nature communications

Article

https://doi.org/10.1038/s41467-026-73605-z

Diversity and distinctive characteristics of the global RNA virome in urban and peri-urban environments

Received: 25 November 2025

Accepted: 12 May 2026

Published online: 16 July 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

RNA viruses represent an integral component of human-associated environments and human health. However, the ecology of environmental RNA viruses remains largely unexplored. Here, we analyzed 2922 metatranscriptomic samples collected from urban and surrounding environments—including human-dense settings (e.g., transit hubs, hospitals, banks), alongside peri-urban settings – across 102 cities in 31 countries and constructed the Urban & Peri-urban RNA Virus Atlas (UPVAtlas), comprising 54,945 RNA viruses, 77% of which had not been previously observed. Phylogenetic reconstruction based on RNA-dependent RNA polymerases from UPVAtlas greatly expanded the evolutionary diversity of RNA viruses, leading to the identification of two potential candidate phyla, one candidate class, and several unclassified clades. Host association analyses further revealed the ecological complexity of environmental RNA viruses, with the diversity of vertebrate-related and ESKAPE pathogen-related viruses underscoring the importance of continued monitoring of urban environments for tracking RNA viral prevalence and dynamics, with direct relevance to future public health.

RNA viruses represent one of the most evolutionarily diverse and ecologically ubiquitous groups of biological entities on Earth. However, the vast majority of their diversity, often referred to as the viral "dark matter", remains uncharacterized and understudied across global ecosystems[1]. It is estimated that Earth harbors upwards of 10^{31} viral particles, but fewer than 1% of this staggering diversity has been formally described[2]. Among these, RNA viruses cause more concern due to their intrinsically high mutation rates, up to a million times higher than their hosts[2–4]. More importantly, the novel viruses discovered comprise over two-thirds of all human pathogens[5]. Hence, identifying new RNA viruses is a top priority towards elucidating the spectrum of viral diversity and expediting their integration into various research domains, including pharmacology, genetics, agriculture, and other areas[6–9]. Unveiling the distribution pattern of RNA viruses is of utmost importance for a comprehensive assessment of their potential ramifications on human health and the environment.

Recent advances in sequencing technologies and bioinformatic methods have enhanced our ability to uncover unknown RNA viruses and understand their phylogeny in various environments. For example, Hillary et al. identified 3462 viral contigs in RNA virome from purified virus-like particles in five soil types, suggesting that RNA viral communities have the potential to exert influence on inter-kingdom interactions across terrestrial biomes[10]. Chen et al. collected 442 metatranscriptomic samples from 32 diverse environments in 16 provinces and regions of China to expand the known diversity of the RNA virome. The authors identified 6624 putative novel viral operational taxonomic units (vOTUs), classified into at least 62 families[11]. Oceans are also known to be a reservoir of viral genetic diversity[12]. To expand the knowledge of global marine RNA virus diversity, ecology and ecosystem roles, Dominguez-Huerta et al. analyzed the Tara Ocean samples and identified 5504 RNA viruses[13]. Neri et al. analyzed more than 330,000 RNA-dependent RNA polymerases (RdRPs) uncovered from 5150 diverse metatranscriptomes, identifying two candidate novel phyla and revealing that the diversity of RNA viruses on Earth is at least five times larger than previously recorded[14]. To investigate the evolutionary origins of marine RNA viruses, Zayed et al. analyzed 771

e-mail: jwu@bio.ecnu.edu.cn; linfa.wang@duke-nus.edu.sg; chm2042@med.cornell.edu; tlshi@bio.ecnu.edu.cn

ocean metatranscriptomes, and identified 44,779 RdRP-encoding contigs doubling the size of the orthornaviran phyla[15]. Because ~75% of emerging human infectious diseases originate from animals, He et al. performed a metatranscriptomic analysis of 1941 game animals, representing 18 different wildlife species traded or consumed as food across China and identified 102 mammalian-infecting viruses, more than half of which were previously uncharacterized[16]. Finally, Hou et al. innovatively employed artificial intelligence methods to discover 161,979 potential RNA virus species[17], and these and other studies have revealed a high phylogenetic diversity among RNA viruses. To gain a deeper understanding of this diversity, researchers use a phenomenon known as amino acid directional selection[18,19]. This form of natural selection promotes extreme shifts in allele frequencies over time. Variations often emerge from the fitness advantages organisms gain within specific ecological niches, shaped by factors such as geolocation, climate, and environmental conditions. By exploring these dynamics, researchers have provided valuable insights into the evolutionary pressures and adaptations experienced by viruses, including SARS-CoV-2, Zika, and Dengue[20–24]. The findings emphasize the intricate nature of viral evolution and highlight the pivotal role of directional selection in shaping interactions between viruses and their environments.

Crucially, since the Neolithic Revolution, the trend of urbanization has been consistently rising, propelling substantial human migrations towards cities and giving rise to heightened population density, along with increasingly interconnected transportation networks that serve as dissemination hub of epidemics, thereby emerging as a significant threat to humanity[25,26]. RNA viruses have historically served as influential agents in the survival or demise of human populations as they drove the evolution of our immune system[27]. Despite their global impact on humans, little attention has been paid over the past century to the distribution of RNA viruses in rapidly expanding urban environments, where 56.2% of the world's population dwells[28]. This knowledge gap in human-virus interactions is particularly concerning, as urban environments may serve as undiscovered reservoirs and sources for novel RNA viruses. These environments can facilitate the persistence, transmission and diversification, with broad implications ranging from public health crises to privacy concerns[29]. In particular, the COVID-19 pandemic has shown that densely populated cities, such as New York and Shanghai, often bear the brunt of emergencies[30,31]. To improve our preparation for epidemics and reduce their economic costs, it is imperative to carry out large-scale surveys of the microbiome to convey early warnings for epidemics[32] and will facilitate "smart cities" envisioned. Additionally, considering the continuous spread and evolution of viruses, there is an urgent need to develop scalable, cost-effective and long-term passive surveillance systems. Indeed, genomic catalogs of RNA viruses can improve our understanding of viral phylogeny and origins, their roles in the environment and seasonality, and their pathogenicity.

Previously, we reported that 94.1% of the identified DNA viruses did not match any viral sequence at the species level in the IMG/VR database, highlighting the vast number of unidentified DNA viruses in urban environments[33]. Hence, it is reasonable to believe that numerous unidentified RNA viruses also remain to be discovered in and around cities. Thoroughly understanding the RNA virus genomes from diverse urban environments would provide valuable insight into the viral ecology of urban areas and their impact on public health, urban planning and sustainability. To explore this notion, we analyzed 2922 metatranscriptomes obtained from a wide range of urban environments, leading to the discovery of over 40,000 previously unseen RNA viral operational taxonomic units (RvOTUs). This large collection of unknown viruses has enabled the identification of two potential candidate phyla, a novel class within *Kitrinoviricota*, and several previously unseen clades under established phyla. Furthermore, we characterized the evolutionary and directional selection patterns within five established phyla. Finally, by cross-referencing two virus-host interaction databases and two CRISPR spacer databases, we proposed potential hosts for approximately one-third of these viruses and investigated the diversity and genomic evolution of viruses associated with ESKAPE pathogens[34,35] and vertebrates.

Results

A genomic catalog of RNA viruses from diverse environments

From March to July 2020, the MetaSUB Consortium[36] collected 1,862 metatranscriptomes from cities worldwide. Sampling sites were chosen in high-population-density urban areas to serve as a proxy for the urban microbiome. All samples were obtained from commonly touched surfaces (e.g., handrails, benches, ticket machines) and subjected to shotgun metatranscriptomic sequencing (see Methods and Materials). In addition, to broaden the representation of human-associated environments in our sampling, we selectively retrieved non-built, peri-urban environmental samples from the Sequence Read Archive (SRA) database, thereby maximizing coverage of the RNA virome across environments accessible in daily human life. Notably, certain animal-associated samples were also included, as these animals are frequently traded as edible commodities in local markets of the corresponding regions. Accordingly, a series of metatranscriptomes with geographic records, originating from multiple BioProjects and spanning diverse environments worldwide, were incorporated into our analytical framework.

In total, we analyzed 2922 metatranscriptomic samples, spanning 102 cities across 31 countries, comprising a dataset of more than 100 billion reads. The extensive datasets under investigation encompass a broad spectrum of environments, including surfaces of urban environments consisting of public transit stations (e.g., bus stations, subway stations, and train stations) and other high-population-density locations (e.g., hospitals and banks, etc.), supplemented by some urban perimeter environments (e.g., soil, green spaces, water, wastewater, etc.) (Fig. 1A, Supplementary Data 1). Following the initial processing, which adhered to established quality-control standards[33], the sequencing data were subsequently assembled into an extensive contig set comprising nearly 300 million contigs, including 22.21 million contigs not shorter than 1 kb. Subsequently, these 22.21 million contigs were fed into geNomad[37] to extract the putative viral contigs. This process yielded a final set of 675,322 putative viral contigs, of which 241,082 (35.7%) were identified as members of the *Orthornavirae* kingdom (Fig. 1B), encompassing viruses characterized by RNA-based genomes. The *Orthornavirae* contigs were further grouped into 84,438 clusters based on 90% average nucleotide identity (ANI) and 90% alignment fraction (AF), similar to thresholds used in previous studies[14,15]. Of these 54,945 had a RdRP core domain detected, passed the QC filters (see Methods and Materials), and were hereafter designated the Urban & Peri-urban RNA Virus Atlas (UPVAtlas). The longest contigs in each cluster were selected as the representatives, which were regarded as RNA viral operational taxonomic units (hereinafter referred to as RvOTUs).

While our data revealed a substantial number of RvOTUs, the accumulation curve indicates that global RNA virus diversity has not yet reached saturation for almost all sample types (Fig. 1C, D). Although the richness of RvOTUs in the samples collected from artificial environments, such as stores, wastewater facilities, public transit stations, and hospitals, is lower than that in natural samples, RNA virus diversity always remains far from saturation (Fig. 1D, Supplementary Fig. 1).

By comparing between the RvOTUs with known RNA viruses from various catalogs and studies including 7064 RNA viral genomes in the ICTV (International Committee on Taxonomy of Viruses) database[38], 5504 RvOTUs identified in Zayed et al.'s study, 124,742 RvOTUs in Neri et al.'s study and 27,986 RNA viral genomes used as references in Neri et al.'s study, we found that only 12,487 (22.73%) RvOTUs contained at

https://doi.org/10.1038/s41467-026-73605-z

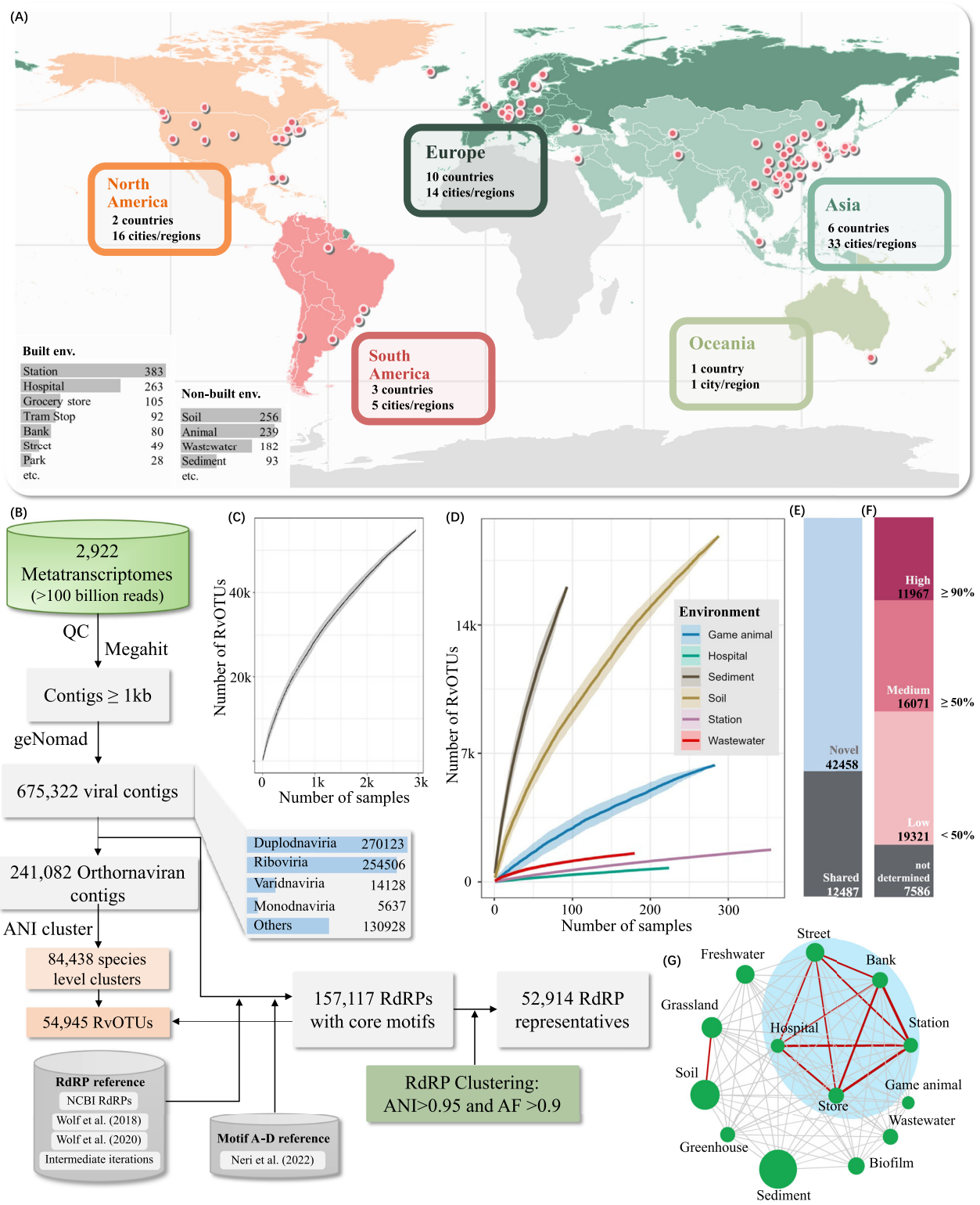

Fig. 1 | Pipeline for discovering RNA viruses in metatranscriptomic datasets. **A** World map shows the locations of samples with recorded geographical location. The map was generated using R packages 'ggplot2' (v4.0.0) and 'maps' (v3.4.3). **B** Pipeline used for uncovering RNA viruses using metatranscriptomic data. **C** Urban-RvOTU rarefaction curves were generated using all the samples in this study. The values were obtained via bootstrapping. The shaded areas represent the mean number of RvOTUs ± standard deviation of measured Urban-RvOTUs through 100 random subsampling. The line represents the mean of 100 random samples. **D** Urban-RvOTU rarefaction curves using the samples collected in major environments (station, hospital, soil, game animal, wastewater and sediment), colors indicate the environment type. The shaded areas represent the mean ± standard deviation for each environment, calculated as in (**C**). **E** Numbers of novel and shared (generated species level match with reference library of RNA viruses composed of 4 distinct datasets) Urban-RvOTUs identified in this study. **F** Numbers of Urban-RvOTUs with different genome completeness. **G** The network graph illustrates the similarity among varied environments concerning prevalent Urban-RvOTUs. Green nodes represent distinct environments, with node size corresponding to the number of collected samples. Edges colored in red signify a Jaccard index exceeding 0.1 and a BH-adjusted $p < 0.001$.

least one previously known sequence, at the ANI90 level of diversity (RdRP core domain sequences, Supplementary Fig. 2, Fig. 1E, Supplementary Data 2). The lengths of these RvOTUs followed a truncated power law distribution, with the vast majority (92.03%) being shorter than 5 kb (Supplementary Data 3); however, more than half of the RvOTUs (28038, 51%, Fig. 1F) still achieved a medium level of genome completeness (> 50%).

We further explored the similarity across different environments with respect to the prevalent RvOTUs (Fig. 1G, see Methods and Materials). We found that prevalent RvOTUs exhibited distinct profiles across various environments, although a notable level of co-occurrence was observed among urban-prevalent RvOTUs in artificial environments characterized by heightened human activity, including public transit stations, stores, banks, streets, and hospitals.

Two potential candidate phyla expand the phylogenetic diversity of RNA viruses

The RdRP enzyme is crucial for the replication of RNA viruses and boasts a remarkable degree of conservation across various RNA virus families, making it a key marker for constructing RNA virus phylogenetic trees. To investigate the phylogenetic diversity of the RNA virome and categorize the newly identified RNA viruses, we gathered RdRP core domain sequences from all the viral contigs classified as *Orthornavirae* (see Methods and Materials), resulting in the recovery of 160,016 RdRPs. Clustering with a 95% amino acid identity threshold over a 90% alignment fraction resulted in 52,914 representatives referred to as the RCR95 set. To avoid false positives in RdRP identification, we applied a similar strategy used in previous studies[39,40], which consists of a homology search to exclude group II introns and reverse transcriptases (RTs)—the primary sources of false positive signals for RdRP detection. The RCR95 representatives were compared with the 3992 group II introns and RTs collected from the NCBI nr database using blastp, and no positive matches were identified (Supplementary Data 4).

We employed the geNomad[37] to perform taxonomic annotation on the RCR95 representative, and the results showed that geNomad offers consistent taxonomic annotation at both phylum and class levels (Supplementary Fig. 3A, B) within clusters generated under different identity thresholds. Ultimately, 91.51% RCR95 representatives got taxonomic annotated at phylum level or lower by geNomad, attributing them to five ICTV-recognized phyla in the megataxonomy of *Orthornavirae*: *Lenarviricota* (46.52%), *Pisuviricota* (23.70%), *Kitrinoviricota* (14.37%), *Duplornaviricota* (4.74%), and *Negarnaviricota* (2.19%). It is noteworthy that, despite the substantial expansion of our dataset, the five phyla previously established continue to dominate the overall phylogenetic diversity of the RNA virome.

