alignment using fast Fourier transform (MAFFT) (51) and trimmed to the coding regions of each segment. PhyML v.3.3.2 was used to generate maximum likelihood phylogenies of each segment under an HKY + Γ4 (52, 53) model. Each tree was rooted via least squares regression of tip dates against divergence from root in TreeTime (54).

To confirm sufficient molecular clock signal in WuMV-6, maximum likelihood phylogenies (55) were inferred from all available aligned segments under a GTR + Γ model and rooted via least-squares regression using TreeTime (54) and visualized using root-to-tip plots for each segment. The genome-wide root-to-tip plot was made by only focusing on strains for which we had complete genomes and summing the divergence of each sequence from the root of their segment tree for any given strain.

We also ran molecular clock analyses on all available segment sequences individually using BEAST v.1.10.4 (56) with tip date calibration, GTR + Γ4 nucleotide substitution model, strict molecular clock, and a constant population size tree prior. The date of collection for strain QN-3 (the first WuMV-6 strain to be discovered) was sampled uniformly from the interval between years 2013 and 2014. Analyses were run for 50 million states, sampling every 5,000 states. With convergence confirmed visually in Tracer v.1.7.1 (57) and 10% of the states discarded as burn-in before, the posterior distribution of trees was summarized using TreeAnnotator with the common ancestors option (58). To confirm the diverged WuMV-6 sequences from Sweden were not influencing the molecular clock rate, we repeated the entire analysis but excluded the samples from Sweden.

Twenty-seven complete WuMV-6 genomes (13 from California, 7 from Australia, 3 from Cambodia, 3 from Sweden, and 1 from China) were analyzed using the reassortment network method (17) implemented in BEAST v2.6 (59). For the smallest segment coding for the hypothetical 3 protein, two Ns were inserted after the 349th nucleotide from the initiation codon ATG to account for the presence of a suspected splicing site (16) that brings a substantial portion of this segment back to being coding. Each

segment was partitioned into codon positions 1 + 2 and 3 evolving under independent HKY + Γ4 (52) models of nucleotide substitution and independent strict molecular clocks calibrated by using tip dates. By default, a constant effective population size coalescent tree prior is applied to the reassortment network. Default priors were left in all cases except for effective population size (set to exponential distribution with mean at 100 years) and reassortment rate (set to exponential distribution with mean 0.001 events/branch/year) to get conservative estimates and prevent exploration of complicated parameter space. Markov chain Monte Carlo (MCMC) was run for 200 million states, sampling every 20,000 states in triplicate, after which, all chains were combined after discarding 10% of the states as burn-in and confirmed to have reached stationarity using Tracer (57). The reassortment network was summarized using the native BEAST v2.6 tool (ReassortmentNetworkSummarizer) provided with the package. Posterior embeddings of each segment within the network (in the form of clonal phylogenetic trees) were summarized using TreeAnnotator v.1.10.4 after combining independent runs after discarding 10% of the states as burn-in.

In our personal experience ReassortmentNetworkSummarizer is overly conservative when summarizing reassortment networks due to reassortant edges requiring conditioning on both the origin and destination clades. As such, we removed reassortant edges with ≤0.1 posterior support and summarized posterior supports by first extracting the embedding of each segment from the summarized network by carrying out the subtree prune-and-regraft procedures implied by reassortant edges and then finding how many of the same clades are found in posterior summaries of segment embeddings and how many of those are supported with posterior probability ≥0.95 to produce Fig. 1. Similarly, the 95% highest posterior density interval for most recent common ancestor of the "Pacific rim" clade (i.e., genomes collected outside of Sweden) produced by ReassortmentNetworkSummarizer in Fig. 1 do not overlap perfectly with estimates we extracted from the posterior distribution of reassortment networks though.

All trees were visualized using baltic (https://github.com/evogytis/baltic) and matplotlib (60); all maps were produced using cartopy v0.21.1 which sources vector map data from Natural Earth (public domain) and Global Self-consistent, Hierarchical, High-resolution Geography Database released under the GNU Lesser General Public License.