To assess the presence of novel phyla within the RCR95 set, we partitioned the 52,914 RCR95 representatives into 4958 clusters using MMSeq2[41] applying a sequence identity threshold of 0.3 to achieve the minimal identity threshold for phylum level clustering (see Methods and Materials, and Supplementary Data 5). Within each cluster, we evaluated the annotation purity across different taxonomic levels. The results demonstrate the high internal annotation consistency at both the phylum and class levels of the geNomad annotation (Supplementary Fig. 3B). Consequently, we uniformly assigned each RCR95 representative within a cluster to the dominant phylum. Only 25 clusters (0.5%), each comprising over 20 members and totaling 1338 representatives in the RCR95 set, lacked phylum-level annotation. To address this gap, we subsequently constructed a maximum likelihood phylogenetic tree to elucidate the phylogenetic relationships between these unannotated RCR95 representatives and the five well-established phyla within *Orthornavirae* (Fig. 2A, see Methods and Materials). The resulting RdRP tree comprised 1984 representative sequences, including 1338 RCR95 representatives lacking geNomad annotation at the phylum level, 643 RdRP core domain sequences extracted from ICTV genomes distributed across five well-established phyla of *Orthornavirae* (*Pisuviricota*: 180, *Kitrinoviricota*: 86, *Lenarviricota*: 168, *Negarnaviricota*: 193, and *Duplornaviricota*: 16), and 3 reverse transcriptases (RTs) included as an outgroup. In addition, we incorporated the 33 RdRPs from a recently identified phylum (*p.0002*), as reported by Neri et al. in their virome study[14], into the phylogenetic tree. We compared the candidate viral phyla proposed in previous studies[14,15] with representatives of the RCR95 set (see Materials and Methods). This preliminary comparison revealed blastp matches between Neri et al.'s *p.0002* and our RCR95 members; accordingly, *p.0002* was included in our phylogenetic reconstruction alongside ICTV-recognized reference viral sequences. All other candidate phyla from the aforementioned studies were excluded from this analysis, as they have not yet been certified by the ICTV.

Ultimately, we found that 18 out of the 25 clusters were phylogenetically assigned to the phylum *Kitrinoviricota*, while the remaining seven clusters formed two deeply branching clades in the RdRP phylogenetic tree, representing two proposed candidate phyla, designated as *upv.p.001* and *upv.p.002*. Our findings demonstrate that the candidate phylum *upv.p.002* clusters closely with the candidate phylum *p.0002* identified by Neri et al.[14] After examining the composition of different phyla within the samples, we observed that two novel candidate phyla exhibited relatively low abundance, yet we still examined their environmental niches (Fig. 2B). The findings revealed that *upv.p.001* was predominantly presented in the wastewater, soil and sediment samples, while *upv.p.002* was primarily identified in grassland samples.

In terms of geographical distribution, the sequences represented by RCR95 members of both *upv.p.*001 and *upv.p.*002 were derived from metatranscriptomic samples collected in the Northern Hemisphere (Fig. 2C). Specifically, *upv.p.*001 was detected across North America, Europe, and Asia, whereas *upv.p.*002 was recovered only from a limited number of samples in North America and Europe.

Subsequently, we explored the domain content within the two novel phyla, *upv.p.001* and *upv.p.002*, to provide further insights into their evolutionary significance. The ORFs were first detected within the contigs of the two novel phyla, and then functionally annotated (see Methods and Materials). Overall, due to their exceptionally low similarity to any known sequences, many genes in these genomes could not be annotated. The genomes of *upv.p.001* exhibited a length of approximately 13kbp and contained 14–15 open reading frames (ORFs), whereas those of *upv.p.002* are about 6kbp in length and comprise 2–4 ORFs. Specifically, the RdRPs were consistently located at the terminus of *upv.p.001* genomes (Fig. 2D), and an icosahedral capsid was detected upstream of the genomes, suggesting that viruses within *upv.p.001* likely possess the classic icosahedral or helical capsid structure (Fig. 2D). A screen of bacterial CRISPR-Cas spacers targeting *upv.p.001* revealed that its *Clostridioides difficile*-infectious members are widely distributed on the phylogenetic tree (Supplementary Fig. 4, see Methods and Materials). Spacers from at least seven different *C. difficile* targeted *upv.p.001* members, indicative of an ongoing evolutionary arms race. Additionally, we also observed *upv.p.001* members linked to *Listeria monocytogenes* and *Lachnospira eligens*. In summary, our observations suggest that the *upv.p.001* clade is more characteristic of a bacteriophage lineage, with no evidence found to support any eukaryotic infection potential.

Regarding *upv.p.002*, unexpectedly, a collagen domain was found within a member, an unusual occurrence since collagen proteins are not typically identified in the viral genome as per the ICTV reference sequences (Fig. 2D). This domain exhibited a high similarity (identity = 0.96) to a collagen-like gene from *Bacillus clarus*. This similarity is explained by a match between the viral genome and a CRISPR-Cas spacer from *B. clarus*, indicating that this virus likely captured a fragment of its host's gene during infection. These results strongly suggest

https://doi.org/10.1038/s41467-026-73605-z

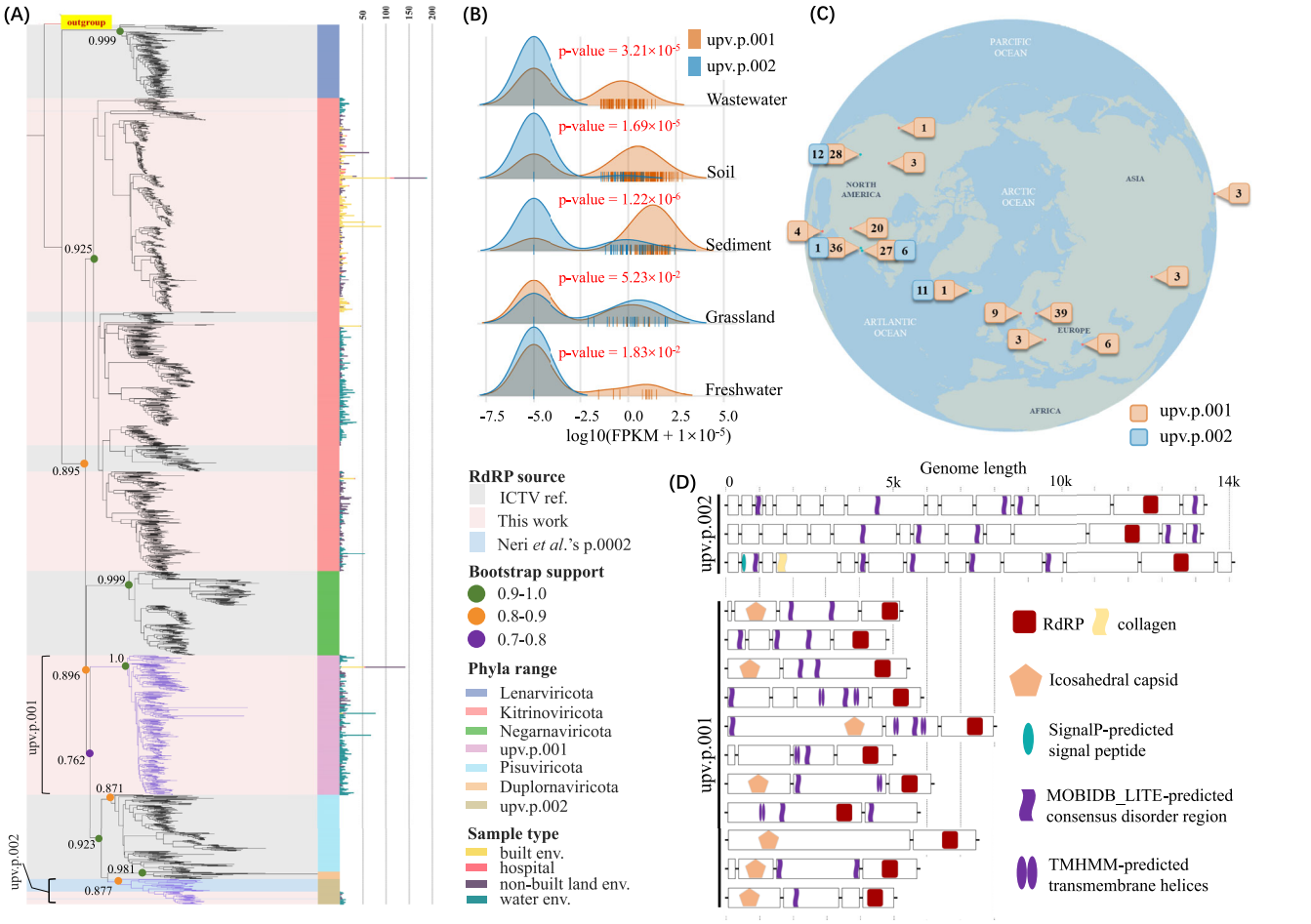

Fig. 2 | Phylogenetic reconstruction of the RNA viruses using the conserved RdRP core motifs. A Phylogenetic tree of RdRP core sequences of RCR95 representatives lacking phylum annotation and 643 RdRP domain sequences from the ICTV database. Thirty-three RdRP domain sequences from Neri et al.'s *p.0002* were also included. The tree was re-rooted using reverse transcriptases as an outgroup and visualized with iTOL. Clades representing two novel candidate phyla is marked in purple. The bars indicate the number of samples in which RdRPs were detected. (In sample type, built env. for station, park, street, etc. Non-built env. for soil, grassland, etc. Water env. for river water, waste water, sediment, etc.) **B** Relative abundance density of the two novel phyla across different environments. Environments where the relative abundance of both *upv.p.001* and *upv.p.002* is zero in over 90% of samples were omitted. The p-values were computed utilizing the two-sided Welch t-test. **C** Global distribution of source samples for the two potential novel phyla. Numbers indicate sample counts. The map was generated using R packages 'ggplot2' (v4.0.0), 'sf' (v1.0-20) and 'rnaturalearth' (v1.1.0). Made with Natural Earth. **D** Genome structures of *upv.p.001* and *upv.p.002* representatives. Similar domains are shown as the same shape and color.

that this *upv.p.002* member likely functions as a bacteriophage targeting *B. clarus*.

Phylogenetic nature and diversity of established phyla

To explore how the high taxonomic-level unclassified viral vOTUs in UPVAtlas could expand the known diversity of RNA viruses, we extracted members from RCR95 that lacked class-level classification and integrated them with RdRPs from ICTV reference sequences of the corresponding phyla to construct phylogenetic trees. Phylum *Duplornaviricota* had no suitable members meeting this criterion; given its recently reported potential polyphyly, we included all *Duplornaviricota* members from RCR95 in a comprehensive phylogenetic analysis alongside ICTV reference sequences from all known phyla. Overall, the phylogenetic tree faithfully reflected the structure of existing taxonomy in general, although certain taxonomic expansions and potential polyphyly also indicated the presence of novel candidate taxa (Fig. 3).

ICTV representatives formed four clusters corresponding to the four established classes within Phylum *Kitrinoviricota* (Fig. 3A), while the RCR95 members constituted a distinct clade, suggesting the existence of a potential candidate class, *k.c.c.01*. Members of this clade were derived from highly diverse sources (soil, 41.5%; animal, 20.8%; sediment, 11.3%; etc.). Given the broad host range of known *Kitrinoviricota* members (plants, invertebrates, vertebrates, etc.) and the absence of CRISPR spacer match in the genomes of this clade, we propose that the *k.c.c.01* members are likely eukaryote-hosted viruses, similar to other members of the phylum.

Phylum *Negarnaviricota* is divided at the subphylum level into *Haploviricotina* and *Polyploviricotina*. The former includes the classes *Yunchangviricetes*, *Chunqiuviricetes*, and *Monjiviricetes*, which possess mono-segmented genomes; the latter comprises the classes *Ellioviricetes* and *Insthoviricetes*, which have multi-segmented genomes. This taxonomic division, based on genomic architecture, inevitably renders *Haploviricotina* polyphyletic in the RdRP phylogenetic tree. However, from the perspective of RdRP evolution, members with multi-segmented genomes likely emerged later than the divergence of several classes with mono-segmented genomes. Moreover, RCR95 members formed a cluster adjacent to the sole member of *Yunchangviricetes* (Yuyuevirus). Since its formal recognition by ICTV in 2018, this group of invertebrate-associated viruses has contained only one genus with two species, all originating from China[42]. Our 27 RCR95 members significantly expand the known diversity of this group, with

https://doi.org/10.1038/s41467-026-73605-z

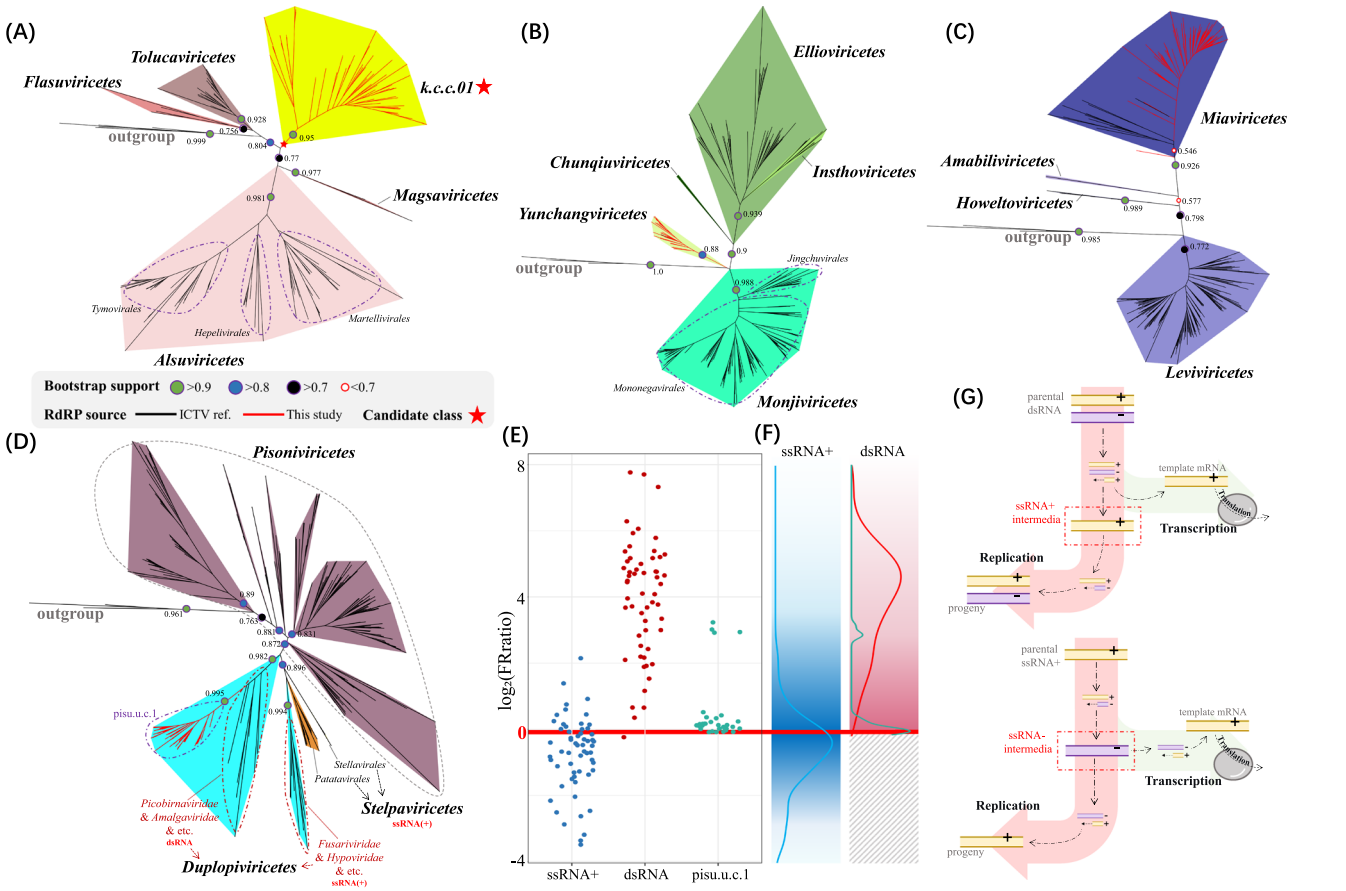

Fig. 3 | Phylogenetic reconstruction of established Phyla. A *Kitrinoviricota*, **B** *Negarnaviricota*, **C** *Lenarviricota*, **D** *Pisuviricota* phylogenetic trees of RdRP core domain sequences of RCR95 representatives lacking class level annotation and reference sequences from the ICTV database. Color ranges in each tree represent different classes. **E** LogFRratio of *pisu.u.c.1* members and two control groups of ssRNA+ and dsRNA RvOTUs. **F** Kernel density curves of LogFRratio distribution of three groups. The colored backgrounds show theoretical ranges of logFRratio for ssRNA+ and dsRNA viruses in shotgun metatranscriptomic sequencing with strand-specific library preparation. **G** Schematic overview of replication cycles of ssRNA+ and dsRNA viruses. The distinct polarity of viral genomes and their replicative intermediates accounts for their different logFRratio distributions.

novel sequences derived from multiple countries across North America and Europe (e.g., United States, Finland, Sweden), demonstrating a broader geographic distribution of these viruses than previously recognized.

Regarding *Lenarviricota* (Fig. 3C), all 102 RCR95 members were ultimately grouped around the reference sequences of *Miaviricetes*, expanding the diversity of this class. Among them, 99 members formed a distinct, novel clade. It is worth noting that one outlier RCR95 member was positioned outside the main *Miaviricetes* clade; however, given that its separation from the primary *Miaviricetes* cluster was supported by only a 0.546 bootstrap value, we consider it more likely that this outlier is part of *Miaviricetes* rather than representing a potential novel class.

Phylum *Pisuviricota* primarily consists of positive-sense single-stranded RNA (ssRNA+) viruses (94%), one of its classes, Class *Duplopiviricetes*, includes members with double-stranded RNA (dsRNA) genomes (e.g., *Picobirnaviridae*, *Amalgaviridae*), while others (e.g., *Fusariviridae*, *Hypoviridae*) possess positive-sense ssRNA genomes like other *Pisuviricota* classes. In the phylogenetic tree constructed with RCR95 members and ICTV reference sequences (Fig. 3D), *Duplopiviricetes* exhibited polyphyly: all of its ssRNA members diverged from the main cluster and instead grouped within a clade containing *Stelpaviricetes*, another class of ssRNA viruses. Both phylogenetic evidence and genomic architecture suggest that these ssRNA families within *Duplopiviricetes* are more closely related to *Stelpaviricetes*. Furthermore, *Pisoniviricetes* also displayed complex polyphyly from the perspective of RdRP phylogeny. The intricate evolutionary relationships and varied genome types within this phylum may have contributed to potential taxonomic challenges. In addition, the RCR95 members formed an evolutionarily distant, distinct cluster within the double-stranded clade of *Duplopiviricetes*, which we designated as *pisu.u.c.1*. We propose that these viruses also possess double-stranded genomes, suggesting additional dsRNA lineages in this phylum.

To verify their genome molecular type, we performed read mapping on samples containing *pisu.u.c.1* members, together with two control groups of confirmed high-quality ssRNA+ and dsRNA RvOTUs. Since these samples were prepared using a strand-specific library preparation protocol, this allowed us to determine the ratio of positive-sense to negative-sense reads and distinguish the molecular type of their genomes. The results showed that the log2 ratio of positive-sense to negative-sense reads (logFRratio, see Methods and Materials) was almost universally greater than 0 for the dsRNA control group (Fig. 3E), with a few exceptions likely attributable to sequencing technological stochasticity. In contrast, the logFRratio for the ssRNA+ control group was distributed on both sides of 0. This distribution pattern arises because dsRNA viruses encapsidate a 1:1 ratio of positive-sense to negative-sense strands in their genomes. The single-stranded positive-sense intermediates produced during their life cycle can only shift the baseline logFRratio (which is 0) towards positive

values (Fig. 3F, G), with the degree of shift depending on viral transcription activity. Conversely, ssRNA+ viral genomes have a 1:0 ratio. The varying levels of negative-sense RNA intermediate templates generated during transcription and replication can result in a logFR-ratio that is either positive or negative (Fig. 3G).

The logFRratio distribution of *pisu.u.c.1* members almost entirely fell on the positive side of zero, confirming their dsRNA nature. However, compared to the dsRNA control group, the logFRratio values for *pisu.u.c.1* members were generally smaller, suggesting lower viral transcription activity of *pisu.u.c.1* within these samples. It is noteworthy that a small number of samples exhibited marginally negative logFRratio values (minimum = −0.086). Given the low read counts of *pisu.u.c.1* in these samples, these negative values are most likely due to the stochasticity inherent in PCR amplification.

Phylum *Duplornaviricota* is composed entirely of members with double-stranded genomes. However, according to current RNA virus taxonomy, dsRNA viruses do not share a common double-stranded ancestor, which is well-illustrated by the dsRNA members within Phylum *Pisuviricota*. More recently, the monophyly of *Duplornaviricota* has been challenged, Zayed et al. suggested that viruses within *Duplornaviricota* are polyphyletic, warranting subdivision into three distinct phyla: (i) *Duplornaviricota*; *Chrymotiviricetes*, (ii) *Duplornaviricota*; *Vidaverviricetes* and (iii) *Duplornaviricota*; *Resentoviricetes*[15]. In addition, Neri et al. also found that some double-strand RNA virus families, e.g., *Reoviridae* and *Picobirnaviridae*, often diverge from their respective phyla[14]. To examine this proposal, we constructed a phylogenetic tree using RCR95 representatives that exhibited positive matches to RdRPs of *Duplornaviricota* in the ICTV database, along with referenced RdRPs across the five established phyla (Supplementary Fig. 5). Our results corroborated the polyphyletic nature of *Duplornaviricota*, aligning with the previous observations[14,15]. Two clades diverged from *Duplornaviricota* main lineage and clustered closely with branches of *Pisuviricota* and *Kitrinoviricota*. This pattern closely resembles the phenomenon we observed within *Pisuviricota*; *Duplopiviricetes*, and further suggests that the emergence of viruses with double-stranded RNA genomes was not an isolated event in RNA virus evolution.

Our findings in *Pisuviricota* and *Duplornaviricota* underscore the likelihood of multiple independent origins of double-stranded RNA viruses. The polyphyly observed in *Pisuviricota*; *Duplopiviricetes* and *Duplornaviricota* from the perspective of RdRP phylogeny indicates potential conflicts within the current taxonomic framework.

Directional selection of RNA viruses

Directional selection represents a form of natural selection where a phenotype is favored over other phenotypes. As a result, the allele frequency shifts over time in the direction of that favored phenotype. This process often targets specific genes or genomic regions that enhance the fitness of organisms within their niches (such as urban environments). In this study, we used FADE (FUBAR Approach to Directional Evolution), which is implemented in Hyphy[43]. Recently, FADE has been employed in viral evolutionary dynamic studies using the RASCL application to detect near-real-time selective pressures in viral clades[44]. The purpose of FADE is to assess whether specific amino-acid sites within a protein alignment are evolving toward a particular amino-acid residue at faster rates along certain branches, in comparison to a baseline model. Our observations revealed extensive diversity of RNA viruses across different environments, potentially reflecting evidence of directional selection in these viruses. We employed FADE to detect any evidence of directional selection in viral phyla present within the RCR95 set to compare evolutionarily complex and diverse phylum-level viral communities across different environments (see Methods and Materials for additional details on FADE). Overall, a total of 104 amino acid sites in RdRP exhibited evidence for directional selection across all the examined phyla, as detailed in Supplementary Data 6. The distribution of these sites varied among phyla (Table 1, Fig. 4A–G). *Lenaviricota* accounted for the majority with 57 sites (54.8%), followed by *Negarnaviricota* with 30 sites (28.8%), *Kitrinoviricota* at 7 sites (6.7%), *upv.p.001* with 6 sites (5.8%), *Pisuviricota* at 2 sites (1.9%), and both *upv.p.002* and *Duplornaviricota* contributing 1 site each (1.0%). Notably, when assessing the preference for specific amino acids, we observed a marked trend towards Aspartic acid [D] and Phenylalanine [F], which were favored in 12.5% of cases, closely followed by Asparagine and Cysteine, each accounting for 11.5% of the substitutions.

Considering the potential impact of sequence number imbalance s and variability in sample genetic diversity, our analysis prioritized the most prevalent amino acid substitutions. This approach revealed 47,778 substitutions within the statistically significant dataset. Among amino acid sites with at least four times at all amino acid targets, we identified 3272 instances of directional selection in the RdRp protein. Notably, certain sites demonstrated repeated substitutions indicative of multiple amino acid exchanges:

Lenarviricota amino acid site 76 (Targets: F, W): Evidence suggested a directional selection favoring Phenylalanine [F] and Tryptophan [W], with 7 noted substitutions from Tryptophan to Phenylalanine. Phenylalanine's role in modulating hydrogen bonding interactions within the RdRp pocket is significant, particularly at terminal nucleotides[45]. This site comprises Phenylalanine [F], Tryptophan [W], and Tyrosine [Y], all bearing hydrophobic aromatic side chains crucial for biological function.

Lenarviricota amino acid site 727 (Target: M): This amino acid site favors Methionine [M], a minor residue (2.2%) with 5 substitutions from Lysine. Methionine's hydrophobic side chain is implicated in the RdRp complex's functionality, particularly at the RNA template exit site[46].

Lenarviricota amino acid site 728 (Target: Y): Tyrosine [Y] is favored at this site, a minor residue (8.1%) with 22 substitutions from Phenylalanine. Tyrosine's unique phosphorylation capability distinguishes it from Phenylalanine.

Lenarviricota amino acid site 917 (Targets: F, N, Y): This site exhibits a diversifying selection preference, with 24 instances of Tyrosine [Y] to Phenylalanine [F] changes out of 33 substitutions (73%).

Negarnaviricota amino acid site 211 (Target: F): This amino acid site favors Phenylalanine [F], occupying ~20% of the site, with 9 substitutions from Tryptophan [W]. Phenylalanine's benzyl side chain facilitates hydrophobic interactions without causing steric hindrance.

Negarnaviricota amino acid site 212 (Targets: C, N): At this amino acid site, with 10 occurrences of Asparagine [N] to Cysteine [C] substitutions, the site is dominated by Asparagine and Serine [S] residues (80%), both hydrophilic with polar uncharged side chains. The N to C exchange maintains physiochemical properties, with cysteine's ability to form disulfide bonds enhances protein structure stability.

Biogeographical distribution of RNA viruses across diverse environments

Understanding the distribution patterns, evolutionary history, infectivity and genomic sequences of specific virus types can provide valuable insights into their potential to be widespread or endemic and help establish a virus-host model[47,48]. Our previous study demonstrated that DNA viruses in urban environments exhibit a non-uniform spatial and geographical distribution[33]. Here, we sought to explore the biogeographic patterns of urban & peri-urban RNA viruses on a global scale, drawing on the extensive data from our surveys of RNA virus across 102 cities in 31 countries on four continents.

First, we investigated the viral community composition and diversity of the samples from different countries (Supplementary Fig. 6A, B), revealing distinct patterns of viral composition and diversity. The results showed that *Kitrinoviricota* and *Duplornaviricota* were the dominant RNA viruses in the samples across most countries. We

https://doi.org/10.1038/s41467-026-73605-z

Table 1 | Amino acid sites (top hits) with evidence of directional selection

Phylum	Amino acid	Site Index	Rate	Bias	Amino acid site composition	Substitutions
Duplornaviricota	I	338	0.89	9.57	A2,E2,I57,K2,L19,Q8,R2,S13,T9,V2	I → E(1)K(1)L(1)Q(1)R(1)S(2)T(1)V(1), L → I(2)V(1), S → A(1)E(1)I(1)K(1), T → A(1)Q(1)R(1)
Kitrinoviricota	F	432	0.72	7.62	A19,D6,F42,G187,H2,L2,R1,S14,V1,W8	A → D(1)F(1)S(1), F → A(1)L(1), G → A(4)D(1)F(1)H(2)L(1)R(1)S(4)V(1)W(2), S → A(1), W → F(2)
Kitrinoviricota	G	338	0.75	5.34	C4,D74,G20,H5,I1,L2,M1, N36,P2,Q2,S63,T66,V5	C → G(1), D → G(1)H(1)N(1), N → D(4)G(4)M(1)T(2), P → D(1), S → C(1)D(1)G(3)L(2)N(2)P(1)Q(2)T(5), T → D(1)G(1)I(1)S(4)V(5)
Kitrinoviricota	G	546	1.01	3.21	A80,C23,F10,G22,H1,I42,K1,L52,M16,P4,Q5,R3,S3,T2,V16,Y1	A → C(4)G(12)H(1)I(5)K(1)L(1)M(7)P(1)Q(2)R(1)T(2)V(5)Y(1), C → L(1)M(1)V(1), I → L(1)V(1), L → A(2)C(1)F(2)I(3)V(1), M → F(1)Q(1)S(1)V(1), V → C(1)F(1)I(2)P(1)
Kitrinoviricota	F	832	0.69	3.03	F251,I5,K1,L1,M1,P3,S1,Y5	F → I(1)L(1)P(1)Y(3), I → M(1), P → K(1)S(1)
Kitrinoviricota	G	557	1.22	2.84	A6,C35,D1,E4,F19,G16,H1,I7,L7,M1,N2,Q8,R7,S60,T40, V51,Y13	C → A(1)F(2)L(2)T(1)V(2), I → L(1), L → I(1), Q → A(1)D(1)R(1), S → A(2)C(2)E(1)F(1)G(11)H(1)N(1)Q(2)R(3)T(3)Y(1), T → A(1)C(2)F(1)L(1)Q(1)R(1)S(4)V(3)Y(1), V → A(1)C(1)E(1)F(3)G(1)I(4)L(2)M(1)N(1)S(1)T(1)Y(3)
Lenarviricota	Y	917*	0.15	47.61	C1,F98,H1,K5,L5,N229,Y207	F → Y(4), Y → C(1)F(24)H(1)K(1)L(1)N(1)
Lenarviricota	N	100	1.02	46.16	E1,F1,N57,P1,R1,S21	N → E(1)F(1)P(1)R(1)S(6), S → N(1)
Lenarviricota	F	917*	0.08	45.61	C1,F98,H1,K5,L5,N229,Y207	F → Y(4), Y → C(1)F(24)H(1)K(1)L(1)N(1)
Lenarviricota	F	136	0.12	41.48	F2,K129,P1	K → F(2)P(1)
Lenarviricota	D	203	0.92	36.18	D97,G1,K2,P20,R2,S14,T1	D → G(1)R(1)S(9)T(1), P → D(2)K(2)R(1)
Negarnaviricota	C	212*	0.07	46.62	C69,G17,L4,N221,S120	L → N(1), N → C(10), S → C(2)G(1)L(1)N(3)
Negarnaviricota	C	893	0.07	42.63	C2,D19,E208,K137,S20,T2	D → E(2), E → D(1), K → E(1)S(1)T(1), S → C(2)E(1)T(1)
Negarnaviricota	N	212*	0.14	40.97	C69,G17,L4,N221,S120	L → N(1), N → C(10), S → C(2)G(1)L(1)N(3)
Negarnaviricota	P	97	1.01	30.66	A1,G12,P42,S1	G → P(1), P → A(1)G(1)S(1)
Negarnaviricota	D	60	0.70	30.37	D36,I1,K3,L16,V1	D → K(1)L(1), K → V(1), L → D(1)I(1)
Pisuviricota	W	97	0.54	13.17	F5,G3,H2,L2,V1,W21,Y138	Y → F(5)G(1)H(2)L(2)V(1)W(9)
Pisuviricota	W	353	0.30	8.18	C69,F321,M1,Q3,W38,Y36	C → W(1), F → C(2)M(1)Q(1)W(6)Y(22), W → F(2)
urv.p.001	W	1203	1.07	24.82	A1,E4,F3,I10,L118,T1,V5,W5	L → A(1)E(1)F(2)I(8)T(1)V(1)W(4)
urv.p.001	D	48	0.59	8.89	A5,C1,D22,E11,G39,K17,L2,M3,N1,P2,R377,S3,T11,V1,Y1	D → G(1), E → D(2)K(1)S(1)T(1), G → C(1)S(1), R → A(2)D(10)E(4)G(4)K(15)L(2)M(3)N(1)P(2)S(1)T(6)V(1)Y(1)
urv.p.001	W	466	0.7	6.41	A3,C5,D1,E6,F4,G8,H15,I3,K23,L2,M2,N28,P3,Q9,R482,S2,T8,V14,W9,Y9	E → A(1), G → T(1), H → K(1)R(1), K → H(2), N → C(1)F(1)Q(1)R(1)S(1), Q → T(1), R → C(3)D(1)E(3)F(3)G(3)H(3)I(3)K(16)L(2)M(2)N(11)P(1)Q(5)S(1)T(4)V(8)W(7)Y(5)
urv.p.001	D	1137	0.4	4.58	A1,D68,E175,F2,G1,H1,K9,Q9,T3,V4,W1	D → E(1)H(1)K(1)Q(1), E → A(1)D(15)F(2)G(1)K(4)Q(4)T(3)V(4)W(1)
urv.p.001	D	781	2.81	2.81	A1,C3,D128,E13,F1,G50,H11,I9,K45,L7,M4,N301,Q8,R12,S17,T12,V1,Y2	D → C(1)E(2)G(1)H(1)L(1)N(3)Q(1)R(1)S(3), E → A(1)K(1)Q(1)S(1), G → D(1)R(1)T(1)Y(1), H → L(1), I → V(1), K → I(1)M(1)R(4)S(1), L → I(1), N → C(2)D(47)E(5)F(1)G(14)H(7)I(4)K(17)L(4)M(1)Q(1)R(4)S(7)T(9)Y(1), Q → S(1)
urv.p.002	A	251	1.18	9.81	A37,C1,E1,F1,I1,L13,M1,N1,V3	A → F(1)L(4)M(1)N(1)V(1), L → E(1), V → C(1)I(1)

This table shows the top 5 amino acid sites with evidence of directional selection in each phylum, based on the bias parameter which suggests a significant increase in substitutions toward a specific amino acid at that site. Only *Pisuviricota* had less than 5 hits, in which case we only show as many sites as we have available. Table columns are as follows - phylum: the test clade, amino acid: the particular target amino acid with evidence of directional selection at that site, site index: the amino acid position in the alignment, rate: mean posterior relative rate at a site, bias: mean posterior bias parameter at a site, composition: the amino acid makeup of that site showing all amino acids present, substitutions: the composition of directional amino acid changes at a site e.g., A- > D (1), with the number of occurrences in parentheses. Note (*) that some sites repeat due to directional selection for multiple amino acids at that site.

speculated that the differences in RNA viral community composition might correlate with the urban locales where the samples were collected from. For example, all French samples were taken from greenhouses, most of the Korean samples were collected from hospitals (124 of 229) and public transit stations (93 of 229), and all Danish and 97.80% of the Polish samples were from subway or bus stations. The Swiss samples included 60 collected from train stations, while the environmental details for the remaining 61 samples were missing. To test this hypothesis, we investigated the dispersion of RNA viral diversity and composition across various types of environments. We first conducted an alpha diversity analysis to assess RNA viruses' diversity in different environments (Supplementary Fig. 6C). The results showed that sediment and soil samples exhibited the highest Shannon diversity index (SDI), consistent with previous studies[49,50]. Additionally, wastewater exhibited one of the highest viral diversities, indicating a potential health hazard through waterborne and aerosolized wastewater pathways. This finding highlights the importance of implementing a risk assessment and management framework for specific pathogenic viruses through wastewater surveillance[51]. Generally, samples collected from urban environments (i.e., streets, grassland, banks, stations and stores) displayed a lower RNA viral diversity compared to those gathered from the natural environments. These results suggest that the unique selection and survival characteristics of urban environments, coupled with human activities, may influence and alter the composition of RNA viruses.