Orthomyxovirus segmentation and surface proteins

For each clonal WuMV-6 segment embedding within the reassortment network, 1,000 trees from the posterior distribution were extracted after removing 10% burn-in and combining all three independent runs. These trees were then used as empirical trees to be sampled from in a BEAST v.1.10.4 (56) renaissance counting analysis (61) run for 10 million states, sampling every 1,000 states.

Orthomyxovirid PB1 protein sequences from each genus—isa-, influenza, thogoto-, and quaranjaviruses—were used as queries in a protein BLAST (62) search with influenza A, B, C, and D viruses excluded from the search. Having identified the breadth of PB1 protein diversity and having downloaded representative PB1 proteins of influenza A, B, C, and D viruses, we aligned all sequences using MAFFT (51) (E-INS-i mode) and removed sequences that were identical or nearly identical, as well as short or poorly aligning sequences. We repeated this procedure with blast hits to capture as much PB1 diversity as is publicly available. Partial, poorly aligning, or insufficiently distinct PB1 sequences were removed from the analysis.

We used the same data gathering technique for surface proteins. To identify HEF-like proteins we used isavirid HE and influenza C and D virus HEF proteins as queries but did not identify any additional proteins. The claimed hemagglutinin proteins of Rainbow and Steelhead trout isaviruses did not resemble anything on GenBank except each other and did not produce any significant hits via HHpred (21). BLAST searches using orthomyxovirid gp64 relatives identified thogoto- and quaranjavirus surface proteins, as well as baculoviruses and rhabdoviruses with identifiably related proteins. The presumed

gp64 proteins found within the clade encompassed by Ūsinis, Astopletus (discovered in California), and WuMV-6 with Guadeloupe mosquito quaranja-like virus 1 (previously described), referred to here as the Asto-Ūsinis clade within quaranjavirids, did not resemble anything on GenBank via protein BLAST but were all inferred to strongly resemble gp64 proteins via HHpred and as such were aligned using MAFFT in G-INS-i mode.

The PB1 data set was then reduced to viruses for which gp64 sequences were largely available, members of the Asto-Ūsinis clade, and more diverged members. Phylogenetic trees for both PB1 and gp64 proteins were inferred using PhyML v.3.3.2 and rooted on isavirids for PB1 sequences and depicted unrooted for gp64.

For each PB1 blast hit, we searched GenBank for the rest of the genome, ignoring any genomes that appear to have fewer than six segments on account of the three RdRp segments, nucleoprotein, and occasionally surface proteins being far easier to identify and all of the best-studied orthomyxovirids having at least six segments. We visualized PB1 and gp64 trees using baltic and annotated tips with number of segments identified and category of surface protein used, where available. For annotating genome organization, we further marked the earliest plausible common ancestors that must have possessed a given genome organization and highlighted all of their descendants as a prediction for which other data sets might have the missing segments.

Phylogenetic diversity estimation

The larger PB1 sequence data set (prior to reduction) was used to infer a maximum likelihood tree using PhyML v.3.3.2 which was rooted on isavirids. For each protein, the date of either its publication in literature or on GenBank was noted. For each year of discovery available, tree branches were marked with the evolutionary path uncovered that year, starting from oldest published sequences. The sum of branch lengths contributed by any given sequence to the tree is what we call phylogenetic diversity. As well as the relationship between year of discovery and maximum PD contributed in Fig. 3A, we looked at successive and unique PD contributions by each newly discovered orthomyxovirid in comparison to a neutral PD discovery curve.

ACKNOWLEDGMENTS

G.D. acknowledges the support of European Molecular Biology Organization (EMBO) installation grant IG-5305–2023. We would like to acknowledge the contributions of Amy Kistler, Maira Phelps, Cristina Tato, and Fabiano Oliveira in setting up sample logistics, experimental design, and data analysis for the Californian mosquito virome study. We would like to thank Darren Obbard for numerous and fruitful discussions.

We are grateful to Jessica Manning, Dara Kong, Sreyngim Lay, Alex Greninger, Mang Shi, Eddie C Holmes, Dana Price, and John Pettersson for sharing assembled sequence data.

AUTHOR AFFILIATIONS

[1]Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania
[2]Chan Zuckerberg Biohub, San Francisco, California, USA

AUTHOR ORCIDs

Gytis Dudas http://orcid.org/0000-0002-0227-4158
Joshua Batson http://orcid.org/0000-0002-9244-2142

FUNDING

Funder	Grant(s)	Author(s)
European Molecular Biology Organization (EMBO)	IG-5305-2023	Gytis Dudas

AUTHOR CONTRIBUTIONS

Gytis Dudas, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing | Joshua Batson, Conceptualization, Methodology, Supervision, Visualization, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

Data and scripts to replicate analyses are publicly available at https://github.com/evogytis/orthomyxo-metagenomics.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental Material (JVI01056-23-s0001.pdf). Fig. S1 to S10.

REFERENCES

1. Roux S, Matthijnssens J, Dutilh BE. 2021. Metagenomics in virology. Encyclopedia of Virology:133–140. https://doi.org/10.1016%2FB978-0-12-809633-8.20957-6
2. Obbard DJ, Shi M, Roberts KE, Longdon B, Dennis AB. 2020. A new lineage of segmented RNA viruses infecting animals. Virus Evol 6:vez061. https://doi.org/10.1093/ve/vez061
3. Shi C, Beller L, Deboutte W, Yinda KC, Delang L, Vega-Rúa A, Failloux A-B, Matthijnssens J. 2019. Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. Microbiome 7:121. https://doi.org/10.1186/s40168-019-0734-2
4. Li C-X, Shi M, Tian J-H, Lin X-D, Kang Y-J, Chen L-J, Qin X-C, Xu J, Holmes EC, Zhang Y-Z. 2015. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. Elife 4:e05378. https://doi.org/10.7554/eLife.05378
5. Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K, Wang W, Eden J-S, Shen J-J, Liu L, Holmes EC, Zhang Y-Z. 2018. The evolutionary history of vertebrate RNA viruses. Nature 561:E6. https://doi.org/10.1038/s41586-018-0310-0
6. Wang Z-D, Wang B, Wei F, Han S-Z, Zhang L, Yang Z-T, Yan Y, Lv X-L, Li L, Wang S-C, Song M-X, Zhang H-J, Huang S-J, Chen J, Huang F-Q, Li S, Liu H-H, Hong J, Jin Y-L, Wang W, Zhou J-Y, Liu Q. 2019. A new segmented virus associated with human febrile illness in China. N Engl J Med 380:2116–2125. https://doi.org/10.1056/NEJMoa1805068
7. Qin X-C, Shi M, Tian J-H, Lin X-D, Gao D-Y, He J-R, Wang J-B, Li C-X, Kang Y-J, Yu B, Zhou D-J, Xu J, Plyusnin A, Holmes EC, Zhang Y-Z. 2014. A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors. Proc Natl Acad Sci U S A 111:6744–6749. https://doi.org/10.1073/pnas.1324194111
8. Ge X-Y, Wang N, Zhang W, Hu B, Li B, Zhang Y-Z, Zhou J-H, Luo C-M, Yang X-L, Wu L-J, Wang B, Zhang Y, Li Z-X, Shi Z-L. 2016. Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. Virol Sin 31:31–40. https://doi.org/10.1007/s12250-016-3713-9
9. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. 2020. A new coronavirus associated with human respiratory disease in China. Nature 580:E7. https://doi.org/10.1038/s41586-020-2202-3
10. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JFY, Sharp PM, Shaw GM, Peeters M, Hahn BH. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313:523–526. https://doi.org/10.1126/science.1126531
11. Wheeler DC, Waller LA, Biek R. 2010. Spatial analysis of feline immunodeficiency virus infection in cougars. Spat Spatiotemporal Epidemiol 1:151–161. https://doi.org/10.1016/j.sste.2010.03.009
12. Parry R, Wille M, Turnbull OMH, Geoghegan JL, Holmes EC. 2020. Divergent influenza-like viruses of amphibians and fish support an ancient evolutionary association. Viruses 12:1042. https://doi.org/10.3390/v12091042
13. Garry CE, Garry RF. 2008. Proteomics computational analyses suggest that baculovirus GP64 superfamily proteins are class III penetrenes. Virol J 5:28. https://doi.org/10.1186/1743-422X-5-28
14. Shi M, Neville P, Nicholson J, Eden J-S, Imrie A, Holmes EC. 2017. High-resolution metatranscriptomics reveals the ecological dynamics of mosquito-associated RNA viruses in Western Australia. J Virol 91:e00680-17. https://doi.org/10.1128/JVI.00680-17
15. Pettersson J-O, Shi M, Eden J-S, Holmes EC, Hesson JC. 2019. Meta-transcriptomic comparison of the RNA viromes of the mosquito vectors *Culex pipiens* and *Culex torrentium* in northern Europe. Viruses 11:1033. https://doi.org/10.3390/v11111033
16. Batson J, Dudas G, Haas-Stapleton E, Kistler AL, Li LM, Logan P, Ratnasiri K, Retallack H. 2021. Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. Elife 10:e68353. https://doi.org/10.7554/eLife.68353
17. Müller NF, Stolz U, Dudas G, Stadler T, Vaughan TG. 2020. Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. Proc Natl Acad Sci U S A 117:17104–17111. https://doi.org/10.1073/pnas.1918304117
18. Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, Brouqui J-M, Bayne EH, Longdon B, Buck AH, Lazzaro BP, Akorli J, Haddrill PR, Obbard DJ. 2015. The discovery, distribution, and evolution of viruses associated with drosophila melanogaster. PLoS Biol 13:e1002210. https://doi.org/10.1371/journal.pbio.1002210
19. Allison AB, Ballard JR, Tesh RB, Brown JD, Ruder MG, Keel MK, Munk BA, Mickley RM, Gibbs SEJ, Travassos da Rosa APA, Ellis JC, Ip HS, Shearn-Bochsler VI, Rogers MB, Ghedin E, Holmes EC, Parrish CR, Dwyer C. 2015. Cyclic avian mass mortality in the Northeastern United States is associated with a novel orthomyxovirus. J Virol 89:1389–1403. https://doi.org/10.1128/JVI.02019-14
20. Batts WN, LaPatra SE, Katona R, Leis E, Ng TFF, Brieuc MSO, Breyta RB, Purcell MK, Conway CM, Waltzek TB, Delwart E, Winton JR. 2017. Molecular characterization of a novel orthomyxovirus from rainbow and steelhead trout (*Oncorhynchus mykiss*). Virus Res 230:38–49. https://doi.org/10.1016/j.virusres.2017.01.005
21. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37. https://doi.org/10.1093/nar/gkr367