We further examined the dispersion of RNA viral composition across various environments (Fig. 5A). Our findings revealed that the relative abundance of *Kitrinoviricota* is higher in terrestrial environments, such as greenhouses, banks, and public transit stations, compared to aquatic environments, such as freshwater, biofilm, and sediment. Notably, the high relative abundance of *Kitrinoviricota* in wastewater samples suggests a connection between terrestrial and natural aquatic environments. Wastewater systems, by their nature, receive inputs from both terrestrial (human and agricultural runoff) and aquatic sources. Given that *Kitrinoviricota* predominantly infects terrestrial plants and animals, its presence in wastewater suggests a transport pathway from terrestrial sources to aquatic environments[39,52,53]. A previous study unexpectedly found a considerable number of virome RdRP sequences classified under *Kitrinoviricota* in water samples from the vicinity of Yangshan Deep-Water Harbor near Shanghai[40]. Our findings suggests that the original observation at Yangshan Deep-Water Harbor could potentially be attributed to human-derived pollution, despite the general expectation of finding more *Kitrinoviricota* in terrestrial environments. Notably, *Duplornaviricota* demonstrated a striking predominance in samples obtained from environments closely linked to human activities, particularly hospitals (52.01%). Meanwhile, *Pisuviricota*, which include numerous vertebrate viruses (e.g., Berne virus, Scotophilus bat coronavirus 512 and BtRs-BetaCoV/HuB2013), were observed to dominate in certain natural environments, such as biofilm (54.29%), freshwater (43.07%), sediment (32.92%) and greenhouses (32.81%). Additionally, *Pisuviricota* were also predominantly present in the animal samples (40.37%).

To further evaluate the differences in viral communities, we performed nonmetric multidimensional scaling (NMDS) analysis based on the Bray-Curtis dissimilarity to assess the similarity in RNA viral composition across various environments (Fig. 5B). To assess the statistical significance of the observed patterns, a permutational multivariate analysis of variance (PERMANOVA) test was performed. In general, the analysis revealed a statistically significant distinction in RNA viral composition among environments (ANOSIM statistic R: 0.561, $p < 0.001$). Furthermore, a comparative assessment of dissimilarity ranks within each environment was conducted. Notably, samples from environments with substantial human influences exhibited the lowest dissimilarity, suggesting a heightened homogeneity within the RNA viral communities (Fig. 5C). This trend indicates a possible homogenization effect of viral communities as a consequence of anthropogenic influences, potentially diminishing biodiversity and ecological distinctiveness within these ecosystems.

Furthermore, we quantified the dissimilarity among pairs of environments, demonstrating that urban, non-urban, and aquatic environments each form distinct groups (Fig. 5D). Notably, animal and wastewater environments were identified as separate entities. Upon identifying the core viromes for the three main groups (see Methods and Materials), we found that viruses hosted by fungi, plants, and bacteria are the predominant constituents, comprising 82%, 76.7%, and 65.7% of the core viromes in urban, non-urban terrestrial, and aquatic environments, respectively. In non-urban environments, we observed a Rotavirus-related RvOTU targeting vertebrates as a significant component of the terrestrial core virome. Moreover, we observed the SARS-CoV-2 in the urban environment, likely due to the collection of most samples during the pandemic. These observations further highlight the necessity for environmental virus surveillance, which is critical for identifying and mitigating the risks associated with these potentially pathogenic RNA viruses. These results also indicated that the soil environment acts as the largest connecting role among different environments.

Host spectrum of UPVAtlas

The virus-host connection range provides valuable insights into viral ecology, transmission, and pathogenesis, which ultimately can inform the development of effective treatment and prevention strategies[54]. To identify putative hosts for the RNA viruses uncovered in this study, we assessed the similarity between the RdRP sequences of each representative in the RCR95 set and those of viruses with documented host associations in the ICTV database[38,55] (see Methods and Materials).

Only 33.54% of RCR95 representatives (17,746 out of 52,914) were assigned to hosts, limiting our ability to fully understand the diversity of potential hosts. Previous studies have suggested that RNA viruses primarily target eukaryotic hosts[14,56]. However, our findings revealed that more than half of RNA viruses assigned host information were likely bacteriophages (50.33% of the uncovered RNA viruses with host information) (Fig. 6A). The observed discrepancy between our findings and previous studies likely stems from the broader scope of host assignments achieved in our study compared to the more limited focus of prior research. A substantial percentage of *Lenarviricota* viruses were associated with bacteria (69.41%), *Kitrinoviricota* with plants (85.81%), and *Duplornaviricota* with fungi (90.05%). Among the RNA viruses that have the potential to infect vertebrates, *Pisuviricota* viruses, which encompass notable viruses like SARS-CoV-2, Rhinoviruses, Polioviruses, and Enteroviruses, exhibited the highest prevalence (95.30%). To determine the specific lineages of vertebrates infected by RNA viruses, we mapped the related viruses to the VIRION database[57], which was curated by the Verena Consortium, an interdisciplinary group focused on zoonotic disease prediction. This database supports proactive measures by predicting human-infecting viruses as well as identifying their host animals, and potential emergence locations. Our analysis revealed 60 viruses with a high likelihood of infecting humans, along with 47 viruses likely infecting other mammals (Fig. 6A). Of these, 19 out of 60 human-associated RCR95 representatives are found in urban environments, and 34 of them are found in animal samples; 1 out of 47 mammalian-associated RCR95 representatives are found in urban environments, and 36 of them are found in animal samples.

To further investigate the prokaryotic hosts of the UPVAtlas members, we aligned the 29,857,318 CRISPR spacers obtained from the IMG database[58,59] with all the contigs corresponding to the RCR95 representatives (see Methods and Materials). This process yielded 9466 matches that included 946 RCR95 representatives and 2446 spacers, with the latter originating from 423 bacterial and 28 archaeal hosts (Fig. 6B, Supplementary Data 7). The results revealed that the

https://doi.org/10.1038/s41467-026-73605-z

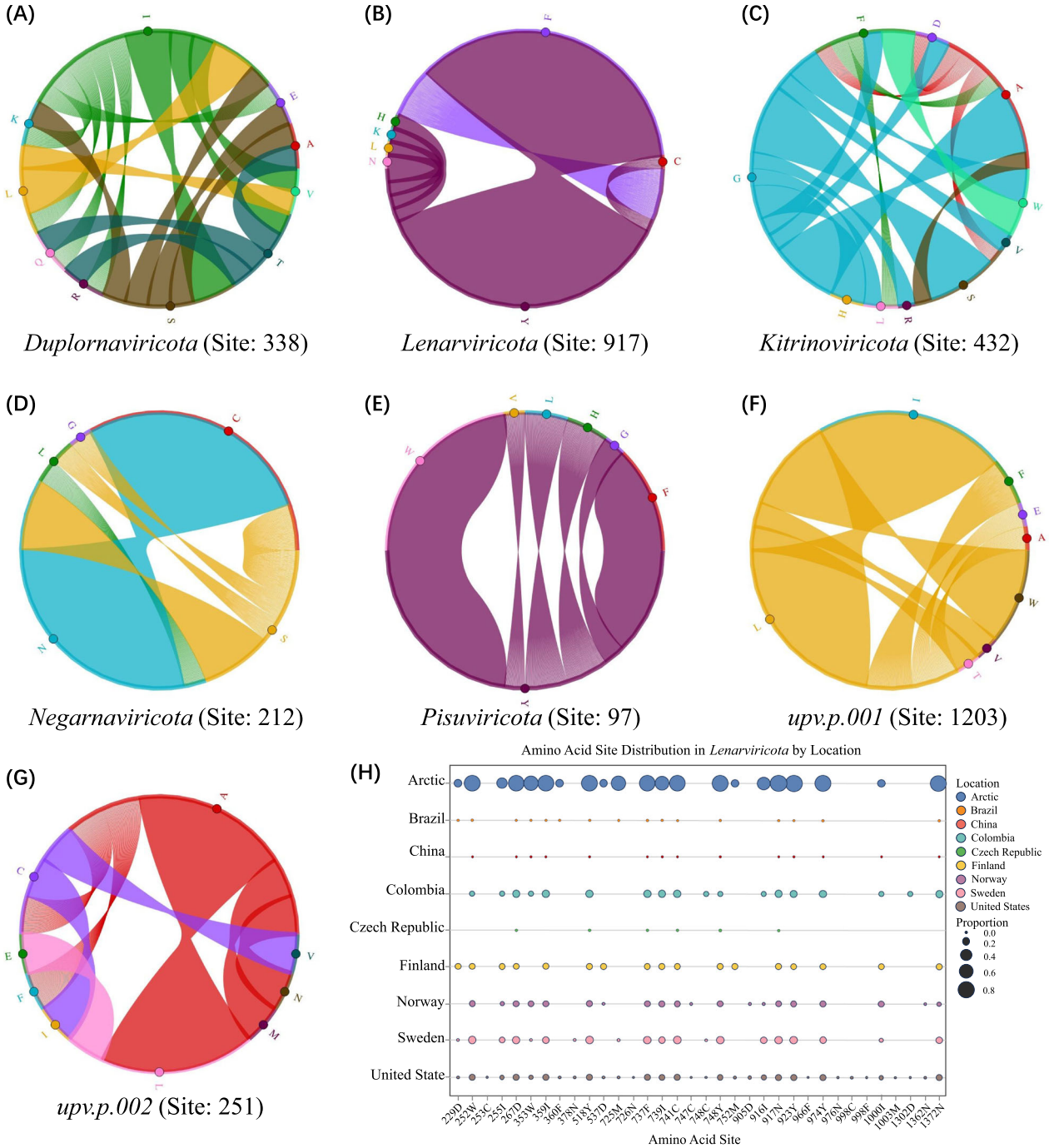

Fig. 4 | Exchanges of amino acid sites under directional selection and their global distribution. A–G Composition and exchange of amino acids at specific top hits for each phylum. The chord diagrams are colored by the respective "To" amino acid on the inner ring and is colored for each amino acid in the outer ring. The thickness of the chord indicates the amount of exchange between two amino acid residues. Sites show a difference in the diversity and volume of exchanges at a particular site, indicative of types of selective pressures at that site such as selective sweeps, consensus replacement sites, repeated substitutions, rare residue, or highly polymorphic sites. **H** Examining the global distribution of amino acid sites under directional selection in *Lenarviricota*. For each site with evidence of directional selection (Table 1) in *Lenarviricota* we examine the global distribution of the target amino acid. The proportion of samples with the target amino acid of each location is normalized by the total number of samples from each country, to avoid bias from oversampled areas.

vast majority of RCR95 representatives are highly specific, with nearly two-thirds showing infection histories exclusive to a single host, and 81.08% infect no more than three species. Furthermore, the discovery of 133 RCR95 representatives hosted by more than 10 species, mostly belonging to the *Pisuviricota* phylum, indicated a potential interaction or co-evolution between these viruses and their target species. Notably, we identified 166 RNA viruses that showed infection histories with four of the ESKAPE pathogens, specifically *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii*. Among these viruses, 60.84% were classified as *Pisuviricota*, and 18.07% were classified as *Kitrinoviricota* (Fig. 6C). Only one virus showed matches with *Acinetobacter baumannii*, while nearly one hundred viruses had infection histories with *Staphylococcus aureus* (96 viruses) and *Klebsiella pneumoniae* (92 viruses), with the majority of

https://doi.org/10.1038/s41467-026-73605-z

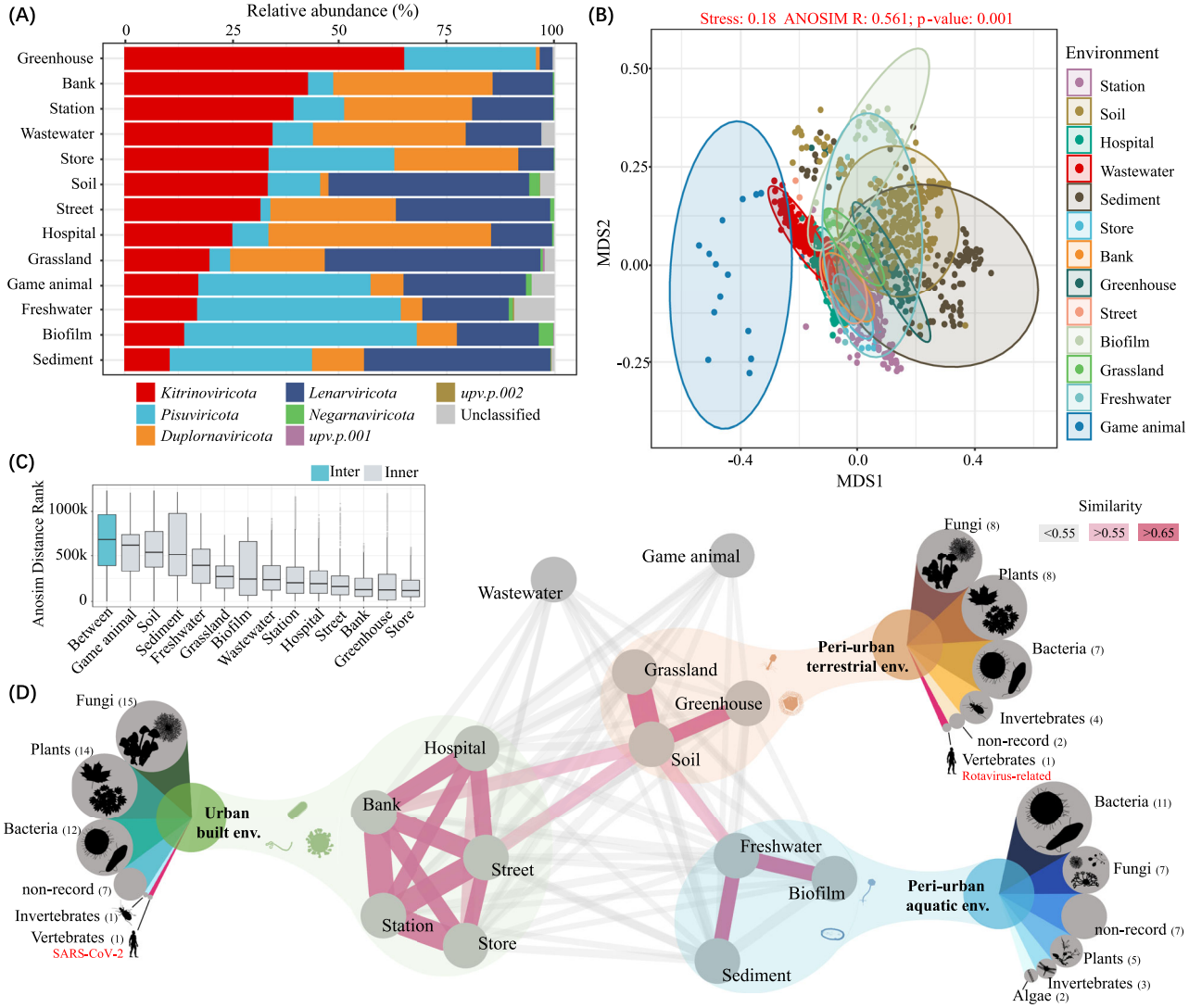

Fig. 5 | Urban & peri-urban environmental RNA viral composition and diversity. **A** Phyla composition of environmental RNA viruses for samples collected from different environments. Relative abundance was quantified as fragments per kilobase million reads (FPKM). **B** Nonmetric multidimensional scaling (NMDS) plots of RNA viral communities grouped by urban environment. The stress value of the NMDS ordination is 0.18. The 95% confidence ellipses are displayed around the sample groups. **C** Pairwise ANOSIM distance boxplot of samples in different environments (Between, $n = 1031224$; game animal, $n = 105$; soil, $n = 40755$; sediment, $n = 4656$; freshwater, $n = 153$; grassland, $n = 496$; biofilm, $n = 861$; wastewater, $n = 13041$; station, $n = 110215$; hospital, $n = 21321$; street, 1128; bank, $n = 2278$; greenhouse, $n = 1378$; store, 2485). The center bar of box indicates median, bounds represent lower and upper quartiles, and the line indicate minimum and maximum of the dataset for each group. **D** The central network shows the similarity of the RNA viral communities collected from different environments. Color and width of edges indicate the similarity between each pair of environments was measured by ANOSIM based on the RNA viral abundance. The three sector plots around illustrate the proportion of members aiming different hosts within the core-virome of the three clusters (urban built env. cluster, peri-urban terrestrial env. cluster, and peri-urban aquatic env. cluster) and the numbers represent the quantity of RvOTUs. The core-virome of a cluster is defined as the intersection of RvOTU members that exhibits a prevalence greater than 10% in each environment. Organism silhouettes are sourced from PhyloPic (phylopic.org). Acer silhouette by T. Michael Keesey, published under a Creative Commons Attribution-ShareAlike 3.0 Unported license (CC BY-SA 3.0, https://creativecommons.org/licenses/by-sa/3.0/); the silhouette has been resized and rotated; this work is distributed under the same license as the original. Aralia silhouette by Tree of Life App, published under a Creative Commons Attribution 4.0 International license (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/); the silhouette has been resized and rotated. Proso millet silhouette by Stefan Lefnaer and T. Michael Keesey, published under a Creative Commons Attribution 4.0 International license (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/); the silhouette has been resized and rotated. Clostridium silhouette by Gareth Monger, published under a Creative Commons Attribution 3.0 Unported license (CC BY 3.0, https://creativecommons.org/licenses/by/3.0/); the silhouette has been resized and rotated. Spirogyra silhouette by Matthew Crook, published under a Creative Commons Attribution-ShareAlike 3.0 Unported license (CC BY-SA 3.0, https://creativecommons.org/licenses/by-sa/3.0/); the silhouette has been resized and rotated; this work is distributed under the same license as the original. The other silhouettes are all licensed under CC0.

these viruses belonging to the *Pisuviricota* phylum. These results underline the potential of RNA viruses as platforms for developing therapeutic strategies like phage therapy to combat drug-resistant bacterial infections.

We also compared 1.8 million CRISPR spacers extracted from 286,997 genomes in UHGG database[60] with RCR95 representatives, to explore the association between urban & peri-urban RNA virome and human health (see Methods and Materials). This resulted in 793 potential host-virus interactions, encompassing 1.15% (606 out of 52,736) of RCR95 representatives and 107 human gut microbes, which include 106 bacteria and one archaeon (*Methanobrevibacter smithii*) (Fig. 6B, Supplementary Data 8). The majority of the human gut

https://doi.org/10.1038/s41467-026-73605-z

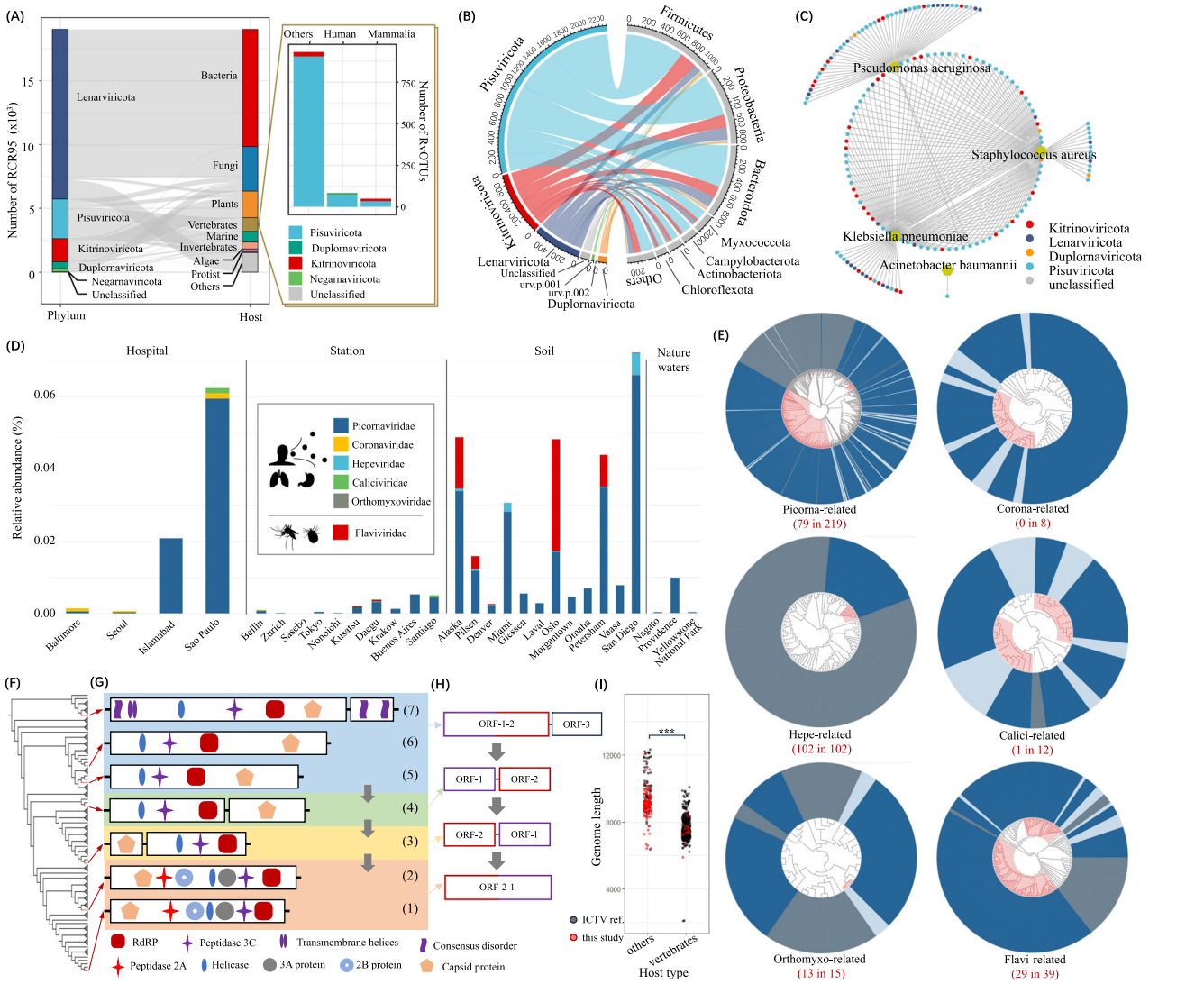

Fig. 6 | Host associations and vertebrate RNA viruses in UPVAtlas. A Sankey diagram displays the distribution of uncovered RNA viruses and their hosts. Only RNA viruses that matched to viruses with ICTV host information were included. The subplot shows the phyla of RNA viruses potentially hosted by vertebrates. **B** Chord diagram shows the targeting interaction between viruses and their corresponding host organisms, as determined by matching the spacers in the IMG database to the viral contigs. The color of the ribbons represents the virus phylum. **C** Network of potential interactions between RNA viruses and ESKAPE pathogens based on spacer to protospacer matches. ESKAPE pathogens were represented by citrine hexagons and labeled. Viruses are colored by phylum. **D** Relative abundance of six major vertebrate pathogenic families across key environmental types in different cities. In the legend, families are categorized by transmission routes, including gastro-intestinal/respiratory and vector borne. Human and insect silhouettes are sourced from PhyloPic (phylopic.org) and are all licensed under CC0 or PDM 1.0 (https://creativecommons.org/publicdomain/mark/1.0/). **E** Phylogenetic tree constructed using core motif RdRP domain sequences of reference genomes and RCR95 representatives. Reference viruses are shown in dark blue, viruses with larger than 50% identity to the reference viruses are shown in light blue, and the potential novel RNA viruses are shown in gray. Trees are midpoint rooted. Numbers under the label indicated the number of RNA viruses with less than 50% identity to the reference viruses in all the positive matched RNA viruses. **F** Phylogenetic tree of Picorna-related RCR95 members. **G**, **H** The evolutionary progression of structures within Picorna-related viruses. Homologous domains are visually represented using identical shapes and colors, as delineated in the legend provided at the bottom. Genome structure evolutionary trajectory of Picorna-related viruses experienced three major transformations: 1. Separation of structural ORF from non-structural ORF (from 5, 6, 7 to 4); 2. Shift of structural and non-structural ORFs (from 4 to 3); 3. Emergence of peptidase 2 A and fusion of ORFs. **I** Genome length of *Picornavirales* members targeting different hosts. Genomes with completeness over 70% from this study are included. A two-sided Welch *t*-test were performed to evaluate the difference in means between groups ($p = 0.000416$). *** denote a *p*-value lesser than 0.001.

bacteria infected by the RNA viruses were *Clostridia* (52.83%) and *Bacteroidia* (17.92%), the two of the most abundant bacterial classes in the human gut[61,62]. The dominant RNA viruses that can infect human gut bacteria were *Pisuviricota* (41.09%) and *Lenarviricota* (26.90%). Using the spacers extracted from the 4644 species-level representatives in the UHGG database, only 0.05% of these RCR95 representatives were covered by the host-virus interactions, suggesting considerable CRISPR diversity between bacterial strains and active community infection.

In adherence to the "One-Health" framework, our investigation extended to the association between 1477 World Development Indicators[63] (WDIs), reflective of a country's status in terms of public health, education, and economic parameters, and the abundance of RNA viruses hosted by various organisms (see Methods and Materials). Our analysis found that 219 WDIs, mainly associated with health, economy and the environment, exhibited statistically significant correlations with the abundance of RNA viruses in various hosts (Supplementary Fig. 7, Supplementary Data 9). From the results, we

observed that most of these WDI were significantly associated with the abundance of the fungal, plant and vertebrate RNA viruses, and only a few of them correlated with the bacteria and invertebrate RNA viruses. Specifically, many of the WDIs related to the topics of Economic Policy & Debt and Social Protection & Labor were significantly and positively correlated with the abundance of vertebrate RNA viruses (e.g., Population of all age groups, Unemployment with basic education, etc.). Furthermore, the negative correlation ($\rho = -0.5$, $P = 0.02$) observed between "Rural population growth (annual %)" and vertebrate viral abundance suggests additional viral risks associated with urbanization and population density. Overall, urban areas in less developed countries were found to carry a higher burden of vertebrate viruses.

To examine the diversity and phylogenetic expansion of several major vertebrate pathogenic families, we retrieved 396 RCR95 members belonging to six families. Hospital samples from Islamabad and São Paulo exhibited substantially higher abundances of vertebrate pathogens compared with those from Baltimore and Seoul, while station samples from Buenos Aires and Santiago displayed the highest vertebrate-pathogen levels, exceeding those observed in developed countries in Europe and East Asia (Fig. 6D). As expected, *Picornaviridae* —one of the largest families of human and animal viral pathogens[64]— dominated nearly all urban and peri-urban environments across cities, whereas vector-borne *Flaviviridae* members were predominant only in a subset of soil samples (Fig. 6D). *Coronaviridae* were detected in hospital environments in multiple cities, which is unsurprising given our sampling period: the COVID-19 pandemic clearly influenced the urban RNA virome composition, although not sufficiently to challenge the dominance of Family *Picornaviridae*.

We further constructed maximum likelihood phylogenetic trees for each family. These trees were composed of RCR95 representatives mentioned above and the related referenced RNA viruses (Fig. 6E). However, it is noteworthy that more than half of these RNA viruses exhibited low similarity to the known viruses (sequence identity less than 50%). For example, all the RNA viruses annotated as *Hepeviridae* showed low similarity to the known *Hepeviridae* members. This underscores the potential presence of novel and divergent strains within this viral family.

Moreover, protein domains within the genomes of RNA viruses were discerned, revealing substantial genetic diversity across distinct genera and species within the same viral family, thereby underscoring the pivotal role of this diversity in fostering the adaptation and evolution of these viruses (Fig. 6G, Supplementary Fig. 8). As the largest family of vertebrate-infecting RNA viruses, *Picornaviridae* encompasses numerous important human pathogens (e.g., Poliovirus, Echovirus, Hepatitis A virus, etc.). Upon investigation of *Picornaviridae*-related viruses' genomes, three discernible structural arrangements emerged, the initial one closely mirroring the extensively documented structural organization of *Picornaviridae* viruses archived in the ICTV database. In this arrangement, peptidase 2 A and 3 C are strategically positioned upstream and downstream of the helicase, respectively. However, an intriguing finding involves the absence of peptidase 2 A in the alternate structural arrangements. Considering the pivotal roles played by peptidase 2 A and 3 C in the cleavage of the viral precursor polyprotein, and their contribution to the characteristic protein arrangement[65], coupled with insights from the phylogenetic relationships among *Picornaviridae*-related viruses, we can infer potential evolutionary trajectories in the genome structure of *Picornaviridae*-related viruses. In the initial phases of phylogenetic development, the original ORF was cleaved, producing one structural and one non-structural protein (From 5,6,7 to 4 as shown in Fig. 6F–H). Subsequently, a reversal in the positions of these proteins occurred (from 4 to 3 as shown in Fig. 6F–H). Ultimately, the introduction of the 2 A protein facilitated the emergence of this merged-ORF genomic architecture. (from 3 to 1,2 as shown in Fig. 6F–H). Previous research has indicated that 2A-like peptides emerged independently during the evolutionary history of various viral families[66]. From the evolutionary history of picorna-related viruses, the appearance of the picornavirus 2 A protein may indicate potential reassortment events. In summary, this series of changes ultimately led to a compact genomic architecture characterized by one single ORF.

Takada et al. reported that vertebrate viruses in multiple families (e.g., *Flaviviridae*) exhibit more simplified genomes compared to invertebrate viruses, which may confer a selective advantage in vertebrate hosts—which possess more sophisticated adaptive immune systems—by reducing the number of potential immune targets encoded in the genome[67]. Here, we observed that *Picornaviridae*-related viruses also exhibit a trend toward simplifying their genomes throughout their evolutionary history (Fig. 6F-H). Besides, *Picornaviridae* members exhibited markedly shorter genomes compared to their non-vertebrate-infecting relatives (Fig. 6I), providing further evidence for adaptive evolution associated with vertebrate infection.

Discussion

RNA viruses are a major favor in human and environmental health, most especially in urban environments with higher population densities and more interconnected transportation networks, which can facilitate the propagation of an epidemic. In this study, we analyzed 2922 metatranscriptomic samples spanning diverse environments across 102 cities (31 countries) before, during, and after the COVID-19 pandemic, and uncovered 54,945 RvOTUs, forming a large and geographically widespread assemblage of RNA viruses, termed the Urban & Peri-urban RNA Virus Atlas (UPVAtlas), of which ~77% had not been previously reported. Our findings suggested that urban environments harbor a rich and diverse RNA virome, with highly frequent RvOTUs present in both natural and urban environments. We found that the UPVAtlas members' prevalence in these environments was distinct, likely because different factors contribute to their diversity and distribution. Nonetheless, a considerable number of RvOTUs were found to be prevalent across multiple environments, especially those closely associated with human activities. This indicates the cross-transmission of RNA viruses across these environments, highlighting their adaptability and raising concerns regarding risks of human infection.

Revising the RNA virome phylogeny

Utilizing the phylogenetic tree derived from the complete RdRP (RNA-dependent RNA polymerase) core domain sequences of the environmental RNA virome, two novel phyla were identified, with one showing a close clustering relationship to the previously identified phylum *p.0002* from Neri et al.'s work. Moreover, the polyphyletic nature of *Duplornaviricota* was further corroborated in this study, aligning with the Zayed et al.'s observation. Furthermore, the discovery of two distinct clades within *Duplornaviricota*, potentially indicative of novel subphyla or classes, has been elucidated. Besides, the polyphyly detected within *Duplopiviricetes* aligned with the situation observed in *Duplornaviricota*. Together, these findings provide further evidence for the multiple independent origins of double-stranded RNA viruses in the evolutionary history of RNA viruses and highlight potential issues inherent in taxonomic classification scheme based on genome molecular type.

Using the phylogenetic annotation, we generated a global landscape of the RNA virome revealing the biogeographical specificity of RNA viruses across diverse urban environments. Our finding highlighted the influence of human activities in forging interconnected environments that manifest a comparatively reduced viral diversity. For instance, samples collected from urban environments such as streets, banks, and subway stations demonstrate a diminished RNA viral diversity relative to those obtained from natural habitats like sediment, soil and grassland. Natural environments, exemplified by soil, offer a more stable milieu for the persistence of microorganisms compared to surfaces within the urban environment.

The high diversity observed in the RNA virome of wastewater samples supports the potential of wastewater monitoring systems, particularly in the context of their widespread implementation across countries in response to the COVID-19 pandemic. This heightened diversity underscores the capacity of such systems to detect pathogens that may be absent or present at lower levels in other environments, including hospitals. While the candidate phyla exhibited lower relative abundance compared to well-established phyla in the majority of samples, their presence was nonetheless indicated by distinctive ecosystem niches.

It is noteworthy that, although we employed an iterative RdRP search strategy and used geNomad to assist in recovering the RNA virome from our metatranscriptomes as comprehensively as possible, these methods also have limitations in capturing highly divergent RNA viruses and their RdRPs. Substantial evidence suggests that a vast number of unknown RNA viruses remain unobserved[68]. Therefore, it is not surprising that the previously established five phyla dominate from the perspective of current methods; more advanced approaches will need to be created to uncover the "hidden" RNA viruses.

The predominance of short viral contigs in the UPVAtlas, with over 92% of RvOTUs shorter than 5 kb, warrants careful interpretation. While approximately half of the RvOTUs achieved medium genome completeness (> 50%), the truncated length distribution likely reflects a combination of biological and technical factors. On one hand, many RNA virus lineages, particularly within *Lenarviricota*, genuinely possess compact genomes. Narnaviruses and related capsidless RNA elements can be as small as ~2.5 kb, encoding little more than an RdRP. Our recovery of numerous short yet apparently complete *Lenarviricota* genomes is consistent with this biology. On the other hand, the limited sequencing depth inherent to broad metatranscriptomic surveys where viral reads compete with vastly more abundant host and microbial transcripts, inevitably constrains the assembly of low-abundance viral genomes, yielding fragmentary contigs that capture the conserved RdRP domain but miss flanking structural and accessory genes. Future studies employing targeted viral enrichment strategies such as virus-like particle purification or sequence-independent single primer amplification (SISPA) combined with long-read sequencing technologies, would help distinguish genuinely compact viral genomes from assembly artifacts and enable more complete characterization of the structural and functional gene repertoire of these novel RNA viruses.

Environment specific RNA viruses pose a risk to monitoring efforts

A key limitation of this study is the presence of sample bias, which hindered our ability to directly compare the influence of different environments across countries. The uneven distribution of samples across countries and environment types created a significant confounding factor. For instance, all the French samples were collected from greenhouse, while in Poland, 97.80% of the samples were from subway or bus stations. This make it difficult to determine whether observed viral community differences are due to the specific environment or to underlying country-level factors. This bias prevents us from definitively attributing observed patterns to country of origin. Future studies should prioritize a more balanced and stratified sampling design, ensuring adequate representation of all countries and environment types of interest, to address this limitation. Despite this constraint, our current findings provide valuable insights into the influence of environment type on RNA virus diversity.