22. Koonin EV, Dolja VV. 2014. Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 78:278–303. https://doi.org/10.1128/MMBR.00049-13
23. Kobayashi M, Toyoda T, Ishihama A. 1996. Influenza virus PB1 protein is the minimal and essential subunit of RNA polymerase. Arch Virol 141:525–539. https://doi.org/10.1007/BF01718315
24. Lum D, Rheindt FE, Chisholm RA. 2022. Tracking scientific discovery of avian phylogenetic diversity over 250 years. Proc R Soc B: Biol Sci 289:20220088. https://doi.org/10.1098/rspb.2022.0088
25. Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat 125:1–15. https://doi.org/10.1086/284325
26. Lounibos LP. 2002. Invasions by insect vectors of human disease. Annu Rev Entomol 47:233–266. https://doi.org/10.1146/annurev.ento.47.091201.145206
27. Fonseca DM, Smith JL, Wilkerson RC, Fleischer RC. 2006. Pathways of expansion and multiple introductions illustrated by large genetic differentiation among worldwide populations of the southern house mosquito. Am J Trop Med Hyg 74:284–289. https://doi.org/10.4269/ajtmh.2006.74.284
28. Bataille A, Cunningham AA, Cedeño V, Cruz M, Eastwood G, Fonseca DM, Causton CE, Azuero R, Loayza J, Martinez JDC, Goodman SJ. 2009. Evidence for regular ongoing introductions of mosquito disease vectors into the Galápagos islands. Proc R Soc B: Biol Sci 276:3769–3775. https://doi.org/10.1098/rspb.2009.0998
29. Huestis DL, Dao A, Diallo M, Sanogo ZL, Samake D, Yaro AS, Ousman Y, Linton Y-M, Krishna A, Veru L, Krajacich BJ, Faiman R, Florio J, Chapman JW, Reynolds DR, Weetman D, Mitchell R, Donnelly MJ, Talamas E, Chamorro L, Strobach E, Lehmann T. 2019. Windborne long-distance migration of malaria mosquitoes in the Sahel. Nature 574:404–408. https://doi.org/10.1038/s41586-019-1622-4
30. Di Giallonardo F, Geoghegan JL, Docherty DE, McLean RG, Zody MC, Qu J, Yang X, Birren BW, Malboeuf CM, Newman RM, Ip HS, Holmes EC. 2016. Fluid spatial dynamics of West Nile virus in the United States: rapid spread in a permissive host environment. J Virol 90:862–872. https://doi.org/10.1128/JVI.02305-15
31. Li Z, Fan Y, Wei J, Mei X, He Q, Zhang Y, Li T, Long M, Chen J, Bao J, Pan G, Li C, Zhou Z. 2019. Baculovirus utilizes cholesterol transporter NIEMANN-Pick C1 for host cell entry. Front Microbiol 10:2825. https://doi.org/10.3389/fmicb.2019.02825
32. de Jong JC, Smith DJ, Lapedes AS, Donatelli I, Campitelli L, Barigazzi G, Van Reeth K, Jones TC, Rimmelzwaan GF, Osterhaus A, Fouchier RAM. 2007. Antigenic and genetic evolution of swine influenza a (H3N2) viruses in Europe. J Virol 81:4315–4322. https://doi.org/10.1128/JVI.02458-06
33. Lycett SJ, Duchatel F, Digard P. 2019. A brief history of bird flu. Philos Trans R Soc Lond B Biol Sci 374:20180257. https://doi.org/10.1098/rstb.2018.0257
34. Bedford T, Cobey S, Pascual M. 2011. Strength and tempo of selection revealed in viral gene genealogies. BMC Evol Biol 11:220. https://doi.org/10.1186/1471-2148-11-220
35. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. Trends Ecol Evol 18:481–488. https://doi.org/10.1016/S0169-5347(03)00216-7
36. Longdon B, Wilfert L, Obbard DJ, Jiggins FM. 2011. Rhabdoviruses in two species of drosophila: vertical transmission and a recent sweep. Genetics 188:141–150. https://doi.org/10.1534/genetics.111.127696
37. Anxolabéhère D, Kidwell MG, Periquet G. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of drosophila melanogaster by mobile P elements. Mol Biol Evol 5:252–269. https://doi.org/10.1093/oxfordjournals.molbev.a040491
38. Coatsworth H, Bozic J, Carrillo J, Buckner EA, Rivers AR, Dinglasan RR, Mathias DK. 2022. Intrinsic variation in the vertically transmitted core virome of the mosquito *Aedes aegypti*. Mol Ecol 31:2545–2561. https://doi.org/10.1111/mec.16412
39. Baker SF, Nogales A, Finch C, Tuffy KM, Domm W, Perez DR, Topham DJ, Martínez-Sobrido L. 2014. Influenza A and B virus intertypic reassortment through compatible viral packaging signals. J Virol 88:10778–10791. https://doi.org/10.1128/JVI.01440-14
40. Chare ER, Gould EA, Holmes EC. 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. J Gen Virol 84:2691–2703. https://doi.org/10.1099/vir.0.19277-0
41. Kondo H, Maeda T, Shirako Y, Tamada T. 2006. Orchid fleck virus is a rhabdovirus with an unusual bipartite genome. J Gen Virol 87:2413–2421. https://doi.org/10.1099/vir.0.81811-0
42. Ke R, Aaskov J, Holmes EC, Lloyd-Smith JO. 2013. Phylodynamic analysis of the emergence and epidemiological impact of transmissible defective dengue viruses. PLoS Pathog 9:e1003193. https://doi.org/10.1371/journal.ppat.1003193
43. Arunkumar GA, Bhavsar D, Li T, Strohmeier S, Chromikova V, Amanat F, Bunyatov M, Wilson PC, Ellebedy AH, Boons G-J, Simon V, de Vries RP, Krammer F. 2021. Functionality of the putative surface glycoproteins of the Wuhan spiny eel influenza virus. Nat Commun 12:6161. https://doi.org/10.1038/s41467-021-26409-2
44. Whitfield ZJ, Dolan PT, Kunitomi M, Tassetto M, Seetin MG, Oh S, Heiner C, Paxinos E, Andino R. 2017. The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti* genome. Curr Biol 27:3511–3519. https://doi.org/10.1016/j.cub.2017.09.067
45. Holmes EC, Duchêne S. 2019. Can sequence phylogenies safely infer the origin of the global virome? mBio 10:e00289-19. https://doi.org/10.1128/mBio.00289-19
46. Frey KG, Biser T, Hamilton T, Santos CJ, Pimentel G, Mokashi VP, Bishop-Lilly KA. 2016. Bioinformatic characterization of mosquito viromes within the eastern United States and Puerto Rico: discovery of novel viruses. Evol Bioinform Online 12:1–12. https://doi.org/10.4137/EBO.S38518
47. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinform 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
48. Konstantinidis K, Dovrolis N, Kouvela A, Kassela K, Freitas MGR, Nearchou A, Williams M de C, Veletza S, Karakasiliotis I. 2021. Defining virus-carrier networks that shape the composition of the mosquito core virome of an ecosystem. In Review. https://doi.org/10.21203/rs.3.rs-229254/v1
49. He X, Yin Q, Zhou L, Meng L, Hu W, Li F, Li Y, Han K, Zhang S, Fu S, Zhang X, Wang J, Xu S, Zhang Y, He Y, Dong M, Shen X, Zhang Z, Nie K, Liang G, Ma X, Wang H. 2021. Metagenomic sequencing reveals viral abundance and diversity in mosquitoes from the Shaanxi-Gansu-Ningxia region, China. PLoS Negl Trop Dis 15:e0009381. https://doi.org/10.1371/journal.pntd.0009381
50. Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, Banfield JF, de la Peña M, Korobeynikov A, Chikhi R, Babaian A. 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature 602:142–147. https://doi.org/10.1038/s41586-021-04332-2
51. Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518. https://doi.org/10.1093/nar/gki198
52. Hasegawa M, Kishino H, Yano T-A. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174. https://doi.org/10.1007/BF02101694
53. Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314. https://doi.org/10.1007/BF00160154
54. Sagulenko P, Puller V, Neher RA. 2018. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol 4:vex042. https://doi.org/10.1093/ve/vex042
55. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinform 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033
56. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:vey016. https://doi.org/10.1093/ve/vey016
57. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst Biol 67:901–904. https://doi.org/10.1093/sysbio/syy032
58. Heled J, Bouckaert RR. 2013. Looking for trees in the forest: summary tree from posterior samples. BMC Evol Biol 13:221. https://doi.org/10.1186/1471-2148-13-221
59. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu C-H, Xie D, Zhang C, Stadler T, Drummond AJ. 2019. BEAST 2.5: an advanced software platform for

Bayesian evolutionary analysis. PLOS Comput Biol 15:e1006650. https://doi.org/10.1371/journal.pcbi.1006650
60. Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55
61. Lemey P, Minin VN, Bielejec F, Kosakovsky Pond SL, Suchard MA. 2012. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. Bioinform 28:3248–3256. https://doi.org/10.1093/bioinformatics/bts580
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2