The host analysis found that most RNA viruses uncovered in this study were hosted by bacteria, fungi and plants, highlighting the role of these organisms may play in the ecology and evolution of RNA viruses in urban environments. In alignment with the "One-health" concept, our research has elucidated a significant correlation between societal and economic progression and the prevalence of vertebrate RNA viruses. These finding serve as a cautionary note regarding the impact of biodiversity on the trajectory of sustainable development. Through spacer to protospacer matching, we unveiled that a majority of the human gut bacteria potentially infected by RNA viruses belonged to the *Clostridia* and *Bacteroidia* classes, which are among the most abundant bacterial classes in the human gut[61,62]. Clostridia, constituting a polyphyletic group of Gram-positive, spore-forming anaerobes within the *Bacillota* phylum, significantly influences metabolism and the functioning of the human gastrointestinal tract[69]. Certain members of the *Bacteroidia* class play roles in bile acid metabolism and the transformation of toxic and/or mutagenic compounds[70]. Notably, we identified 166 RNA viruses showing infection histories with four ESKAPE pathogens, with most belonging to the phylum *Pisuviricota*. These findings provide a foundation for developing novel viral-based therapeutic strategies for ESKAPE pathogens[71].

The finding that more than half of the host-assigned RNA viruses in the UPVAtlas are likely bacteriophages challenges the long-held perception that RNA viruses primarily target eukaryotic hosts. Historically, only a small number of RNA phage families, most notably, the *Leviviridae* infecting, Gram-negative bacteria, were recognized, and RNA phages were considered evolutionary curiosities of limited ecological relevance. However, our data have expanded the known diversity of RNA bacteriophages, particularly within *Lenarviricota*, where nearly 70% of host assigned members were associated with bacteria. In the urban context, these findings carry practical implications: RNA phages infecting ESKAPE pathogens may represent an untapped reservoir for phage therapy development, particularly given the growing crisis of antimicrobial resistance in hospital and transit environments where these pathogens are most prevalent. Additionally, the detection of RNA phages targeting *Clostridioides difficile* across multiple members of the candidate phylum *upv.p.001* suggests that these viruses may play an unrecognized role in modulating pathogen populations within urban wastewater and hospital settings, environments where *C. difficile* transmission is a persistent clinical concern.

Molecular profiling using natural selection analysis

Evidence for directional selection at amino acid sites can broadly be categorized based on their impact on protein function or fitness, including beneficial sites leading to adaptive changes, deleterious sites associated with reduced fitness, and neutral sites with no apparent effect on fitness. Directional selection at specific amino acid sites within RdRp may provide insight into the fundamental evolutionary mechanisms that significantly influence the adaptability and virulence of RNA viruses: including SARS-CoV-2 and Influenza. These amino acid changes, often resulting from mutations that confer selective advantages such as enhanced replication fidelity, drug resistance, or evasion from host immune responses, establish crucial molecular signatures within the RdRp gene[72]. Such molecular signatures are vital for tracking viral evolution, understanding pathogen adaptability, and identifying potential therapeutic targets[73]. Additionally, the identification of these molecular markers through global molecular surveillance programs significantly enhances public health responses by facilitating the early detection of potentially pandemic strains, guiding vaccination strategies, and informing global health preparedness and response strategies.

In our investigation of the global distribution of amino acid sites that showed signs of directional selection across all phyla (Fig. 4H), we observed a disproportionate number of sites with high values in the Arctic. However, it should be noted that a significant portion of our dataset, specifically 25%, consists of samples gathered from the United States, while over 10% originate from China. To mitigate potential analytical bias, we normalized the sequence counts of the target amino acids by the total number of samples from each country. Within the phylum *Lenarviricota*, we identified amino acid sites such as 998 C,

998 F, 1003 M, 1302D, and 1362 N that exhibit a relatively low global distribution, contrasting with sites like 252 W, 267D, 353 W, 518Y, and 917 N, which showed a significant global presence. These observations underscore the role of evolutionary pressures in shaping viral fitness and pathogenicity, thereby enriching our understanding of viral adaptability. A comprehensive understanding of how the distribution of amino acid sites affects viral dynamics requires consideration of multiple factors, including temperature, which profoundly influences viral replication rates and stability, and geographical climate shifts. Human activities, such as urbanization, deforestation, and agricultural activities, particularly livestock farming, can also create conditions conducive to the spread of infectious diseases[25,26,74,75].

Our current directional selection analysis, while limited to the available sampling period, exposed significant amino-acid site behavior in terms of composition and substitution diversity, aiming to identify biological covariates in RdRp that influence replication fidelity, immune response, and host factor interaction. The directional selection analysis was conducted to minimize bias towards any oversampled phylum, opting against a normalization approach. Future serial sampling across cities is expected to explore amino acid site behavior under different evolutionary scenarios, offering deeper insights into temporal signals including selective sweeps and frequency-dependent selection.

Our study has several limitations. Our approach involves identifying RNA viruses through metatranscriptome analysis, which is susceptible to artifacts, such as RT errors or chimeric RNA assemblies. Consequently, we focused on evolutionarily conserved groups of RNA virus sequences, rather than singletons. Additionally, we recognize the significance of having virtually complete viral genomes to comprehend viral genome structure and function. However, due to the limitation in sequencing depth, our dataset contains only a limited number of complete or near-complete RNA genomes, hindering further functional analysis of the identified RNA viruses. Limitations related to our directional selection analysis are also important to note. Alternatives to directional selection, such as the confounding effects of diversifying selection, must be considered when interpreting our results. These can be partially elucidated by examining stabilizing selection, which curtails genetic variation to preserve a population's fitness; balancing selection, which maintains multiple alleles at frequencies exceeding those expected by random chance; and neutral evolution, characterized by genomic changes occurring at a steady rate without impacting the organism's fitness. Collectively, these evolutionary mechanisms contribute to the genetic diversity and evolution of viruses, shaping their adaptability, virulence, and resistance to therapeutic interventions. Consequently, these factors may circumscribe the interpretability of our analytical outcomes. Furthermore, despite the substantial effort dedicated to host analysis, our current knowledge can only attribute proper hosts to about half of the identified RNA viruses. Also, due to the lack of a widely accepted gold standard for host inference of viruses obtained from metagenomes, the broad host groupings (bacteria, plants, vertebrates, invertebrates) we applied also prevent us from revealing more nuanced relationships between viruses' relative abundance and meteorological conditions. Additionally, the rapid and frequent mutations characteristic of RNA viruses, coupled with variations in host range among closely related species[76], caution against relying solely on sequence similarity for host predictions. This highlights the potential for misinterpretations and underscores the critical necessity for developing and implementing more sophisticated tools in virus-host prediction, aiming to enrich our comprehension of these intricate interactions.

A potential limitation of the phylum-level relative abundance comparison in this study lies in the use of aggregated relative abundance, which may introduce biases in interpreting the ecological dominance of viral taxon. First, taxonomic diversity varies substantially across viral phyla, with certain phyla (*Lenarviricota*, *Pisuviricota*) harboring a far larger number of RvOTUs than others. This uneven distribution of RvOTU richness could inflate the aggregated relative abundance of more diverse phyla; thus, the observed higher abundance of these phyla may not fully reflect their genuine ecological dominance, but rather a statistical artifact derived from greater taxonomic representation. Second, for a phylogenetically broad taxa group whose members span a wide host range—from prokaryotes to humans (e.g., *Pisuviricota*)—the use of aggregated abundance to assess its ecological dominance masks the distinct ecological functions and host specificities of individual OTUs. Therefore, calculating the average abundance of OTUs per phylum, as an alternative metric, might provide a more ecologically relevant perspective for cross-phylum comparisons, and this approach could be validated in future studies. Besides, the viral relative abundance inferred from metatranscriptomic data represents a mix of viral particle abundance and viral transcription activity, rather than pure viral biomass. Further, viral phyla differ in genome polarity (e.g., dsRNA, ssRNA + , ssRNA- viruses), and these differences are associated with distinct transcriptional efficiencies and RNA stability in host cells. Such inherent variations further complicate the direct comparison of relative abundance across taxa, as the detected signals may not be fully comparable in their biological implications.

Overall, our UPVAtlas constitutes the largest assemblage of RNA viruses from metatranscriptomics data to date, and provides important insights into the diversity, distribution, and host range of RNA viruses in urban environments across more than 100 cities worldwide. These findings hold potential implications for public health, as the extensive phylogenetic diversity of pathogen-related viruses and their broad environmental distribution suggest possible risk associated with human interaction in these environments. Identifying potential candidate phyla and discovering unknown RNA viruses hosted by ESKAPE pathogens provide new insights into the mechanisms underlying the virulence and pathogenicity of these bacteria, which could inform the development of new treatment strategies. The expansion of the phylogenetic diversity within pathogen-related viral families also emphasizes the need for continued surveillance and monitoring of urban environments to identify and characterize previously undiscribed viruses that may threaten public health. The MetaSUB consortium and its analyses of microbiomes and metagenomes in urban environments, provide a rich dataset for understanding the diversity, distribution, and evolution of viruses in cities. These types of data, combined with viral sequences from publicly available databases may be useful for creating molecular signatures of viruses and their local adaptations. This involves identifying unique genetic sequences or patterns that distinguish one virus from another or one strain from its counterparts in different global regions. Among other things, the MetaSUB consortium seeks to highlight how viruses have evolved in response to environmental pressures, host immune systems, and treatments. These molecular signatures (which leverage our existing toolkit of selection analyses) can then inform vaccine development, tracking of viral spread, and understanding of how viruses adapt to different hosts or geographical areas. Our study contributes towards a crucial aspect of modern molecular epidemiology, which may help to predict and control viral outbreaks by understanding viral evolution and dynamics in urban environments and contributes to our knowledge and management of viral diseases.

Methods

Sample collection and processing

All MetaSUB samples were processed under a standard operating procedure (SOP) as our previous study[33]. Samples were collected using Isohelix RNA buccal swabs, ensuring standardized procedures across diverse cities and sampling teams. Prior to extraction, each sample was preserved in a tube containing 400 µl of RNA Shield (Zymo Research). Notably, samples stored in the DNA/RNA Shield can be frozen at

temperatures of −20 °C to −80 °C for extended periods, while nucleic acids remain stable at ambient temperature (RNA up to 1 month). The collection process adhered strictly to the SOP developed collaboratively by all MetaSUB consortium members[33,77], aiming to maintain high-quality control throughout the sampling procedure.

Subsequently, the swabs were extracted using the easyMag instrument (bioMérieux), following the manufacturer's instructions. The input volume for extraction was 300 µl, with an elution volume of 110 µl. For library preparation, the total nucleic acid (TNA) from all samples underwent treatment with DNAse I (Zymo Research, Catalog # E1010), which effectively cleaves both double-stranded and single-stranded DNA. Post-DNAse digestion, the samples were subjected to the NEBNext rRNA depletion v2 (Human/Mouse/Rat) protocol, utilizing Ultra II Directional RNA (10 ng) and Unique Dual Index Primer Pairs from New England Biolabs. All kits were sourced from a single manufacturer lot.

Completed libraries were quantified using Qubit and analyzed for size determination on a Bioanalyzer (Agilent). These libraries were subsequently pooled and dispatched to either the WCM Genomics Core or HudsonAlpha for final quantification using a Qubit fluorometer (ThermoFisher Scientific), TapeStation 2200 (Agilent), and qRT-PCR with the Kapa Biosystems Illumina library quantification kit. The NovaSeq 6000 (Illumina) facilitated 150-base pair paired-end sequencing at the UCSF Center for Advanced Technology. Additionally, a Human Reference RNA Standard (Agilent) was incorporated for quality assurance.

Data preprocessing, assembly and establishment of RNA virus operational taxonomic units (RvOTU)

A total of 2922 metatranscriptomic samples were collected, including 1863 samples collected by the MetaSUB Consortium[36] during the COVID-19 pandemic, and 1059 samples from peri-urban environment were collected from the Sequence Read Archive (SRA) database by keyword searching (e.g., environmental metatranscriptome, city metatranscriptome and <*city name*> metatranscriptome, etc.). The metadata of these samples was also downloaded and manually curated according to the sample descriptions.

The raw sequencing data were initially processed with Trimmomatic[78] (version 0.39) with default parameters to remove low-quality reads and those with ambiguous bases. The quality-controlled reads were then aligned to the human genome (GRCh38, including alternate contigs) using Bowtie2[79] (version 2.4.4) with default setting to remove the host-related sequences. No filtering for low-complexity reads was performed prior to the assembling step. Subsequently, for each sample, reads were de novo assembled into contigs using MEGAHIT[80] (version 1.2.9) with default parameters.

Contigs with length no shorter than 1000 nt were first screened for viral contigs using geNomad[41] (version 1.6.1) end-to-end command with default parameters. Contigs annotated as Orthornaviran by geNomad were further clustered into species level units using MMseq2[41] (release 13-45111) with parameters "--min-seq-id 0.9 -c 0.9 --cluster-mode 2 --cov-mode 1 --seq-id-mode 2". And then clusters with RdRPs detected in the "Identification of RNA-dependent RNA polymerases (RdRPs)" section were defined as RvOTUs.

Identification of potential strain contamination

Negative controls of MetaSUB samples were processed in the same way as viral contigs detection. Then, we obtained only five viral contigs. These contigs did not produce blastn matches meeting the contamination threshold[33] (identity > 99.99%) with our RvOTUs and, therefore, no RvOTU was excluded as contamination.

Identification of RNA-dependent RNA polymerases (RdRPs)

To maximize the retrieval of RdRP sequences, we implemented an iterative search approach similar to that used in Neri et al.'s study. Briefly, a comprehensive RdRP database was constructed by keywords searching (RdRP, RNA-directed RNA polymerase, RNA-Dependent RNA Polymerase) in the NCBI nr database, along with RdRPs provided by Neri et al.[14]. This database was then used as queries to search a 6-frame end-to-end viral-contig-translation BLAST database, which was generated using Transeq (EMBOSS version 6.6.0). Translations from RNA viral contigs (mainly *Orthornavirae* contigs, and a set of *Riboviria* contigs, which only classified as *Riboviria* and have no kingdom-level annotation, as a supplement in case of omission) that exhibited at least one hit with an alignment length over 100nt were retained as RdRP-drafts for subsequent inspections and filtrations. After that, we detected the RdRP motif A-D against the RdRP-drafts using hmmsearch (HMMER, version 3.3) based on the published profile HMMs of RdRPs[39,40]. Considering the variability of motifs in some viruses[81], only RdRP drafts with at least three of the motifs (A–D) were considered reliable. These reliable RdRP drafts were subsequently trimmed to approximate the core domain, referred to as the core-ref.

All RdRP sequences in core-ref were used as queries to search contig translations again for deep exploration of RdRP. Acceptance criteria for this round were: profile coverage ≥ 75%, E-value ≤ 0.00001 and identity ≥ 0.45 (0.15, 0.30, 0.45, 0.60, 0.75 were tested to determine the optimal value for distinguishing true RdRPs from group II introns; see "False positive RdRP filtering process" for details). Then, all accepted translations were trimmed as before and combined with the core-drafts to constitute the ultimate reliable RdRP core domain sequences. All vOTUs containing RdRP core domain sequences in at least one member are labeled as RvOTUs.

Finally, considering the higher conservation at the amino acid level, core-RdRPs were further clustered at a higher 95% amino-acid identity than nucleotide acid using MMseq2 with parameters "--min-seq-id 0.95 -c 0.333 --cluster-mode 2 --cov-mode 1 --seq-id-mode 2" and the representative sequences given by MMseq2 were selected as the RCR95 set.

In parallel, based on the virus taxonomy database published by the ICTV Virus Metadata Resource (VMR) (version VMR_20-190822_MSL37.2), we selected representative genomes of known *Orthornavirae* viruses at approximately genus level and employed the same method to search RdRP as their representatives for following phylogenetic reconstruction.

False positive RdRp filtering process

To filter out potential misidentified sequences caused by false positive motif matches from reverse transcriptases. We constructed a database containing various non-*Orthornavirae* sequences with domains homologous to RdRPs' domain by keywords searching (reverse transcriptase, group 2 intron and group II intron) from NCBI nr database. All RdRPs retrieved from our samples were compared to this database using blastp to prevent contamination and got a best threshold for RdRP acceptance.

RdRP-based phylogenetic reconstruction

To facilitate the search for high taxonomic level unknown viruses, we re-clustered the RCR95 members using MMseq2 with sequence identity ranged from 0.8 to 0.1 and other parameters kept the same as those used during the construction of RCR95. Then, annotation purity of each cluster set showed that the RCR30 set (clustering identity equals 0.3) achieved minimal acceptable identity threshold for phylum level clustering because the purity experienced a sharp decline when the clustering identity reached 0.2 (Supplementary Data 5). So, the RCR30 set was analyzed for taxonomic purity and the proportion of unknown elements at different taxonomic levels (Supplementary Fig. 3A, B). Subsequently, the RCR30 sets were proven to be suitable as a phylum-level clustering representative, exhibiting high purity at phylum level. Twenty-five clusters from the RCR30 sets, comprising members without geNomad taxonomic annotations, were selected for subsequent phylogenetic reconstruction.

On the other hand, we incorporated an average of five representative sequences per family from ICTV sequences into the phylogenetic tree construction. These sequences provided a reliable reference for taxonomic classification of clades. In total, 643 sequences were selected (Pisuviricota: 180, Kitrinoviricota: 86, Lenarviricota: 168, Negarnaviricota: 193, and Duplornaviricota: 16).

Considering the candidate phyla provided in multiple studies[14,15], we compared these candidates with our RdRPs using blastp. The results indicated that only Neri et al.'s *p.0002* yielded positive hits with our candidate RdRPs, so only Neri et al.'s *p.0002* were included in this phylogenetic reconstruction together with reference sequences from ICTV, and other candidate phyla were excluded considering that they are not certified by ICTV yet.

Passed 25 RCR30 sets (unknown part), 5 groups of core-RdRPs of different *Orthornavirae* phyla (known part) from ICTV Virus Metadata Resource (VMR)[38], an extra candidate phylum (validation part) from Neri et al.'s work[14] and an outgroup consisting of three reverse transcriptase sequences (RT) were passed to MUSCLE[82] (version 5.1) Multiple Sequence Alignment (MSA) separately and yielded 32 MSA files. Consensus sequences of each MSA file were generated by HHconsensus[83] (version 2.0.16), and the sites with more than 95% gap characters were excluded. Then, all 32 consensus sequences were aligned together using MUSCLE, generating a template MSA file. Following this, each gap position in the template MSA file was expanded to the corresponding column of all 32 MSA files and merged into the final one.

Finally, the merged MSA file was used to reconstruct an approximate maximum likelihood tree using the FastTree[84] tool (V.2.1.4) with the WAG evolutionary model and gamma-distributed site rates[85]. The tree was re-rooted between RTs and RdRPs and was visualized using the online tool iTOL[86].

Phylogenetic reconstruction for *Duplornaviricota* polyphyly analysis

To validate the polyphyly nature of *Duplornaviricota* claimed by Zayed et al.[15] we constructed another tree with all RdRPs from *Duplornaviricota* vOTU annotated by geNomad and RdRPs from ICTV representatives. We made an MSA file out of them using MUSCLE as before and then built the phylogenetic tree with FastTree program.

Clade taxonomic annotation of phylogenetic tree

The principles of phylogenetic tree clade taxonomic annotation follow the work of Neri et al.[14].

Tree leaves, composed of ICTV VMR members (see "RdRP-based phylogenetic reconstruction"), were used as reference leaves for taxonomic affiliation of clades.

All leaves were assigned to certain taxa based on the taxonomic composition of ICTV VMR members within their respective clades, adhering to the following principles:

- All descendants of their last common ancestor were assigned to a single taxon and sequences descending from deeper nodes should be assigned to a different taxon of the same taxonomic rank.
- The depth of different taxa rank depends on their upper rank nodes, which means, e.g., depths of classes vary among different phyla.

Notably, application of these principles is different among sections. In the tree generated in "RdRP-based phylogenetic reconstruction", all clade annotations followed these rules. But during "Phylogenetic reconstruction for *Duplornaviricota* polyphyly analysis", leaves annotated as *Duplornaviricota* members were excluded while others still stick to these principles above, and then all *Duplornaviricota* members and corresponding clades were assigned taxa at class rank separately.

In the two sections above, phylum and class rank taxa were assigned to trees and only 11 leaves out of 2021 were found nested into wrong place. We used the most agnostic way and delabeled them to resolve the conflict.

Finally, the remaining unlabeled clades were assigned to novel taxon according to their depth. In this work, two candidate phyla were named as *upv.p.*xxx (i.e., prefixed by upv for urban & peri-urban RNA virus and p for phylum, e.g., *upv.p.*001 is the first potential candidate phylum).

Extra phylogenetic reconstruction for highly concerned pathogenic viral families

To examine the phylogenetic diversity of certain pathogenic viral families. We incorporated the genomes of all corresponding family members provided by ICTV Virus Metadata Resource (VMR), along with RvOTUs annotated as the same family member, to construct the RdRP-based phylogenetic trees. Core RdRPs of ICTV reference genomes were extracted by being searched using core-RdRPs from their RvOTU relatives as queries and trimmed into the mapped regions of the best hits given by tblastn.

Among 6 pathogenetic familied we chose, *Orthomyxoviridae* has specificity as its RdRP is composed with three parts (PB1, PB2, PA), and we chose PB1 as representative of RdRP. Moreover, *Thogotovirus thailandense* has its PB1 divided into two subunits (MN095540), so we lumped them into one sequence during the phylogenetic reconstruction.

Relative abundance estimation for Urban-RvOTU

Bowtie2[79] (version 2.4.4) was used to map the reads back to the corresponding contigs using end-to-end alignment for each sample. After that, the alignments were sorted and indexed using SAMtools[87] (version 1.14). The command *coverage* in checkM[88] tool (version 1.2.0) was used to calculate the average coverage depth and quantity of mapped reads for each contig in each sample. The relative abundance of each contig was estimated using FPKM metric (Fragments Per Kilobase Million), which was computed by normalizing the mapped read count based on the contig length and the total number of reads in the sample. Specifically, the relative abundance of a phylum in a sample was calculated by aggregating the relative abundance values of the contigs classified under that phylum. Similarly, for a specific Urban-RvOTU, its relative abundance in a sample was calculated as the sum of the relative abundance values of the contigs attributed to that particular Urban-RvOTU. A threshold of relative abundance ≥ 0.001% was set to indicate the presence of the Urban-RvOTU in a sample[89].

ORF calling and domain annotation

For each viral contig, we used Prodigal[90] (v2.6.3) with the metagenomic mode ('anonymous") and the default parameters to predict the ORFs. The predicted ORFs were then subjected to domain annotation using InterPro, SignalP, TMHMM and MobiDBLite[91–93].

Genome molecular type prediction

To validate the genome molecular type of the distinctive double-stranded genome clade within the phylum *Pisuviricota*, we employed the ratio of the number of reads mapping to the positive strand versus those mapping to the negative strand (i.e., reads matching the reverse complement of the positive strand). This metric is defined as the log2FRratio and is calculated as follows:

$$\log FRratio = \log_2\left(\frac{RC_f}{RC_r}\right) \quad (1)$$

where RC_f represents the number of reads mapping in the forward orientation to the reference genome, and RC_r represents the number of reads mapping in the reverse orientation. Since all samples were

prepared using a strand-specific library preparation protocol, this ratio allows for a reasonable approximation of the positive-to-negative strand ratio in the original sample, although PCR amplification bias during sequencing may still influence the absolute value.

To ensure high-quality estimates, we selected control groups from the samples containing *pisu.u.c.1* to serve as references. The control set comprised high-quality ssRNA+ and dsRNA RvOTUs with genome completeness >80% and read counts >10,000, which also had definitive taxonomic annotations at the sub-Class level. Prior to analysis, the strand orientation of all selected genomes was corrected according to their gene-coding direction. Besides, we included only reads where both ends of a paired-end read pair successfully mapped. Therefore, alignments recorded in the SAM file with a flag of 99 or 147 were recorded as one forward mapping event, while those with a flag of 83 or 163 were counted as one reverse mapping event.

In silico host prediction for RNA viruses

We employed best-hit approach and CRISPR spacer detection to predict the host of RNA viruses based on sequence similarity to the isolates and assembled genomes in the ICTV database as well as CRISPR spacer to protospacer matching. For the former, we mapped the RdRP sequences to the ICTV database using tblastn and matches with identity greater than 50% and query coverage over 75% remained as effective hits. The host information from the ICTV-VMR database was cloned to the matched RCR95 representative based on the highest score. For the vertebrate hosts, we further subdivided them based on the knowledge recorded in the VIRION database[57], an atlas of the vertebrate-virus associations.

The RNA virus sequences were also compared to known bacterial and archaeal CRISPR spacer sequences to predict the potential prokaryotic host of these viruses. We first retrieved 29,867,318 CRISPR spacers from IMG/VR[94] and 1,846,441 CRISPR spacers predicted from 79,735 UHGG collection of genomes[95]. To control false-positive hits, CRISPR spacers encoded in a predicted CRISPR array containing less than three spacers and spacers shorter than 20 bp were excluded. Spacers were searched against viral genomes using blastn (version 2.12.0, blastn -dust no -word_size 7 -evalue 1e-3)[96–98]. Hits with zero or one mismatch and zero or one gap over the whole spacer length were considered.

Host assignment conflicts arising between ICTV-VMR and VIRION are resolved in favor of the ICTV-VMR standards. Between best-hit approach and CRISPR spacer detection, no assignment conflict occurred, and the results were merged (Supplementary Data 10).

Environmental core-virome extraction

To extract the core-virome for each environment set (built, non-built, water environments), we conducted an iterative search for highly prevalent RvOTUs under prevalence thresholds ranging from 5% to 50%. Initially, we extracted RvOTUs with prevalence greater than the threshold within each environment of an environment set, and then we took the intersection of the RvOTUs extracted from each environment to obtain the core-virome for this environment set. To balance the size of core-viromes of each environment set, we calculated the Coefficient of Variation (CV) for the size of core-viromes in each environment set after each iteration. The results showed that at a prevalence threshold of 10%, a relatively balanced number of core-viromes was obtained across all environment sets (Supplementary Fig. 9, Supplementary Data 11).

Constructing environmental networks with prevalent urban-RvOTUs

For a particular environment, Urban-RvOTUs present in more than 10% of the samples collected from that environment were considered as prevalent. We then determined the similarities between environments with respect to the prevalent Urban-RvOTUs by calculating the Jaccard index and a one-tailed p value of observing at least c prevalent Urban-RvOTUs in common between each pair of environments. The Jaccard index was calculated as follows,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where set A represents the highly present Urban-RvOTUs in environment A and set B represents the set of highly present Urban-RvOTUs in environment B.$|A \cap B|$ and $|A \cup B|$ denote the intersection and union of two sets, respectively. The one-tailed p value was calculated based on the following hypergeometric equation as per Bin Jang et al.[99]:

$$p(X \geq c) = \sum_{i=c}^{\min(a,b)} \frac{C_a^i C_{n-a}^{b-i}}{C_n^b} \tag{3}$$

in which c is the number of prevalent Urban-RvOTUs in common; a and b are the number of prevalent Urban-RvOTUs in environment A and B, respectively; n is the total number of prevalent Urban-RvOTU in environment A or B; and C_a^i denote the number of combinations of a taken i at a time. The hypergeometric formula calculated the probability of sharing a number of common prevalent Urban-RvOTUs between two environments at or above the number (c) under the null hypothesis that the observation is likely to occur by chance. Environment pairs with $J(A,B)$ greater than 0.1 and p value less than 0.001 were regarded as similar.

Directional selection analysis

FADE, which is implemented in Hyphy[43] (version 2.4.0 and above), was used to conduct the directional selection analysis. The purpose of this computational method is to assess whether specific sites within a protein alignment are evolving toward a particular residue at faster rates along certain branches, in comparison to a baseline (background) model. FADE specifically evaluates each site in an alignment, determining if a group of foreground branches exhibits a preferential substitution toward a certain amino acid compared to background branches. A high bias parameter suggests a significant increase in substitutions toward a specific amino acid at that site. Additionally, Empirical Bayes Factors (EBF) are used in FADE to gauge statistical significance (EBF threshold of ≥ 100), offering robust evidence of directional selection at work. Utilizing a random effects model combined with approximation methods inspired by Latent Dirichlet allocation (LDA), FADE efficiently assigns sites to different rate classes. In contrast with most Hyphy methods, FADE does not utilize a reversible Markov model as its goal is to pinpoint directional selection, necessitating a rooted phylogeny for analysis.

Statistical analysis

A multivariate analysis of all samples was conducted to assess for significant differences in the RNA viral communities collected from different environments using non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities. Analysis of similarities (ANOSIM) was performed to test for significant differences in RNA viral communities, generating a sample statistic (R) that represents the degree of separation between the assessed groups. The NMDS and ANOSIM were performed in R v4.2.2 using the "*vegan*" package[100].

To assess the significance of the prevalence of a specific RNA virus in a certain environment, we conducted a hypergeometric test as follow,

$$p_{v,e} = \frac{C_{m_v}^{k_{v,e}} C_{N-m_v}^{n_e-k_{v,e}}}{C_N^{n_e}} \tag{4}$$

where N is the total number of samples used in this study, n_e is the number of samples collected in a specific environment, e, m_v denotes the number of samples where the specific RNA virus is present in, $k_{v,e}$

is the number of samples in which a specific virus is found in the environment e. Values of $p_{v,e}$ below 0.05 were considered significant.

Socioeconomic and climate-related factors accession

The World Development Indicators (WDI) data were accessed from World Bank Group (https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators). We used data from 2020, and missing data were filled using records from the most recent recorded year.

Silhouettes in figures

Silhouettes of animals, microbes and insects in Fig. 5D and Fig. 6D were obtained from PhyloPic (http://www.phylopic.org).

In Fig. 5D, the acer silhouette was created by T. Michael Keesey, published under a Creative Commons Attribution-ShareAlike 3.0 Unported license (CC BY-SA 3.0); the silhouette has been resized and rotated; this work is distributed under the same license as the original. Aralia silhouette by Tree of Life App, published under a Creative Commons Attribution 4.0 International license (CC BY 4.0); the silhouette has been resized and rotated. Proso millet silhouette by T. Michael Keesey, published under a Creative Commons Attribution 4.0 International license (CC BY 4.0); the silhouette has been resized and rotated. Clostridium silhouette by Gareth Monger, published under a Creative Commons Attribution 3.0 Unported license (CC BY 3.0); the silhouette has been resized and rotated. Spirogyra silhouette by Matthew Crook, published under a Creative Commons Attribution-ShareAlike 3.0 Unported license (CC BY-SA 3.0); the silhouette has been resized and rotated; this work is distributed under the same license as the original. The other silhouettes in Fig. 5D are all licensed under CC0.

Silhouettes of human and insects in Fig. 6D are all licensed under CC0 or PDM 1.0 (https://creativecommons.org/publicdomain/mark/1.0/).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data generated in this study are available in the National Genomics Data Center of China under accession code PRJCA026924. Other generated data are available at https://zenodo.org/records/17823213. Source data are provided in this paper. Source data are provided with this paper.

Code availability

The assembled viral contigs and the custom codes utilized in this study are available at the following GitHub repository: https://github.com/MetaSUB/Urban-RNA-Virome.

References

1. Varanda, C., Felix, M. D. R., Campos, M. D. & Materatski, P. An overview of the application of viruses to biotechnology. *Viruses* **13**, 2073 (2021).
2. Mushegian, A. R. Are there 10(31) virus particles on earth, or more, or fewer? *J. Bacteriol.* **202**, e00052–20 (2020).
3. Ye, S. et al. An atlas of human viruses provides new insights into diversity and tissue tropism of human viruses. *Bioinformatics* **38**, 3087–3093 (2022).
4. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2**, 63–77 (2012).
5. Woolhouse, M. & Gaunt, E. Ecological origins of novel human pathogens. *Crit. Rev. Microbiol.* **33**, 231–242 (2007).
6. Harvey, E. & Holmes, E. C. Diversity and evolution of the animal virome. *Nat. Rev. Microbiol.* **20**, 321–334 (2022).
7. Wang, H. et al. Genetic diversity of RNA viruses infecting invertebrate pests of rice. *Sci. China Life Sci.* **67**, 175–187 (2024).
8. Kourelis, J., Marchal, C., Posbeyikian, A., Harant, A. & Kamoun, S. NLR immune receptor-nanobody fusions confer plant disease resistance. *Science* **379**, 934–939 (2023).
9. Sourimant, J. et al. 4′-Fluorouridine is an oral antiviral that blocks respiratory syncytial virus and SARS-CoV-2 replication. *Science* **375**, 161–167 (2022).
10. Hillary, L. S., Adriaenssens, E. M., Jones, D. L. & McDonald, J. E. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Commun.* **2**, 34 (2022).
11. Chen, Y. M. et al. RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat. Microbiol.* **7**, 1312–1323 (2022).
12. Suttle, C. A. Marine viruses–major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
13. Dominguez-Huerta, G. et al. Diversity and ecological footprint of Global Ocean RNA viruses. *Science* **376**, 1202–1208 (2022).
14. Neri, U. et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023–4037.e4018 (2022).
15. Zayed, A. A. et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
16. He, W. T. et al. Virome characterization of game animals in China reveals a spectrum of emerging pathogens. *Cell* **185**, 1117–1129.e1118 (2022).
17. Hou, X. et al. Using artificial intelligence to document the hidden RNA virosphere. *Cell* **187**, 6929–6942.e6916 (2024).
18. Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).
19. Stern, A. et al. The evolutionary pathway to virulence of an RNA virus. *Cell* **169**, 35–46.e19 (2017).
20. Duraes-Carvalho, R., Ludwig-Begall, L. F., Salemi, M., Lins, R. D. & Marques, E. T. A. Influence of directional positive Darwinian selection-driven evolution on arboviruses Dengue and Zika virulence and pathogenesis. *Mol. Phylogenet Evol.* **140**, 106607 (2019).
21. Martin, D. P. et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200.e5187 (2021).
22. Wang, W., Zhao, H. & Han, G. Z. Host-virus arms races drive elevated adaptive evolution in viral receptors. *J. Virol.* **94**, 16 e00684–16.e00620 (2020).
23. Li, N. K., Corander, J., Grad, Y. H. & Chang, H. H. Discovering recent selection forces shaping the evolution of dengue viruses based on polymorphism data across geographic scales. *Virus Evol.* **8**, veac108 (2022).
24. An, D. et al. Receptor affinity-selective differential dynamics of membrane fusion initiation govern deltacoronavirus cross-species transmission. *Sci. China Life Sci.* **68**, 3756–3774 (2025).
25. Gong, P. et al. Urbanisation and health in China. *Lancet* **379**, 843–852 (2012).
26. Baker, R. E. et al. Infectious disease in an era of global change. *Nat. Rev. Microbiol.* **20**, 193–205 (2022).
27. Liston, A., Humblet-Baron, S., Duffy, D. & Goris, A. Human immune diversity: from evolution to modernity. *Nat. Immunol.* **22**, 1479–1489 (2021).
28. Li, X. et al. Global urban growth between 1870 and 2100 from integrated high resolution mapped data and urban dynamic modeling. *Commun. Earth Environ.* **2**, 201 (2021).
29. Elhaik, E., Ahsanuddin, S., Robinson, J. M., Foster, E. M. & Mason, C. E. The impact of cross-kingdom molecular forensics on genetic privacy. *Microbiome* **9**, 114 (2021).

30. Sharifi, A. & Khavarian-Garmsir, A. R. The COVID-19 pandemic: impacts on cities and major lessons for urban planning, design, and management. *Sci. Total Environ.* **749**, 142391 (2020).
31. Barbarossa, L. The post pandemic city: challenges and opportunities for a non-motorized urban environment. An overview of Italian cases. *Sustainability* **12**, 7172 (2020).
32. Rocha, N. P. et al. Smart cities and public health: a systematic review. *Procedia Comput. Sci.* **164**, 516–523 (2019).
33. Danko, D. et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* **184**, 3376–3393.e3317 (2021).
34. Pipito, L. et al. Antimicrobial Resistance in ESKAPE Pathogens: a Retrospective Epidemiological Study at the University Hospital of Palermo, Italy. *Antibiotics (Basel)* **14**, 186 (2025).
35. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. *J. Infect. Dis.* **197**, 1079–1081 (2008).
36. Ryon, K. A. et al. A history of the MetaSUB consortium: tracking urban microbes around the globe. *iScience* **25**, 104993 (2022).
37. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
38. Lefkowitz, E. J. et al. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **46**, D708–D717 (2018).
39. Wolf, Y. I. et al. Origins and evolution of the global RNA virome. *mBio* **9**, e02329–18 (2018).
40. Wolf, Y. I. et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat. Microbiol.* **5**, 1262–1270 (2020).
41. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
42. Krupovic, M., Wolf, Y. I., Koonin, E. V. & Kuhn, J. H. ICTV virus taxonomy profile: yueviridae 2023. *J. Gen. Virol.* **104**, 001904 (2023).
43. Kosakovsky Pond, S. L. et al. HyPhy 2.5-a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).
44. Lucaci, A. G. et al. RASCL: rapid assessment of selection in clades through molecular sequence analysis. *PLoS One* **17**, e0275623 (2022).
45. Shi, W. et al. A nucleobase-binding pocket in a viral RNA-dependent RNA polymerase contributes to elongation complex stability. *Nucleic Acids Res.* **48**, 1392–1405 (2020).
46. Ahmad, J., Ikram, S., Ahmad, F., Rehman, I. U. & Mushtaq, M. SARS-CoV-2 RNA Dependent RNA polymerase (RdRp) – A drug repurposing study. *Heliyon* **6**, e04502 (2020).
47. Jian, H. et al. Diversity and distribution of viruses inhabiting the deepest ocean on Earth. *ISME J.* **15**, 3094–3110 (2021).
48. Chow, C. E. & Suttle, C. A. Biogeography of viruses in the sea. *Annu Rev. Virol.* **2**, 41–66 (2015).
49. Chu, Y., Zhao, Z., Cai, L. & Zhang, G. Viral diversity and biogeochemical potential revealed in different prawn-culture sediments by virus-enriched metagenome analysis. *Environ. Res.* **210**, 112901 (2022).
50. Gao, S. et al. Patterns and ecological drivers of viral communities in acid mine drainage sediments across Southern China. *Nat. Commun.* **13**, 2389 (2022).
51. Bogler, A. et al. Rethinking wastewater risks and monitoring in light of the COVID-19 pandemic. *Nat. Sustain.* **3**, 981–990 (2020).
52. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
53. Dolja, V. V. & Koonin, E. V. Common origins and host-dependent diversity of plant and animal viromes. *Curr. Opin. Virol.* **1**, 322–331 (2011).
54. Bean, A. G. D. et al. Studying immunity to zoonotic diseases in the natural host — keeping it real. *Nat. Rev. Immunol.* **13**, 851–861 (2013).
55. 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).
56. Wu, R. et al. RNA viruses linked to eukaryotic hosts in thawed permafrost. *mSystems* **7**, e0058222 (2022).
57. Carlson, C. J. et al. The global virome in one network (VIRION): an atlas of vertebrate-virus associations. *mBio*. **13**, e0298521 (2022).
58. Mukherjee, S. et al. Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res.* **51**, D957–D963 (2023).
59. Chen, I. A. et al. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* **51**, D723–D732 (2023).
60. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
61. Bhattacharjee, D., Flores, C., Woelfel-Monsivais, C. & Seekatz, A. M. Diversity and prevalence of Clostridium innocuum in the human gut microbiota. *mSphere* **8**, e0056922 (2023).
62. Zafar, H. & Saier, M. H. Jr Gut Bacteroides species in health and disease. *Gut Microbes* **13**, 1–20 (2021).
63. World Bank. World Development Indicators. https://databank.worldbank.org/source/world-development-indicators (2019).
64. Norder, H. et al. Picornavirus non-structural proteins as targets for new anti-virals with broad activity. *Antivir. Res.* **89**, 204–218 (2011).
65. Yi, J. et al. Picornavirus 3C - a protease ensuring virus replication and subverting host responses. *J. Cell Sci.* **134**, jcs253237 (2021).
66. Luke, G. A. et al. Occurrence, function and evolutionary origins of '2A-like' sequences in virus genomes. *J. Gen. Virol.* **89**, 1036–1042 (2008).
67. Takada, K. & Holmes, E. C. Genome sizes of animal RNA viruses reflect phylogenetic constraints. *Virus Evol.* **11**, veaf005 (2025).
68. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol Mol. Biol. Rev.* **84**, e00061-19 (2020).
69. Egan, M., Dempsey, E., Ryan, C. A., Ross, R. P. & Stanton, C. The sporobiota of the human gut. *Gut Microbes* **13**, 1–17 (2021).
70. Thomas, F., Hehemann, J. H., Rebuffet, E., Czjzek, M. & Michel, G. Environmental and gut bacteroidetes: the food connection. *Front Microbiol.* **2**, 93 (2011).
71. Patil, A., Banerji, R., Kanojiya, P., Koratkar, S. & Saroj, S. Bacteriophages for ESKAPE: role in pathogenicity and measures of control. *Expert Rev. Anti Infect. Ther.* **19**, 845–865 (2021).
72. Ilmjarv, S. et al. Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant. *Sci. Rep.* **11**, 13705 (2021).
73. Stevens, L. J. et al. Mutations in the SARS-CoV-2 RNA-dependent RNA polymerase confer resistance to remdesivir by distinct mechanisms. *Sci. Transl. Med.* **14**, eabo0718 (2022).
74. Yang, Z. et al. Crop antiviral defense: Past and future perspective. *Sci. China Life Sci.* **67**, 2617–2634 (2024).
75. Xu, Y., Li, Y. & Wei, F. Conservation medicine: a "Unity of Nature and Man" perspective. *Sci. China Life Sci.* **68**, 2475–2477 (2025).
76. Zhao, L. et al. Existing host range mutations constrain further emergence of RNA viruses. *J. Virol.* **93**, e01385-18 (2019).
77. Afshinnekoo, E. et al. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst.* **1**, 97–97.e93 (2015).
78. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
79. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

80. Li, D. et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
81. Venkataraman, S., Prasad, B. & Selvarajan, R. RNA dependent RNA polymerases: insights from structure, function and evolution. *Viruses* **10**, 76 (2018).
82. Edgar, R. C. Muscle5: h1igh-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
83. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
85. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
86. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
87. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
88. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
89. Chang, D. et al. Gut microbiome wellness index 2 enhances health status prediction from gut microbiome taxonomic profiles. *Nat. Commun.* **15**, 7447 (2024).
90. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
91. Teufel, F. et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
92. Piovesan, D. et al. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* **49**, D361–D367 (2021).
93. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
94. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2023).
95. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
96. Parcey, M., Gayder, S., Castle, A. J. & Svircev, A. M. Function and application of the CRISPR-Cas system in the plant pathogen erwinia amylovora. *Appl Environ. Microbiol.* **88**, e0251321 (2022).
97. Feike, J. et al. Measuring unbiased metatranscriptomics in suboxic waters of the central Baltic Sea using a new in situ fixation system. *ISME J.* **6**, 461–470 (2012).
98. Chao, S. et al. Metatranscriptomic sequencing suggests the presence of novel RNA viruses in rice transmitted by brown planthopper. *Viruses* **13**, 2464 (2021).
99. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
100. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Vegetation Sci.* **14**, 927–930 (2003).

Acknowledgements

We thank all the members in the MetaSUB consortium for generating the raw data used in this study.

Author contributions

Conceptualization, J.W., L.-F.W., C.E.M. and T.S.; methodology, J.W., Z.G., A.G.L., J.O., K.A.R. and L.W.; software, J.W., Z.G.; formal analysis, J.W., Z.G. and A.G.L.; data curation, J.W., Z.G., A.G.L., J.O., L.W., K.A.R., A.J.P., E.E., T.P.V., A.C., C.A.O., M.O., E.D.-N., O.O., M.P., D.M., M.B., J.A.U., A.T., K.H.R., M.A.S., B.T.T., B.P., N.K.S., V.M., S.M., K.C.K., M.U., N.M.-P., K.K., K.M., M.K., R.B.T., W.B., K.P., B.S., A.F., P.P.Ł., J.G.B., Y.D., K.I.U., X.R., L.M.S., N.H. H.-C., H.S., P.K.H.L., A.P.C., N.C.K., L.-F.W., C.E.M. and T.S.; writing – original draft, J.W., Z.G., A.G.L., K.A.R., A.J.P., E.E., D.L., L.-F.W., C.E.M. and T.S.; writing – review and editing, all authors have reviewed and approved the manuscript.

Funding

T.S. discloses support for the research of this work from the National Key Research and Development Program of China [grant number 2023YFC2706503], the National Natural Science Foundation of China [grant number 32370720], the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU, and the Key Laboratory of MEA, Ministry of Education, ECNU. J.W. discloses support for the research of this work from the Natural Science Foundation of Shanghai [grant number 25ZR1401106]. L.W. discloses support for the research of this work from the Key Laboratory of Ecology and Energy Saving Study of Dense Habitat (Tongji University), Ministry of Education - Shanghai Tongji Urban Planning & Design Institute Co., Ltd Joint Research Project [grant number KY-2022-LH-A03], and the Shanghai Tongji Urban Planning & Design Institute Co., Ltd - China Intelligent Urbanization Co-creation Center for High Density Region Research Project [grant number KY-2022-PT-A02]. V.M. and S.M. disclose support for the research of this work from the Ministry of Research, Innovation and Digitization, under Romania's National Recovery and Resilience Plan - Funded by EU - NextGenerationEU program, project "Metagenomics and Bioinformatics tools for Wastewater-based Genomic Surveillance of viral Pathogens for early prediction of public health risks - (MetBio-WGSP)" [grant number 760286/27.03.2024, code 167/31.07.2023], within Pillar III, Component C9, Investment 8. A.P.C. discloses support for the research of this work from the São Paulo Research Foundation (FAPESP) [grant number 2021/10577-0]. A.P.C. and N.C.K. disclose support for the research of this work from the US DOE JGI, a DOE Office of Science User Facility, supported by the Office of Science of the US DOE operated under [grant number DE-AC02-05CH11231].

Competing interests

C.E.M. is a co-Founder of Biotia. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-73605-z.

Correspondence and requests for materials should be addressed to Jun Wu, Lin-Fa Wang, Christopher E. Mason or Tieliu Shi.

Peer review information *Nature Communications* thanks Nicolas Bejerman, José Cuevas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Zihao Gao[1,58], Jun Wu[1,2,58] ✉, Alexander G. Lucaci[3,58], Jian Ouyang[1], Lan Wang[4], Krista A. Ryon[3], Eran Elhaik[5], Alexander J. Probst[6], Xavier Rodó[7], Thirumalaisamy P. Velavan[8,9,10], Anastasia Chasapi[11], Christos A. Ouzounis[11,12], Manuela Oliveira[13,14], Emmanuel Dias-Neto[15], Olayinka Osuolale[16], Michael Poulsen[17], Dmitry Meleshko[18], Malay Bhattacharyya[19], Juan A. Ugalde[20,21], Alba Tull[22], Kathleen H. Rubins[23], Maria A. Sierra[3], Braden T. Tierney[24], Bharath Prithiviraj[25], Nitesh Kumar Sharma[26], Viorel Munteanu[27], Serghei Mangul[27,28,29,30], Kubra Can Kurt[3], Masayuki Ushio[31], Natalia Mazur-Panasiuk[28], Katarzyna Kopera[28], Kamila Marszałek[32,33], Michał Kowalski[32,33,34], Rodolfo Brizola Toscan[28], Wojciech Branicki[28,35], Krzysztof Pyrć[28], Paweł P. Łabaj[28], Balakrishnan Subramanian[36], Alina Frolova[37,38], Joshua G. Burkhart[39], Youping Deng[40], Klas I. Udekwu[41,42], Lynn M. Schriml[43], Nur Hazlin Hazrin-Chong[44], Haruo Suzuki[45], Patrick K. H. Lee[46], Antonio Pedro Camargo[47,48], Nikos C. Kyrpides[47], Di Liu[49], Lin-Fa Wang[50,51] ✉, Christopher E. Mason[3] ✉, Tieliu Shi[1,2,52] ✉ & The International MetaSUB Consortium*

[1]The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai, China. [2]Institute of Wildlife Epidemics, Shanghai, China. [3]Weill Cornell Medicine, New York, NY, USA. [4]College of Architecture and Urban Planning, Tongji University, Shanghai, China. [5]Department of Biology, Lund University, Lund, Sweden. [6]Environmental Metagenomics, Research Center One Health Ruhr of the University Alliance Ruhr, Faculty of Chemistry, University of Duisburg-Essen, Essen, Germany. [7]ICREA and AIRLAB, Climate and Health RG, CANU, ISGlobal, Barcelona, Spain. [8]Institute of Tropical Medicine, Universitätsklinikum Tübingen, Tübingen, Germany. [9]Vietnamese German Center for Medical Research, Hanoi, Vietnam. [10]Faculty of Medicine, Duy Tan University, Da Nang, Vietnam. [11]Biological Computation & Process Laboratory, Chemical Process & Energy Resources Institute, Centre for Research & Technology, Thessalonica 57001, Greece. [12]Artificial Intelligence & Information Analysis Lab, School of Informatics, Aristotle University of Thessalonica, Thessalonica, Greece. [13]i3S—Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal. [14]Ipatimup—Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal. [15]Rutgers Cancer Institute of New Jersey, Newark, NJ, USA. [16]AEMIDR, Department of Biological Science, Elizade University, Ondo State, Nigeria. [17]Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [18]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, NY, USA. [19]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. [20]Center for Bioinformatics and Integrative Biology, Facultad de Ciencias de la Vida, Universidad Andres Bello, Santiago, Chile. [21]SENTINET: Surveillance, Epidemiology and New Technologies for Infectious Emerging Threats, Santiago, Chile. [22]Carnegie Mellon University, Pittsburgh, PA, USA. [23]Trivedi Institute for Space and Global Biomedicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [24]The Two Frontiers Project, Fort Collins, CO, USA. [25]Microbiome & Microbiology R&D Science Platforms, Reckitt Benckiser LLC US, Montvale, NJ, USA. [26]Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, University of Southern California, Los Angeles, CA, USA. [27]Department of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Moldova. [28]Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland. [29]Biological and Life Sciences Division, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. [30]Department of Biological and Morphofunctional Sciences, College of Medicine and Biological Sciences, Stefan cel Mare University of Suceava, Suceava, Romania. [31]Department of Ocean Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China. [32]Jagiellonian University, Małopolska Centre of Biotechnology, Krakow, Poland. [33]Pomeranian Medical University, Department of Genomics and Forensic Genetics, Szczecin, Poland. [34]Pomeranian Medical University, Regional Centre for Digital Medicine, Szczecin, Poland. [35]Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland. [36]Computational Centre, University of Białystok, Białystok, Poland. [37]Institute of Molecular Biology and Genetics of NASU, Kyiv, Ukraine. [38]Kyiv Academic University, Kyiv, Ukraine. [39]Department of Cell & Molecular Biology, John A. Burns School of Medicine, University of Hawai'i, Honolulu, HI, USA. [40]Department of Quantitative Health Sciences, John A. Burns School of Medicine, University of Hawai'i, Honolulu, HI, USA. [41]Department of Biological Sciences, Program for Bioinformatics and Computational Biology, University of Idaho, Moscow, ID, USA. [42]Swedish Environmental Epidemiology Centre & Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden. [43]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [44]Department of Biological Sciences & Biotechnology, Faculty of Science & Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia. [45]Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa, Japan. [46]School of Energy and Environment, State Key Laboratory of Marine Environmental Health, and Climate Impact and Environmental Resilience Research Center, City University of Hong Kong, Hong Kong SAR, China. [47]Department of Energy Joint Genome

Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [48]Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP, Brazil. [49]Computational Virology Group, Center for Bacteria and Virus Resources and Bioinformation, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China. [50]Programme in Emerging Infectious Diseases, Duke-NUS Medical School, Singapore, Singapore. [51]SingHealth Duke-NUS Global Health Institute, Singapore, Singapore. [52]Key laboratory of advanced theory and application in statistics and data science – MOE, East China Normal University, Shanghai, China. [58]These authors contributed equally: Zihao Gao, Jun Wu, Alexander G. Lucaci. ✉e-mail: jwu@bio.ecnu.edu.cn; linfa.wang@duke-nus.edu.sg; chm2042@med.cornell.edu; tlshi@bio.ecnu.edu.cn

The International MetaSUB Consortium

Jacqueline Proszynski[3], Daniel Brennan[3], Chris K. Deng[40], Andrew Wan[40], Masaki Nasu[40], Andre Kahles[53], Eduardo Castro-Nallar[54,55], Paula Quiroga[56], Mariana Massó[56], Harleen Kaur[57] & Angela Ivask[57]

[53]ETH Zurich, Zurich, Switzerland. [54]Centro de Ecología Integrativa, Universidad de Talca, Talca, Chile. [55]Instituto de Ciencias Biológicas, Universidad de Talca, Talca, Chile. [56]Institute of Research in Medical Microbiology and Parasitology, Faculty of Medicine, University of Buenos Aires, Buenos Aires, Argentina. [57]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia.