Microbiome

RESEARCH **Open Access**

Global biogeography of airborne viruses in public transit systems and their host interactions

Huaxin Lei[1], Shicong Du[1], Xinzhao Tong[1,2], Wing Lam Chan[1], Marcus H. Y. Leung[1], Kari O. Bøifot[3,4], Daniela Bezdan[5], Daniel J. Butler[5], David C. Danko[5], David C. Green[6,7], Mark T. Hernandez[8], Frank J. Kelly[6], Alexander G. Lucaci[5], Cem Meydan[5], Marina Nieto-Caballero[8], Krista Ryon[5], Braden Tierney[5], Klas I. Udekwu[9,10], Benjamin G. Young[5], Christopher E. Mason[5,11,12,13*], Marius Dybwad[3,4*] and Patrick K. H. Lee[14,15*]

Abstract

Background There is a diverse assemblage of microbes in air in built environments (BEs), but our understanding of viruses and their interactions with hosts in BEs remains incomplete. To address this knowledge gap, this study analyzed 503 metagenomes isolated from air samples from public transit systems in six global cities, namely Denver, Hong Kong, London, New York City, Oslo, and Stockholm. Viral genomes were recovered from samples via metagenomic binning, and viruses' taxonomy, functional potential, and microbial hosts were determined. The study also investigated correlations between virus and host abundances, the coevolution of clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems and anti-CRISPR (Acr) proteins, and the potential impacts of auxiliary metabolic genes (AMGs) on hosts.

Results Airborne viruses in global BEs exhibited biogeographical variations in diversity, composition, function, and virus–host interactions. Nearly half of the vOTUs analyzed were from the *Caulimoviridae* family, while 31.8% of them could not be taxonomically classified. Diverse functions were identified within the vOTUs, together with antimicrobial resistance genes with the potential to confer resistance to various antibiotics and antimicrobial agents. Strong correlations were observed between vOTU and host abundances, with clear distinctions between virulent and temperate viruses. However, there was limited co-evolution of CRISPR-Cas systems and Acr proteins, which was likely due to the oligotrophic and physical conditions in the BEs and the dominance of vOTUs with a virulent lifestyle. Phage-encoded AMGs appeared to have the potential to enhance host fitness. These findings highlight biogeographical variations in airborne viruses in BEs and that physical and oligotrophic conditions in BEs drive virus survival strategies and virus–host coevolution.

Conclusion There are biogeographical variations in airborne viruses in BEs in global cities, as physical and oligotrophic conditions in BEs drive virus survival strategies and virus–host coevolution. Moreover, the characteristics of airborne viruses in BEs are distinct from those of viruses found in other, more nutrient-rich ecosystems.

*Correspondence:
Christopher E. Mason
chm2042@med.cornell.edu
Marius Dybwad
marius.dybwad@ffi.no
Patrick K. H. Lee
patrick.kh.lee@cityu.edu.hk
Full list of author information is available at the end of the article

Keywords Airborne viromes, Built environments, Metagenomics, Virus–host coevolution

Introduction

The air in built environments (BEs) is oligotrophic and influenced by fluctuating physical and environmental conditions, yet it harbors a diverse assemblage of bacteria, fungi, and viruses [1]. The average concentrations of airborne bacteria and fungi in BEs are as high as ~ 1×10^5 particles m^{-3}, and that of viruses is similar [2, 3]. However, the compositions and metabolic functions of airborne bacteria and fungi in BEs are generally better understood than those of airborne viruses in BEs [4]. In previous work [5], different occupied rooms were found to contain distinct airborne viral communities, with human *papillomaviruses* and *polyomaviruses* being prevalent. Furthermore, airborne viruses in a mechanically ventilated venue were found to exhibit seasonal dynamics, highlighting the influence of season-associated factors such as air exchange rate and outdoor weather conditions on viral compositions in BEs [6].

As viruses can adapt their lifestyles to prevailing environmental conditions, there are substantial variations in the abundances of viruses and their hosts, especially when virulent predation leads to host cell lysis [7]. The equilibria between virus and host abundances in ecosystems can be described by ecological models, such as the Kill-the-Winner and Piggyback-the-Winner models [8], which shed light on viral prey strategies and underscore the complex nature of virus–host interactions [8, 9]. The lysis of microbial cells trigged by viral virulent infections plays a crucial role in the cycling of nutrients within natural ecosystems, ultimately benefiting microbial growth and the replication of future viral generations [10]. Viruses have also evolved the capacity to carry auxiliary metabolic genes (AMGs) for adaptation and survival [11]. These phage-encoded AMGs can enhance viral fitness by augmenting or redirecting host metabolism, such as photosynthesis, carbon metabolism, and nutrient cycling [12]. To counter viral infections, microbial hosts have developed clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems to recognize and degrade invading nucleic acids [13]. In response, viruses have evolved anti-CRISPR (Acr) proteins to evade the activity of CRISPR–Cas systems [14]. There is extensive evidence that in nutrient-rich environments, phage-encoded AMGs can hijack and manipulate essential metabolic pathways in hosts [15–17], and diverse CRISPR–Cas systems and Acr proteins undergo coevolution [18, 19]. For example, in wastewater treatment systems, the expression of diverse phage-encoded AMGs has been observed in microbial hosts, suggesting that phages play a role in pollutant removal [20, 21]. Similarly, in the human gut, a substantial proportion of viruses were found to be linked to specific microbial hosts through spacer sequences, in line with the Red Queen hypothesis [22] and indicative of ongoing coevolution resulting from a continuous battle of defense and counter-defense [23]. Although the host–prey strategies employed by viruses and the dynamics of virus–host coevolution in low-nutrient environments, such as marine [24] and desert ecosystems [25], have been studied, these processes in environments with continuous airflow—such as the air in BEs—are still largely unknown.

To fill the gap in knowledge on airborne viruses in BEs, this study analyzed 503 bulk metagenomes from air samples collected in public transit systems across six major global cities: Denver, Hong Kong, London, New York City, Oslo, and Stockholm. We aimed to determine the (i) diversity, composition, and functional potential of airborne viruses across global BEs; (ii) co-evolution of CRISPR-Cas systems and Acr proteins among microbial hosts and airborne viruses; and (iii) relationships between the temperate and virulent lifestyles of airborne viruses and host abundance. We hypothesized that the biological characteristics of airborne viruses would exhibit biogeographic patterns and that virus–host interactions would be influenced by oligotrophic conditions and the physical environments of BEs.

Methods

Sampling, genomic DNA extraction, and metagenomic sequencing

Five hundred three air samples were collected from public transit systems in Denver ($n = 13$), Hong Kong ($n = 159$), London ($n = 76$), New York City ($n = 96$), Oslo ($n = 127$), and Stockholm ($n = 32$) between June and July in 2018 and 2019 (Table S1). Except for Denver, where samples were collected from the city's rail and bus system, all samples were obtained from subway systems. Sampling was performed on weekdays during working hours, and the selected sampling locations exhibited different building characteristics (Table S1). All samples were collected using a SASS 3100 Dry Air Sampler (Research International; Monroe, WA, USA) equipped with an electret microfibrous filter at a flow rate of 300 L/min for 30 min. The sampler was positioned on a tripod, tilted at a 45° angle facing downward, and placed approximately 1.5 m above the floor. All samples were stored at − 80 °C until they were processed. Two types of negative control samples were prepared, namely field control samples, which were collected in each city by placing a new filter on the air sampler and not operating the sampler, and laboratory control samples, which were prepared by subjecting a new filter to a genomic DNA extraction process [26].

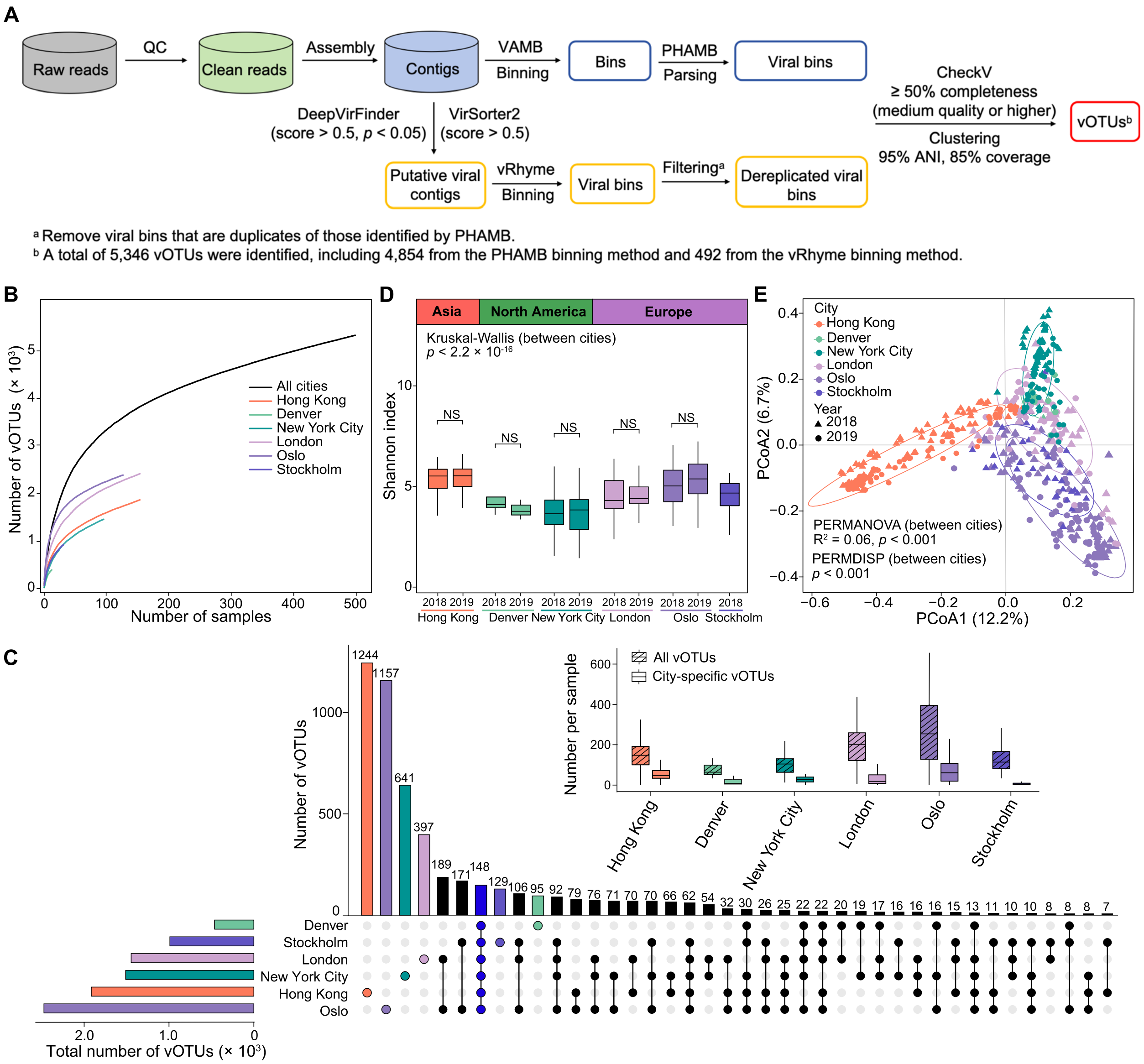
All samples were transported on dry ice to the Norwegian Defence Research Establishment (Kjeller, Norway) for genomic DNA extraction, which was performed via a previously described method [27]. Briefly, the particulates collected on the filters were extracted into NucliSENS Lysis Buffer (BioMérieux; Marcy-l'Étoile, France), and the resulting suspension was centrifuged to pellet the extracted material. The supernatant and pellet were separated, and the pellet was subjected to enzymatic and mechanical lysis to release genomic DNA. Inhibitors were removed from the lysate using a DNeasy PowerSoil Kit (QIAGEN; Germantown, MD, USA) according to the manufacturer's protocol and then combined with the original supernatant. Genomic DNA was extracted from the resulting solution using a NucliSENS Magnetic Extraction Reagents Kit (BioMérieux) according to the manufacturer's protocol, except with an increased volume of magnetic silica suspension (90 μL) and an extended incubation time (20 min) [28]. Alongside the air samples, 14 negative control samples and three positive control samples (ZymoBIOMICS Microbial Community Standard, Zymo Research; Irvine, CA, USA) were processed. Concentrations of genomic DNA were determined via Qubit dsDNA high-sensitivity assays conducted on a Qubit 3.0 Fluorometer (Thermo Fisher Scientific; Waltham, MA, USA). Sequencing libraries were constructed and paired-end 150-bp metagenomic sequencing of all samples was performed on an Illumina HiSeq X System (Illumina Inc.; San Diego, CA, USA) at the HudsonAlpha Genome Center (Huntsville, AL, USA) following a previously described protocol [26].

Quality control of raw sequences and assembly of reads into contigs

Raw reads were subjected to quality control filtering using Trim Galore (v0.6.10) [29] with the parameters "–length 50 –q 20." Subsequently, filtered reads were processed using KneadData (v0.7.4) (https://github.com/biobakery/kneaddata) with the default options and the Genome Reference Consortium Human Build 37 [30] as the reference to remove human sequences. Thereafter, any reads that could be mapped to contigs assembled from any of the negative control samples using MEGAHIT (v1.1.3) in MetaWRAP (v1.3.2) [31] were removed. The remaining reads were analyzed using Kraken2 (v2.1.3) [32] and Bracken (v2.8) [33] to determine their taxonomy and count information, respectively. Potential contaminant reads were identified using the R package decontam (v1.12) [34] in prevalence mode and with a probability threshold of 0.1 and were subsequently removed using the Python script "extract_kraken_reads.py" (https://github.com/jenniferlu717/KrakenTools). This quality filtering process afforded an average of 18.8 ± 7.2 million paired-end clean reads per sample. Finally, MEGAHIT (v1.1.3) in MetaWRAP (v1.3.2) with the default options was used to assemble the reads in each sample into contigs.

Recovery of viral genomes

To recover viral genomes from all the contigs, two binning methods were used (see Fig. 1A for the workflow). In the first method, reads were mapped to the assembled contigs using minimap2 (v2.24) [35], followed by filtering using samtools (v1.6) [36] and determination of contig coverage using the "jgi_summarize_bam_contig_depths" module in MetaBAT2 (v2.12.1) [37]. Subsequently, VAMB (v4.1.3) [38] with the default options was used to cluster the metagenomic contigs into bins. These bins were further parsed using the recommended workflow in PHAMB (v1.0.1) [39] to obtain viral bins. In the second method, putative viral contigs were identified using DeepVirFinder (v1.0) [40] with a score > 0.5 and p value < 0.05 [41], and VirSorter2 (v2.2.4) [42] with a score > 0.5 [43]. These contigs were then processed with vRhyme (v1.1.0) [44] to generate viral bins. Given the lower accuracy of DeepVirFinder (69–74%) and VirSorter2 (30–84%) in excluding non-phage bins compared to PHAMB (93–99%) [39], we also tested higher viral score thresholds (0.6, 0.7, 0.8, and 0.9) to assess their impact on vRhyme binning results. Only viral bins that did not contain contigs present in any bins obtained from PHAMB were retained. The viral operational taxonomic units (vOTUs) were determined by clustering all the viral bins at 95% average nucleotide identity and 85% alignment fraction [16], and the longest sequence within each cluster was used to represent a vOTU. CheckV (v1.0.1) [45] was used to identify proviruses and classify the vOTUs into five quality tiers with the default options, and only those classified as complete, high quality (> 90% completeness), and medium quality (50–90% completeness) were retained [41]. The validity of the derived vOTUs was further assessed using DeepVirFinder (v1.0) [40], VirSorter2 (v2.2.4) [42], and VIBRANT (v1.2.0) [46]. The viral lifestyle of the vOTUs was predicted using VIBRANT (v1.2.0) [46] with the "virome" flag, based on the following criteria: (1) genomes classified as viruses by VIBRANT were categorized as temperate if they contained integrase-like annotations or were identified as prophages, while all others were classified as virulent; (2) genomes not identified as viruses by VIBRANT were assigned an unknown lifestyle.



[a] Remove viral bins that are duplicates of those identified by PHAMB.

[b] A total of 5,346 vOTUs were identified, including 4,854 from the PHAMB binning method and 492 from the vRhyme binning method.

Fig. 1 Bioinformatics workflow and biogeography of viral operational taxonomic units (vOTUs) in air samples from built environments in global cities. **A** Bioinformatics workflow for viral bin identification used in this study. **B** Accumulation curves of vOTUs in relation to the number of samples analyzed. **C** Distribution of the number of vOTUs in samples from a given city and between samples from different cities. The inset figure shows the average total and city-specific numbers of vOTUs per sample from each city. **D** Shannon indices of the viral communities in samples from each city over a two-year period. Kruskal–Wallis tests were performed to examine differences between samples from cities regardless of the year, while Mann–Whitney tests were performed to examine differences between samples from each city in two different years. **E** Principal coordinate analysis of the Bray–Curtis dissimilarity matrix for the samples from the cities over a two-year period. Points are colored by city and shaped by year. Permutational multivariate analyses of variance (PERMANOVA) and permutational multivariate analyses of dispersion (PERMDISP) were performed without considering the year in which the samples were collected. The ellipses show the multivariate normal distribution at a 90% confidence interval for samples from each city. *NS*: not statistically significant

Taxonomic assignment of vOTUs

The open reading frames (ORFs) of vOTUs were predicted using Prodigal (v2.6.3) [47] with the parameter "–p meta." Species-level taxonomy of vOTUs was assigned by searching for the protein coding sequences in the IMG/VR database (v4.1) [48] using Diamond (v2.6.1) [49] (options: –evalue 1e-5 –max-target-seqs 10,000 –query-cover 50 –subject-cover 50) [23], and a customized Python script was used to retain only the top hit. Each vOTU was assigned the most common taxonomy based on the annotation of greater than 20% of its proteins [50]. Family-level taxonomies of vOTUs were assigned by constructing viral clusters (VCs) using average amino acid identity (AAI) and the number of shared viral proteins, as described previously [23]. Briefly, all vOTUs of medium quality or higher were combined with reference viral genomes from the NCBI RefSeq database ($n = 16{,}398$; retrieved on May 4, 2023), and clustering was performed based on greater than or equal to 20% AAI and either eight viral proteins or greater than 20% of viral proteins shared between genomes [23]. Each VC was assigned the most common taxonomy based on greater than 20% of its reference viral genomes. Singletons and VCs without a reference viral genome were novel viral families. The vOTUs within VCs were visualized using the R package Rtsne (v0.17). The phylogenetic tree of vOTUs in a VC, together with reference viral genomes, was constructed as previously described [23].

Estimation of vOTU coverage

To account for differences in read depths, the clean reads from each sample were rarefied to a uniform depth of 4.5 million reads, based on the sample with the least number of reads, using seqtk (v1.4) [51] with the option "–s 100." After rarefaction, five samples were excluded from further abundance and viral diversity analyses. The rarefied reads were then mapped to the vOTUs using Bowtie2 (v2.5.1) [52] with the "very-sensitive" model, and mappings with low identity were removed using CoverM (v0.6.1) (https://github.com/wwood/CoverM) with the parameter "–min-read-percent-identity 95." All filtered mappings were then entered into CoverM (v0.6.1) with the "contig" model and a setting of "–min-covered-fraction 0.7" to calculate the reads per kilobase per million mapped reads (RPKM) values for each vOTU in a sample [53]. The relative abundance of a vOTU in a sample was determined by dividing its RPKM value by the sum of the RPKM values for all vOTUs in that sample [53]. Similarly, the relative abundance of a VC was calculated by summing the relative abundances of all its vOTUs. A vOTU was considered present in the air in a BE in a given city if it was detected in at least one air sample from the BE in that city.

α- and β-diversity analyses

The R package vegan (v2.6–4) was used to assess the α-diversity (in terms of the Shannon index and observed species richness) and the β-diversity (in terms of the Bray–Curtis dissimilarity) of the viral communities. To visualize the dissimilarities in community composition, principal coordinate analysis was performed using the "cmdscale" function in vegan (v2.6–4) based on the Bray–Curtis dissimilarity.

Reconstruction of metagenome assembled genomes and coverage estimation

Metagenome -assembled genomes (MAGs) in each sample were reconstructed using the binning module in MetaWRAP (v1.3.2) with the options "–metabat2 –maxbin2 –concoct" and then further refined using the binning-refinement module in MetaWRAP (v1.3.2). Only MAGs with a completeness of greater than 50% and a contamination of less than 10% were retained [54]. These MAGs were then dereplicated using dRep (v3.4.2) [55] with the option "–sa 99" to generate representative MAGs (rMAGs). Subsequently, the GTDB-Tk (v2.1.1) [56] and the Genome Taxonomy Database (Release 208) were used to taxonomically assign the rMAGs [57]. Then, the ORFs of the rMAGs were predicted using Prodigal (v2.6.3) [47] with the parameter "–p meta." The rarefied reads were aligned against the rMAGs using Bowtie2 (v2.5.1) with the "very-sensitive" model, and mappings with low identity were removed using CoverM (v0.6.1) with the parameter "–min-read-percent-identity 95." Next, all filtered mappings were entered into CoverM (v0.6.1) with the "genome" model to calculate the RPKM values for each rMAG in a sample. The relative abundance of an rMAG in a sample was determined by normalizing its RPKM value in the same way as for vOTUs [58]. An rMAG was considered present in the air of a BE in a city if it was detected in at least one air sample from the BE in that city.

The virus-to-host abundance ratio (VHR) in a sample was calculated by dividing the abundance of vOTUs (in RPKM) by the abundance of their linked hosts represented by rMAGs (in RPKM) [24, 59]. Additionally, the virus-to-microbe abundance ratio in a sample was determined by dividing the total abundance of all vOTUs (in RPKM) by the total abundance of all rMAGs (in RPKM).

Prediction of virus–host links

To gain a comprehensive understanding of the potential hosts of airborne viruses, VirHostMatcher-Net [60], a network-based computational tool with a default set of 62,493 prokaryotic genomes, was first used to predict ex-situ hosts for the viral genomes. In-situ hosts were identified from the rMAGs recovered across all samples, using CRISPR spacer matches and similarities in integrated genome regions [16]. CRISPR spacers from all rMAGs were extracted, as described previously [23], and then mapped to viral genomes using Basic Local Alignment Search Tool (BLAST)–Short Nucleotide with the parameters "E value $\leq 10^{-5}$," "1 maximum target," "18 word-size," "$\geq$95% identity," and "$\leq$one mismatch" [23, 61]. This mapping process established a link between a mapped viral genome and a spacer-derived rMAG, thereby indicating a likely virus–host relationship. Furthermore, viral genome sequences were aligned to in-situ rMAGs using BLAST–Nucleotide with the settings “bitscore ≥ 50,” “*E* value ≤ 10^{-5},” and “identity ≥ 96%,” and viral genomes with a mapped region of greater than or equal to 1000 bp were considered linked to be the corresponding rMAG [23]. Thus, the vOTUs in an air sample of a BE from a given city were exclusively linked to the in-situ hosts present in air samples from the BE in that city, and the links for each city were visualized in a network using Cytoscape (v3.10.0) [62]. vOTUs classified as eukaryotic viruses (at the family level) were excluded from ex-situ and in-situ host analyses to maintain the study’s focus on prokaryotic hosts.

Annotation of viral functions, antibiotic resistance genes, and auxiliary metabolic genes

The functions of viral ORFs were annotated using five protein family databases (Kofam [63], TIGRFAM [64], Pfam [65], the Virus Orthologous Groups Database (VOGDB; http://vogdb.org), and the Earth’s Virome database [66]) through the hidden Markov model search method implemented with the hmmsearch utility in the HMMER package [67] with the default parameters. Each ORF was assigned an annotation based on the top-scoring alignment, which was determined by meeting the criteria (an *E* value ≤ 10^{-5} and a bitscore ≥ 60) [41]. ORFs that did not match any of the databases were classified as having an unknown function. The ORFs were clustered into gene clusters at 30% AAI and 70% alignment coverage [23] using MMseqs2 (v14.7e284) [68]. Gene functions were classified into six categories (i.e., DNA binding/regulation, lysis, replication, structural, transporters, and others) based on Pfam annotations, following the framework of Nayfach et al. [23], with the addition of a “transporters” category [41].

Antibiotic resistance genes (ARGs) in viral genomes were identified by searching three databases. The Comprehensive Antibiotic Resistance Database (v3.2.7) was searched using the Resistance Gene Identifier (RGI; v5.1.0) [69] with the option “–low_quality.” The NCBI Antimicrobial Resistance Finder (AMRFinder) database (v3.11) was searched using the NCBI AMRFinder tool (v3.11.14) [70] with the default options. The Structured Antibiotic Resistance Gene (SARG) Database (v3.0) [71] was searched using BLAST–Protein with thresholds of 80% identity and 70% coverage [21]. An ARG type is a set of genes that confer resistance against a specific class of antibiotics, while ARG subtypes are the individual ARGs comprising an ARG type [72]. All unique ARGs from the three databases were retained, and in the six instances in which different subtypes were identified for an ARG across the databases, annotations from RGI were adopted as they demonstrated greater consistency with other tools. The relative abundance of an ARG was determined based on the relative abundance of the corresponding viral genome [73], while the relative abundance of an ARG subtype in a sample from a BE in a given city was determined by summing the relative abundances of all the ARGs within that subtype present in samples from the BE in that city.

Putative auxiliary metabolic genes (AMGs) in viral genomes were detected using DRAM (v1.4.5) [74] by applying its “DRAM-v” function, following the recommended workflow. Viral genes were annotated using the default databases in DRAM, and those with an auxiliary score of 1 or 2 were classified as putative AMGs. The relative abundance of viral genomes containing putative AMGs was regarded as the relative abundance of putative AMGs [73]. The protein structure of a putative AMG was predicted using Phyre2 (v2.0) [75] in the normal modeling mode.

Identification of CRISPR–Cas systems and anti-CRISPR proteins

The CRISPR–Cas genes and arrays present in the rMAGs were identified using CRISPRCasTyper (v1.8.0) with the default options, which identifies CRISPR subtypes based on Cas genes and CRISPR repeat sequences [76]. The anti-CRISPR (Acr) homologs in viral genomes were identified using Anti-CRISPR-Associated Protein Finder (AcaFinder) [77] and the default Anti-CRISPR Protein Database (AcrDatabase). An Acr homolog was considered to be an Acr protein if a helix–turn–helix domain-containing protein was detected in at least one direction of the viral contig [78]. The resulting identified Acr proteins were mapped against the default AcrDatabase in AcaFinder [77] using Diamond (v2.6.1) [49] with an *E* value threshold of ≤ 10^{-5} to determine their subtype, and only the hits with a bitscore of greater than or equal to 60 were retained. A maximum likelihood phylogenetic tree of the identified Acr proteins with 339 experimentally validated Acr reference sequences obtained from the AcrHub database [79] was constructed using FastTree (v2.1.11) [80] with the default JTT model and using the multiple sequence alignment generated by Mafft (v7.520) [81].

Statistical analyses

All statistical analyses were conducted using R (v4.1.1). Between-group significance was assessed by conducting Mann–Whitney tests for two groups and Kruskal–Wallis tests for more than two groups. To evaluate the differences in viral compositions across cities and sampling years, permutational multivariate analysis of variance (PERMANOVA) was conducted using the “adonis2” function in the R package vegan (v2.6–4) (permutations = 999, method = "bray"). Permutational multivariate analysis of dispersion (PERMDISP) was conducted using the "betadisper" function in vegan (v2.6–4). Pearson's correlations and two-sided p values were calculated using the "stat_cor" function in the ggpubr package (v0.6.0). A p value of less than 0.05 was considered statistically significant in all statistical tests.

Results

Biogeography of airborne viruses in BEs

Analysis of the air samples collected from BEs in the six cities over a 2-year period included 503 metagenomes, from which 303,342 putative viral bins were recovered. After further quality filtering and clustering, 5346 vOTUs of at least medium quality, including 249 proviruses, were obtained, with 3084 (57.7%) classified as high quality or complete (Table S2). The majority of vOTUs (91.8%; $n = 4854$) were recovered using the PHAMB binning method, whereas 492 originated from vRhyme. Given the lower accuracy of DeepVirFinder (69–74%) and VirSorter2 (30–84%) compared to PHAMB (93–99%) [39] in excluding non-phage bins, we applied a stricter viral score threshold to evaluate its impact on vRhyme results. Raising the threshold to 0.9 reduced the number of vRhyme-derived vOTUs to 227 (Table S3), resulting in only a modest 4.9% decrease in the total vOTU count (Fig. S1). All 5346 vOTUs met at least medium-quality standards according to CheckV and were retained for downstream analyses. To further validate their viral identity, we analyzed the vOTUs using three widely adopted tools—DeepVirFinder, VirSorter2, and VIBRANT. Together, these analyses confirmed 82.8% of the vOTUs as viral (Table S4), reinforcing the reliability of both the vOTUs and the binning approaches. Of the 5346 vOTUs, 34.3% ($n = 1832$) were classified as virulent, 12.9% ($n = 692$) as temperate, and 52.8% had an unknown lifestyle (Table S4), consistent with distributions observed in other datasets (Fig. S2). The accumulation curves of vOTUs in samples from each city were unsaturated (Fig. 1B). City-specific vOTUs were found in samples from all cities, with the average number ranging from seven (in Stockholm) to 81 (in Oslo), while only 148 vOTUs were present in samples from all cities (Fig. 1C). Significant differences in α-diversity (the Shannon index and the number of observed species) were observed between samples from different cities regardless of the year (Kruskal–Wallis test, $p < 2.2 \times 10^{-16}$), while no significant differences were found between years in samples from a given city (Mann–Whitney test, $p > 0.05$; except for the number of observed species in samples from Oslo) (Fig. 1D and Fig. S3A). Samples from Hong Kong and Oslo showed significantly higher α-diversities than those from the other cities (Mann–Whitney test, $p < 0.05$). Samples from cities in North America showed significantly lower viral diversities than samples from cities on the other two continents in both years (Mann–Whitney test, $p < 0.01$) (Fig. S3B and S3C). Regarding β-diversity, viral compositions were found to cluster by city (permutational multivariate analysis of variance (PERMANOVA), $R^2 = 0.06$, $p < 0.001$), albeit with some dispersion (permutational multivariate analysis of dispersion (PERMDISP), $p < 0.001$) (Fig. 1E), and by continent, regardless of the year (PERMANOVA, $R^2 = 0.02$, $p < 0.001$; PERMDISP, $p = 0.20$) (Fig. S3D). Samples from a given city (except Denver) or from cities on a given continent showed variations in composition between years (pairwise PERMANOVA, $R^2 = 0.04$ to 0.12, $p < 0.001$; Table S5). Furthermore, viral community similarity showed a negative correlation with geographical distance across all cities and city pairs (Pearson's $r < -0.13$, $p < 0.001$), except for London–Oslo and Stockholm–Oslo ($p > 0.05$) (Table S6). Given that the airborne viral communities' α- and β-diversities were significantly influenced by geography, whereas their α-diversities were not influenced by time, our subsequent analyses focused on geographical differences rather than temporal differences.

Caulimoviridae and taxonomically unclassified vOTUs were dominant in the airborne viruses

Among all the vOTUs of medium quality or higher, only 3116 (58.3%) could be matched to a known genome in the Integrated Microbial Genomes/Virus (IMG/VR) database. The majority (95.9%) of the mapped vOTUs were taxonomically classifiable at the class level only and were primarily members of the *Caudoviricetes* (83.1%), *Revtraviricetes* (10.1%), or *Papovaviricetes* classes (1.4%) (Fig. S4A). The average relative abundance of viruses that were members of *Caudoviricetes* was highest in samples from Hong Kong (73.6 ± 23.4%), while the average relative abundance of viruses that were members of *Revtraviricetes* was highest in samples from Denver (9.2 ± 4.3%) (Fig. S4B). vOTUs from the *Papillomaviridae* family, including five classified genera (*Alphapapillomavirus*, *Betapapillomavirus*, *Dyodeltapapillomavirus*, *Dyothetapapillomavirus*, and *Gammapapillomavirus*), were detected in samples from all cities, with the highest average relative abundance observed in samples from Stockholm (2.3 ± 6.2%) (Fig. S4A and S4B). To improve the taxonomical assignment of the large number of unclassified vOTUs at the family level, all the vOTUs of medium quality or higher were clustered with reference viral genomes from the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database to yield 697 VCs (Table S7). The accumulation curve of VCs and vOTUs was also unsaturated (Fig. S5A). The majority of vOTUs within a given VC were closely clustered and thus distinct from those within other VCs (Fig. S5B). At the family level, 117 VCs, consisting of 3651 vOTUs, could be taxonomically classified, while a family-level taxonomy could not be assigned to the remaining vOTUs, which included 471 singletons (Table S7). VC1, the largest and most abundant VC, consisted of 2430 vOTUs with an average relative abundance of 46.2 ± 26.4% in samples across cities, and was affiliated with the *Caulimoviridae* family within the *Riboviria* realm [82] (Fig. S5C and Fig. S6). Nine hundred three vOTUs, comprising 266 vOTUs from VC2, the second largest cluster, and 637 vOTUs from 79 other VCs, were affiliated with an unclassified family in *Caudoviricetes*, accounting for an average relative abundance of 10.9 ± 11.5% in samples across cities (Fig. S5C and Fig. S6). The corresponding phylogenetic trees of the vOTUs affiliated with the *Caulimoviridae* family and the unclassified *Caudoviricetes* families, along with their corresponding reference genomes, revealed that many of the vOTUs in VC1 and all the vOTUs in VC2 formed a distinct monophyletic cluster in their respective trees (Fig. S7), suggesting that these vOTUs differed from known reference viruses. The relative abundances of some VCs in samples varied significantly between cities, especially VC1 and VC3 (Fig. S6). The average relative abundance of VC1 in samples from Hong Kong (19.7 ± 21.3%) was significantly lower than the average relative abundances of VC1 in samples from Denver (80.1 ± 10.3%), New York City (45.4 ± 24.7%), London (52.1 ± 20.5%), Oslo (71.3 ± 20.6%), and Stockholm (50.6 ± 15.5%) (Mann–Whitney test, $p < 0.001$). In contrast, the average relative abundance of VC3 (consisting of 232 vOTUs) in samples from Hong Kong (17.0 ± 10.0%) was significantly higher than the average relative abundances of VC1 in samples from Denver (1.1 ± 1.9%), London (12.3 ± 8.5%), Oslo (5.3 ± 5.9%), and Stockholm (10.9 ± 7.1%) (Mann–Whitney test, $p < 0.001$). For consistency, the family-level taxonomy derived from the clustering method was applied to all vOTUs in subsequent analyses.

Airborne viruses possessed diverse functions and ARGs

To investigate the functional capabilities of airborne viruses, gene annotation was performed on all vOTUs of medium quality or higher. This revealed that a significant proportion of the 318,966 genes (72.6%) did not have a match in any of the five reference databases, suggesting that these genes may have novel functions (Fig. 2A and B). To improve the functional annotation of the vOTUs, all viral genes were clustered into 157,870 viral gene clusters. The largest cluster contained 450 genes and 37.1% of clusters consisted of singletons (Fig. 2C). Like the taxonomy results, the accumulation curve of viral gene clusters and viral genes was unsaturated (Fig. 2D). Across all cities, the most common viral gene functions were related to DNA binding/regulation (e.g., LAGLIDADG endonuclease and phage integrase family), followed by transporters (e.g., ABC transporter) (Fig. 2E). Other prevalent functions included lysis (e.g., phage portal), structural (e.g., phage capsid), and replication (e.g., DNA polymerase type B), all associated with key viral signatures.

While previous studies have found ARGs in viruses extracted from the human gut [23], fresh water [83], and soil [84], their prevalence in airborne viruses remains unclear. A search in three ARG databases identified 326 unique ARGs from 159 vOTUs in samples across all cities, 74% of which were associated with high-quality vOTUs (Fig. S8A, S8B, and Table S8). In terms of ARG type, the largest number of ARGs in samples from all cities included 80 genes conferring resistance to mercury, 64 genes conferring resistance to glycopeptide antibiotics, and 24 genes conferring resistance to disinfecting agents and antiseptics. Among all ARG subtypes, the most abundant were *merP* and *merR*, which confer resistance to mercury and had an average relative abundance of 0.7 ± 1.3% in samples across all cities, followed by *mdeA* and *adeF*, which confer resistance to fluoroquinolone antibiotics and had average relative abundances of 0.5 ± 1.6% and 0.5 ± 1.5%, respectively, in samples across all cities (Fig. S8C). These four ARG subtypes were most abundant in samples from Hong Kong, with an average relative abundance of 1.8 ± 2.1%, while the *msrA* subtype (conferring multidrug resistance) was exclusively found in samples from New York City, with an average relative abundance of 1.3 ± 3.8%. The *arnA* subtype, which confers resistance to polymyxin (an antibiotic considered the last line of defense against bacterial multidrug resistance [85]), was detected only in a singular vOTU occurring in samples from Hong Kong, but its average relative abundance was low, i.e., 0.06 ± 0.3% (Fig. S8C).

Strong correlations between virus and host abundances were influenced by viral lifestyle

The survival and persistence of viruses are closely tied to their microbial hosts [10]. To shed light on virus–host interactions and coevolution mechanisms in BEs, both ex-situ and in-situ virus–host links were examined to maximize host assignment. Out of all vOTUs, ~ 35.8% (1915 vOTUs) could be linked to an ex-situ host, while ~ 20.1% (1073 vOTUs) could be linked to an in-situ host. Among the predicted in-situ virus–host links ($n = 1109$), the vast majority (95.6%) were identified through the genome region matching method ($n = 1060$), with only 49 links derived from the CRISPR spacer matching method (Table S9). In samples from all cities, the most commonly predicted ex-situ

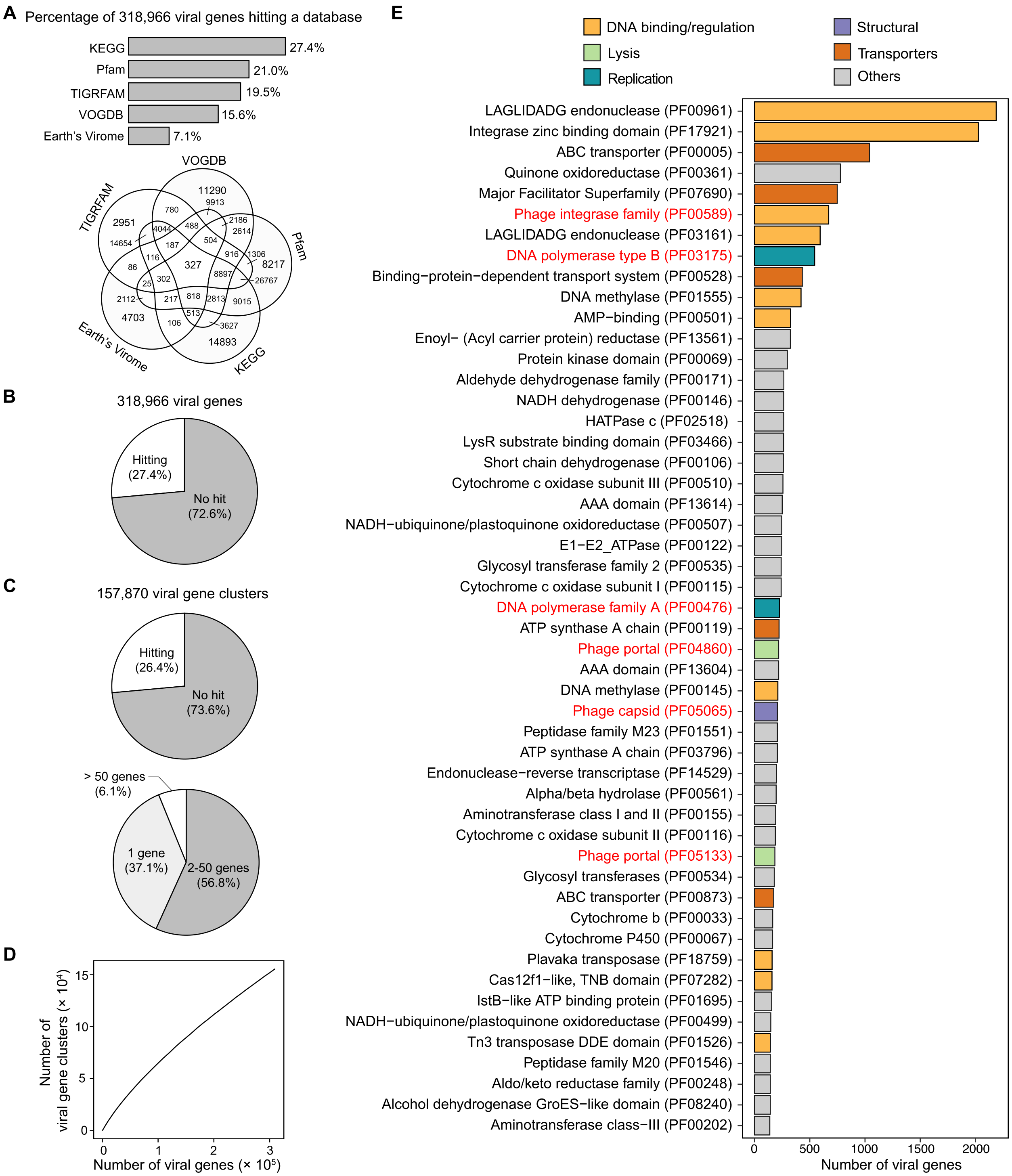


Fig. 2 Functional landscape of the viral operational taxonomic units (vOTUs) in air samples from built environments in global cities. **A** Percentage and number of viral genes identified and shared between the five reference databases. **B** Percentages of viral genes that could and could not be matched, respectively, to any of the five reference databases. **C** Percentages of viral gene clusters that could and could not be matched, respectively, to any of the five reference databases, and the size distribution of viral gene clusters based on the number of genes. **D** Accumulation curve of viral gene clusters. **E** Functional annotation of viral genes according to the Protein Families (Pfam) database. Only the 50 most common viral functions, based on the number of genes matched to the Pfam database, are shown. Gene functions associated with key viral signatures are highlighted in red

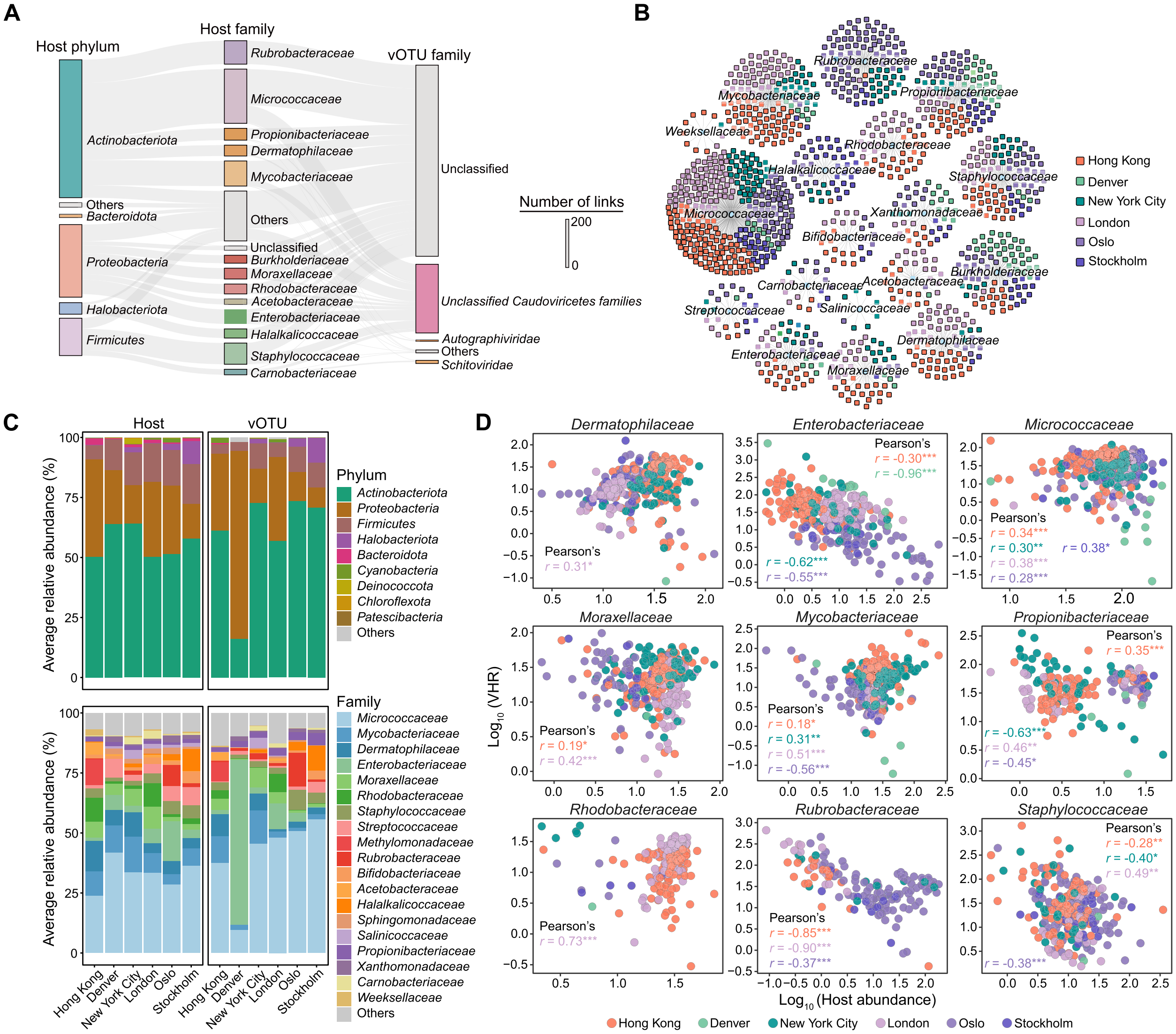


Fig. 3 In-situ virus–host links and correlations between the abundances of viral operational taxonomic units (vOTUs) and hosts in air samples from built environments in global cities. **A** Predicted hosts at the phylum and family levels for vOTUs at the family level in samples from all cities. The length of a bar indicates the number of virus–host links. **B** Network diagram showing the in-situ virus–host links in samples from each city at the host-family level. The central nodes represent the hosts, and the surrounding nodes represent the vOTUs linked to those hosts. The same virus–host links may appear in multiple cities, depending on the abundance of the virus and host. **C** Average relative abundances of vOTUs (right panel) and their linked hosts (left panel) at the phylum and family levels in samples from each city. The bars are colored according to the taxonomy of the linked hosts. **D** Pearson's correlations between host abundances and virus-to-host abundance ratios (VHRs) for six families of hosts in samples from each city. The families shown were present in greater than or equal to 150 samples across all cities, and there were significant correlations between their abundances as hosts and VHRs in samples from at least one city. Only the coefficients and *p*-values that were significant in samples from a given city are shown. *$p<0.05$, **$p<0.01$, ***$p<0.001$

hosts were from the *Mycobacteriaceae* family (number of links = 695) (Fig. S9), while the most commonly predicted in-situ hosts were from the *Micrococcaceae* family (number of links = 225) (Fig. 3A and Table S9). However, the most commonly predicted in-situ hosts in samples differed between cities. Specifically, the *Micrococcaceae* family were the most commonly predicted in-situ hosts in samples from Hong Kong, London, and New York City (numbers of links = 102, 73, and 41, respectively). Meanwhile, the *Burkholderiaceae*, *Rubrobacteraceae*, and *Halalkalicoccaceae* families were the most commonly predicted in-situ hosts in samples from Denver, Oslo, and Stockholm (numbers of links = 31, 81, and 26, respectively) (Fig. 3B and Fig. S10).

To further elucidate the relationships between airborne viruses and their linked hosts, the correlations between the abundances of vOTUs and their linked in-situ hosts were analyzed (Fig. 3C, D and Fig. S11). The average relative abundances of vOTUs that infected hosts from the *Micrococcaceae* family were highest in samples from Hong Kong (35.0 ± 17.0%), London (44.4 ± 16.7%), New York City (42.3 ± 21.8%), Oslo (42.4 ± 26.9%), and Stockholm (52.9 ± 25.4%). However, the average relative abundances of vOTUs that infected hosts from the *Enterobacteriaceae* family were highest in samples from Denver (75.2 ± 22.7%) (Fig. 3C). There were significant correlations between host abundances and virus-to-host abundance ratios (VHRs) in most host families (or phyla) in samples from most cities, although the strength of the correlations varied (Fig. 3D and Fig. S11). For example, there were significant and negative correlations between host abundances and VHRs in the *Enterobacteriaceae* family in samples from Denver (Pearson's $r = -0.96$, $p = 2.0 \times 10^{-6}$), Hong Kong (Pearson's $r = -0.30$, $p = 1.5 \times 10^{-3}$), New York City (Pearson's $r = -0.62$, $p = 2.2 \times 10^{-5}$), and Oslo (Pearson's $r = -0.55$, $p = 2.3 \times 10^{-8}$). However, there were significant and positive correlations between host abundances and VHRs in the *Micrococcaceae* family in samples from all cities (Pearson's $r > 0.28$, $p < 0.032$), except for Denver. Furthermore, the average proportions of vOTUs with either a temperate or virulent lifestyle varied among lineage-specific virus-to-host links (Fig. 4A).

To further investigate the influence of viral lifestyle on microbial abundance, we analyzed the correlations between virus-to-microbe abundance ratios and

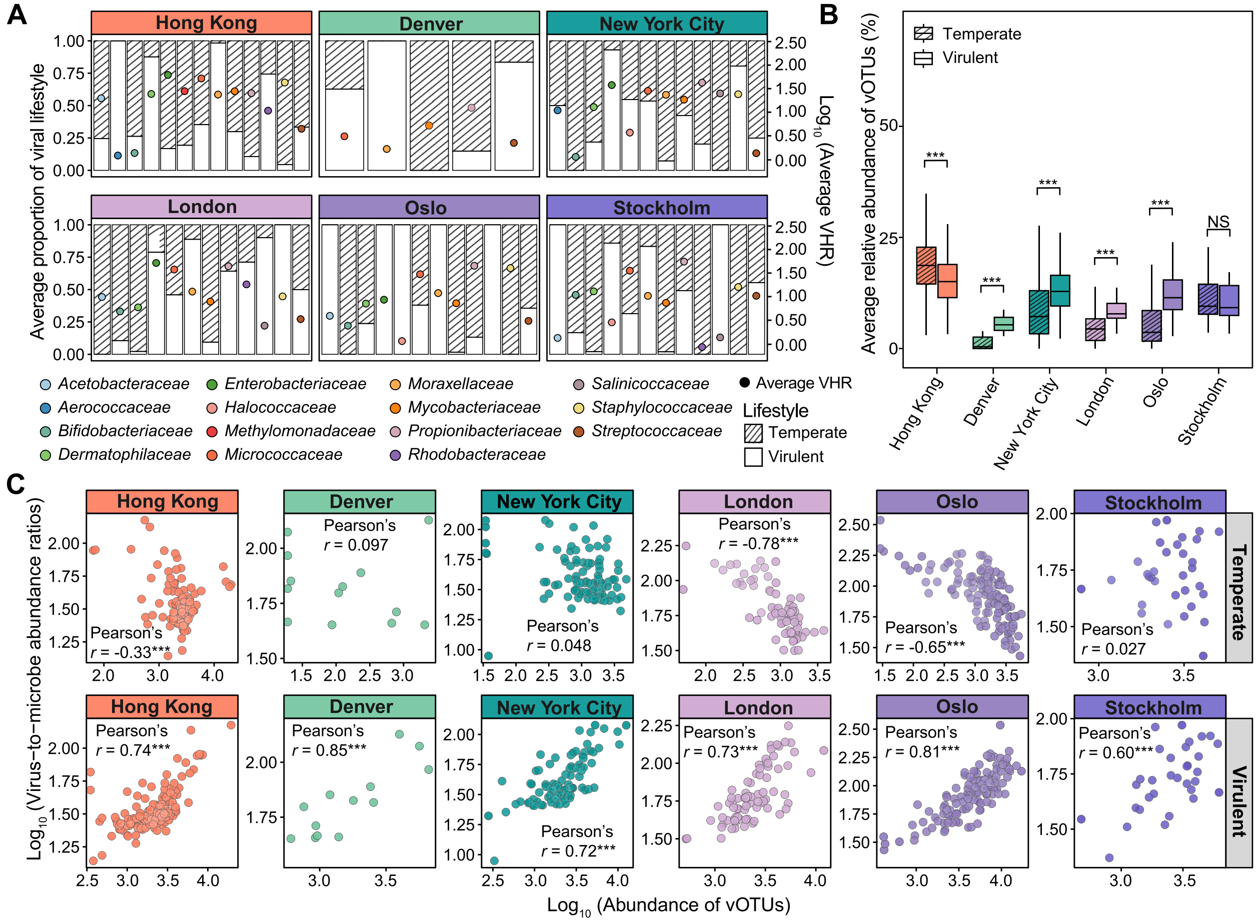


Fig. 4 Associations between the viral lifestyle and host or microbe abundance in air samples from built environments in global cities. **A** Proportions of viral operational taxonomic units (vOTUs) with a predicted virulent or temperate lifestyle (left axis) and the average virus-to-host abundance ratios (VHRs) in different host families in samples from each city (right axis). The proportion of vOTUs with a virulent or temperate lifestyle was calculated by dividing the abundance of vOTUs with a virulent or temperate lifestyle by the total abundance of vOTUs with either a virulent or temperate lifestyle. The vOTUs without a classified lifestyle and their corresponding linked hosts were excluded. **B** Average relative abundance of vOTUs with a predicted virulent and temperate lifestyle in samples from each city. Mann–Whitney tests were performed to assess the difference in the average relative abundance of vOTUs with different lifestyles in samples from each city regardless of the year. **C** Pearson correlations between virus-to-microbe abundance ratios and the abundance (RPKM) of vOTUs predicted to have a temperate (upper) or virulent (lower) lifestyle across samples from each city. *NS*: not statistically significant; ***$p < 0.001$

the abundance of vOTUs predicted as virulent (33.5%, $n = 1791$) or temperate (12.8%, $n = 687$), excluding eukaryotic viruses, across all vOTUs and rMAGs in samples from each city. In samples from most cities (all but Hong Kong and Stockholm), the average relative abundances of vOTUs with a virulent lifestyle were significantly higher than those of vOTUs with a temperate lifestyle (Mann–Whitney test, $p < 0.001$; Fig. 4B). Furthermore, in samples from all cities, the abundances of vOTUs with a virulent lifestyle were significantly and positively correlated with virus-to-microbe abundance ratios (Pearson's $r = 0.60$ to 0.85, $p < 1.0 \times 10^{-4}$), and in samples from Hong Kong, London, and Oslo, the abundances of vOTUs with a temperate lifestyle were significantly and negatively correlated with virus-to-microbe abundance ratios (Pearson's $r = -0.78$ to -0.33, $p < 3.2 \times 10^{-5}$) (Fig. 4C).

Evidence of CRISPR–Acr interactions between airborne viruses and microbial hosts

Viruses and their microbial hosts can evolve their respective CRISPR–Cas systems and Acr systems to serve as countermeasures against each other [14]. To investigate CRISPR–Acr interactions, CRISPR spacers were extracted from all 686 rMAGs in samples across all cities, resulting in the identification of 1988 CRISPR spacers from 155 rMAGs. However, only 52 (~ 2.6%) of the CRISPR spacers could be linked to a vOTU. These CRISPR spacers were predominantly derived from the *Micrococcaceae* (39 rMAGs), *Rubrobacteraceae* (16 rMAGs), and *Halalkalicoccaceae* (14 rMAGs) families. Furthermore, only 53 (~ 7.7%) of the rMAGs were found to carry a CRISPR–Cas system, and a large proportion ($n = 22$) of these rMAGs belonged to the *Micrococcaceae* family (Table S10). The 61 identified CRISPR–Cas systems encompassed four CRISPR types (I, II, III, and V) and were associated with diverse Cas proteins, with type I being the most prevalent ($n = 38$) (Fig. S12A). Most of the identified CRISPR–Cas systems ($n = 29$) were present in samples from all cities, while only some ($n = 9$) were unique to individual cities (Fig. S12B). Specifically, in terms of numbers of city-specific CRISPR–Cas systems, samples from Oslo had four, those from Stockholm had two, those from New York City had two, and those from Hong Kong had one.

To counter the CRISPR–Cas systems of hosts, viruses have evolved Acr proteins [14]. One hundred fifty-five Acr proteins were identified in 79 vOTUs (Table S11), indicating that only ~ 0.015% of all vOTUs encoded a detectable variant of this form of countermeasure against CRISPR–Cas systems. Maps based on the reference Acr protein databases revealed that the Acr proteins of airborne viruses had the ability to counteract various types of CRISPR–Cas systems. Specifically, 53 Acr proteins belonged to subtype II-C, 46 belonged to subtype I-B, and 44 belonged to subtype I-C. The broad distribution of these Acr proteins in a maximum likelihood phylogenetic tree suggested the presence of diverse subtypes (Fig. S12C). Only eight Acr proteins were present in samples from all cities, as there were a significant number of city-specific Acr proteins in samples from Hong Kong ($n = 30$), Oslo ($n = 27$), New York City ($n = 20$), and London ($n = 16$) (Fig. S12D). Of the vOTUs carrying Acr proteins, eight could be linked to a host through CRISPR spacer sequences. For example, a vOTU (S5C4_18hk) and its linked host (rMAG SL467177_bin.10) carried a subtype I-E Acr protein and a subtype I-E CRISPR–Cas system, respectively (Fig. 5A).

Phage-encoded AMGs and their potential impact on host fitness in BEs

Phage-encoded AMGs are important regulators of the metabolism of microbial hosts that enhance their adaptation to diverse environments [8]. In samples from all cities, a total of 1247 putative AMGs were identified in 281 high-quality and 230 medium-quality vOTUs (Table S12), 147 of which had an average relative abundance of 1.7 ± 4.3% and were linked to an in-situ host. Out of the vOTUs carrying putative AMGs, 274 (53.3%) were predicted to have a virulent lifestyle, while 164 (31.9%) were predicted to have a temperate lifestyle. The vOTUs carrying putative AMGs with a virulent lifestyle were primarily linked to hosts in the families *Micrococcaceae* (14 links) and *Mycobacteriaceae* (seven links), while the vOTUs carrying putative AMGs with a temperate lifestyle were primarily linked to hosts in the families *Micrococcaceae* (25 links) and *Enterobacteriaceae* (eight links). Among all the identified putative AMGs, 20% were classified as having miscellaneous functions, 15% were associated with organic nitrogen transformation, and 8.5% were associated with carbon utilization (Table S12). A putative AMG encoding a ribonucleotide reductase had the highest average relative abundance in samples across cities (1.3 ± 5.9%), followed by a putative AMG encoding a DNA (cytosine-5-)-methyltransferase (0.64 ± 2.4%) (Fig. S13A). Certain putative AMGs, such as those encoding aspartate carbamoyltransferase, 2-hydroxyacid dehydrogenase, and outer membrane channel protein TolC, were present only in one vOTU in samples obtained from Hong Kong, indicating the existence of city-specific AMGs (Fig. S13A).

Across samples from all cities, some virus-associated putative AMGs could potentially influence host metabolism and enhance host adaptability. In a sample from Hong Kong, a vOTU (vRhyme_2_S63Ck141_92927_18hk) carried six putative AMGs (i.e., large subunit ribosomal protein L28, cinnamic acid

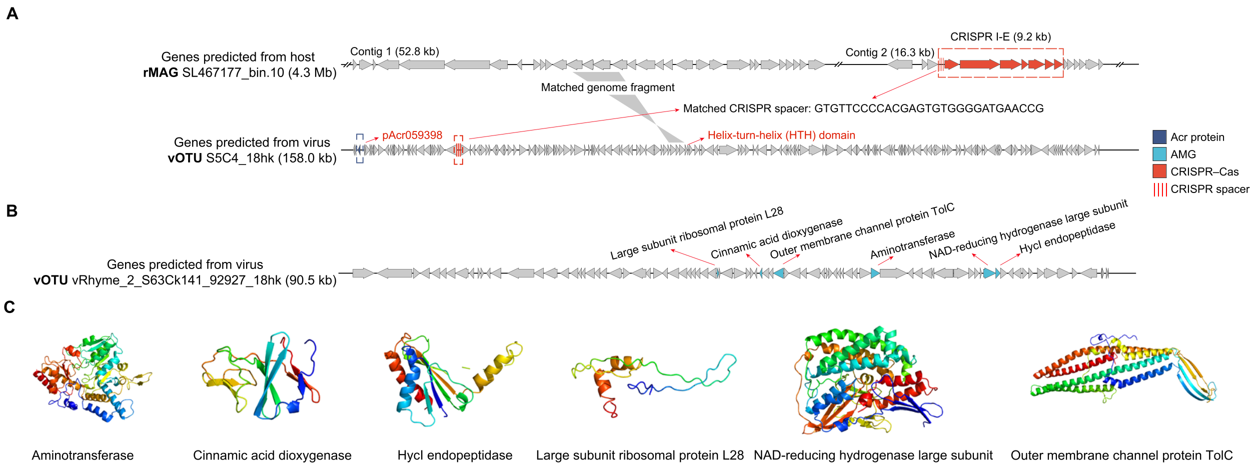

A
Genes predicted from host
rMAG SL467177_bin.10 (4.3 Mb)
Contig 1 (52.8 kb)
Contig 2 (16.3 kb)
CRISPR I-E (9.2 kb)
Matched genome fragment
Matched CRISPR spacer: GTGTTCCCCACGAGTGTGGGGATGAACCG
Genes predicted from virus
vOTU S5C4_18hk (158.0 kb)
pAcr059398
Helix-turn-helix (HTH) domain
🟦 Acr protein
🟦 AMG
🟥 CRISPR–Cas
|||| CRISPR spacer
B
Genes predicted from virus
vOTU vRhyme_2_S63Ck141_92927_18hk (90.5 kb)
Large subunit ribosomal protein L28
Cinnamic acid dioxygenase
Outer membrane channel protein TolC
Aminotransferase
NAD-reducing hydrogenase large subunit
HycI endopeptidase
C
Aminotransferase
Cinnamic acid dioxygenase
HycI endopeptidase
Large subunit ribosomal protein L28
NAD-reducing hydrogenase large subunit
Outer membrane channel protein TolC

Fig. 5 Evidence of virus–host interactions via the clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated (Cas) system, anti-CRISPR (Acr) proteins, and potential functions of putative auxiliary metabolic gene (AMG) in air samples from built environments in global cities. **A** Schematic diagram illustrating the linkage between a viral operational taxonomic unit (vOTU) and its host through CRISPR spacer matching. The Acr protein (pAcr059398) of the vOTU can evade the CRISPR–Cas system (subtype I-E) of its linked host. Three viral proteins were identified within the genome of the linked host. **B** Schematic diagram illustrating six putative AMGs encoded by a viral genome fragment (~33.7 kb). **C** High-confidence protein structures of the six putative AMGs indicated in panel **B**. The colors range from the N terminus to the C terminus in a rainbow pattern

dioxygenase, aminotransferase, outer membrane channel protein TolC, NAD-reducing hydrogenase large subunit, and HycI endopeptidase), and its predicted host belonged to the family *Methylomonadaceae* (SL469770_bin.18) (Fig. 5B). Furthermore, in samples from Denver, London, and New York City, a vOTU (vRhyme_6811_S78Ck141_122871_19ny) connected to a putative host (SL470488_bin.1) from the family *Enterobacteriaceae* carried a putative AMG encoding a glucosyltransferase family 2 cellulose synthase, involved in cell wall synthesis. Additionally, in samples from all cities, a vOTU (S58C2462_18ny) with a putative AMG encoding a multicopper oxidase involved in copper tolerance was linked to a host (SL469684_bin.3) from the family *Xanthomonadaceae* (Fig. S13B). High-confidence protein structure predictions (Fig. 5C and Fig. S13C) supported the potential functions of these putative AMGs, with confidence levels ranging from 82 to 100% (Table S13).

Discussion

BEs are characterized by low nutrient availability and fluctuating physical and environmental conditions and yet they support a diverse community of microbes, including bacteria, fungi, and viruses [1]. Airborne viruses in BEs have received less research attention than airborne bacteria and fungi in BEs [5]. Thus, there is a limited understanding of airborne viruses' distribution, diversity, functions, and interactions with bacterial hosts in BEs in different global locations. This knowledge gap prompted us to collect 503 metagenomic air samples from public transit systems in six cities across three continents over a two-summer period and subsequently analyze these samples. Through metagenomic binning methods, we identified 5346 vOTUs in the samples, including city-specific vOTUs that reflected biogeographical differences in viral community compositions and diversity. Moreover, despite the oligotrophic nature of air in BE, the Shannon indices of the viral communities indicated that they had a high level of diversity. This diversity varied significantly across continents, with lower levels in North America, likely influenced by differences in meteorological conditions, environmental factors, and transit system design and operation [6]. Mapping the vOTUs against the IMG/VR database and clustering them with reference viral genomes revealed a substantial proportion of taxonomically unclassified vOTUs at the species and family levels, suggesting that there are many novel viruses within the air of BEs, similar to other ecosystems [23, 41, 53]. Human-associated *papillomaviruses* [86] were detected in samples from all cities, consistent with these viruses' prevalence in venues with high occupancy [5]. Additionally, viruses from the *Caulimoviridae* family, a group within *Riboviria* characterized by a DNA phase in their life cycle [82], were detected from the air samples, aligning with previous reports of *Caulimoviridae* in outdoor urban air [87]. *Caulimoviridae* viruses were found to be ubiquitous and dominant in samples from all cities, unlike in other ecosystems such as the human gut [23], soil [53] and marine environments [73], and BE surfaces (e.g., bollards, doorknobs, and floors) [41], where viruses of the *Siphoviridae* and *Myoviridae* families are commonly abundant. The distinct viral taxonomies in air samples from BEs and the variations in virus abundance between air samples from BEs in different cities likely stem from BEs' diverse microbial sources, such as human skin, oral microbiota, and pets [5, 88].

The characteristics and dynamics of physical environments have a significant influence on the interactions and functions of viruses with microbial hosts over extended periods [89]. In the current study, although the diversity of viral functions did not reach saturation, the predominant gene clusters in airborne viruses in samples exhibited a significant capacity for infection, replication, and integration within oligotrophic air environments. For example, genes encoding integrases and LAGLIDADG endonucleases were prevalent in airborne viruses, suggesting a high potential for these proteins to facilitate adaptation, survival, and propagation in BEs. This contrasts with the primary functions of viral proteins found in human gut viruses (e.g., packing and assembly functions) [23] and surface-associated BE viruses (e.g., transport functions) [41].While previous studies have suggested that ARGs are rarely found in phages [90, 91], the current study identified diverse types of ARGs within the genomes of airborne viruses. Specifically, there were abundant ARGs that confer resistance to antibiotics commonly used to treat upper respiratory or skin infections (e.g., fluoroquinolones) [92] and to disinfecting agents commonly used in air fresheners or surface disinfectants (e.g., triclosan) [93]. There were also variations in the spatial distribution of ARGs in samples across cities, which is likely partly attributable to differences in antibiotic usage and to the microbial compositions unique to each city [94]. Of particular concern was the detection of the *arnA* subtype in viruses in samples from Hong Kong, albeit at a low relative abundance, which coincided with increased usage of polymyxin [95], a last-resort antibiotic [85]. These findings indicate that airborne viruses may serve as reservoirs of ARGs and possibly contribute to the dissemination of antibiotic resistance within BEs [96].

Virus–host interactions play a crucial role in the coevolution of viruses and their linked hosts within microbial communities [97]. In this study, microbes from the *Mycobacteriaceae* family were the most frequently predicted ex-situ hosts, while they ranked second as predicted in-situ hosts of airborne viruses. This discrepancy is likely due to differences between the rMAGs recovered in this study and the public genomes used for ex-situ predictions. Among the diverse in-situ microbial hosts identified, the *Micrococcaceae* family was the most commonly predicted host. Given the abundance of *Micrococcaceae* in the air of BEs, it is reasonable to posit that this family has a higher probability of being infected by viruses than other host families, thus leading to the generation of virus–host links [98]. However, there were geographical variations in virus–host interactions, namely significant between-city differences in the most commonly predicted host family of virus–host links in samples. This was likely due to diverse ecological and environmental factors [99]. VHRs, which reflect the dynamics of virus–host interactions, varied significantly between different taxa of virus–host links in samples and between samples from different cities, potentially due to the interplay between virulent and temperate viral lifestyles [7]. Among vOTUs with a predicted lifestyle, virulent airborne viruses were more prevalent than temperate ones in samples from most cities, with virulent infections often occurring at low microbial abundances, while temperate infections increased as microbial abundances increased. This suggests that in the oligotrophic conditions typical of BEs, characterized by limited microbial growth [28] and low airborne microbial concentrations [2], airborne viruses may have a greater likelihood of replicating their genomes and lysing host cells upon infection, potentially releasing new viral particles and supporting their continued replication [24]. Conversely, when microbial abundances increase to higher levels, possibly due to improvements in environmental conditions that support the growth of microbial hosts, airborne viruses tend to adopt a temperate lifestyle and utilize host propagation for replicating their viral genomes. The above-described results highlight that virulent infection is a crucial survival strategy for the majority of airborne viruses under oligotrophic conditions in BEs, where carbon sources are extremely limited [100]. In contrast, in other ecosystems (e.g., soil [53], marine [101], and wastewater treatment ecosystems [102]), virulent infection by viruses is more prevalent under conditions of high nutrient availability.

Viruses with a virulent lifestyle are predominant in BEs, as these viruses can efficiently exploit the cellular machinery of their hosts for the production of viral particles, likely resulting in the rapid lysis of infected host cells [103] and preventing the vertical transfer of CRISPR–Cas systems to subsequent generations [104]. Furthermore, microbial hosts primarily acquire CRISPR spacers from non-virulent viruses [105], which rely on the abundance of phage sequences within host cells [106]. As a result, the evolution of CRISPR–Cas systems in BE hosts may be constrained by the low abundance of airborne viruses [107] and the high abundance of virulent viruses. In addition, the low concentration of airborne microbes in BEs means that there is a low virus–host encounter rate [98], which may limit virus–host interactions. Moreover, the constant movement of air in BEs disfavors prolonged contact between viruses and hosts, reducing the probability of coevolution of host CRISPR–Cas systems and viral Acr proteins [13, 14]. The aforementioned factors, along with the incompleteness of both viral and host genomes, likely explain why only a small proportion of vOTUs (~0.92%) could be linked to a host through the CRISPR spacers extracted from the rMAGs, with this proportion being substantially lower than that in the human gut (~81%) [23], where contacts between viruses and hosts are likely to be frequent and persistent. Certain airborne viruses in BEs can evade CRISPR–Cas systems by evolving Acr proteins [14], but the majority of airborne viruses lack these proteins, likely because Acr proteins are primarily derived from temperate phages [108], whereas the predominant airborne viruses in BEs were predicted to exhibit a virulent lifestyle. However, it is possible that the limitations of current Acr databases have contributed to the incomplete detection of these proteins. Additionally, viruses may employ various strategies (e.g., point mutations [109] and large-scale deletions [110]) to minimize phage resistance in microbial hosts and ensure successful viral infections in BEs. The limited coevolution of CRISPR–Cas systems and Acr proteins in oligotrophic BEs may benefit hosts by reducing energetic costs through the downregulation of CRISPR–Cas gene expression [111], thereby enhancing hosts' survival and adaptive capabilities.

Furthermore, viruses can contribute to host adaptation by expressing AMGs upon invading hosts [24]. While these AMGs may enhance virion production during virulent infections, those encoded by temperate viruses are more likely to improve host fitness within ecosystems [112], potentially facilitating the coexistence of viruses and hosts. Moreover, in the current study, virus-associated putative AMGs encoding enzymes such as cinnamic acid dioxygenase and aminotransferase may enhance the host's ability to utilize substrates such as amino acids, which are likely present in low concentrations in BEs [100], thus facilitating phage production [11]. Furthermore, AMGs encoded by airborne viruses may enhance the resistance of microbial hosts to adverse conditions in BEs. For example, the outer membrane channel protein TolC could aid in the removal of toxic compounds, including antimicrobial agents [113], while multicopper oxidase could enhance host resistance to heavy metals [114]. Considering the potential for exposure to various sources of virulence in BEs, such as antimicrobial agents and heavy metals [115, 116], these phage-encoded putative AMGs could aid the survival of their microbial hosts, highlighting that coexistence strategies may be employed by viruses and hosts in BEs as they adapt to and thrive in their environment.

This study sheds light on the viruses found in air samples collected from BEs across six global cities. However, several methodological limitations in viral identification and annotation warrant further investigation. Although default parameters or those from previous studies were used in the bioinformatic workflow, systematically exploring alternative settings could help reduce false positives and false negatives in vOTU identification, improving the accuracy of viral detection [40, 66]. Caution should be exercised when interpreting results from environmental metagenomes, as technical and analytical biases may affect the detection, classification, and quantification of viral populations. Despite the substantial number of air samples analyzed, the workflow allowed for the discovery of only a limited number of vOTUs. To achieve a more comprehensive understanding of viral compositions and diversity, it is crucial to improve the efficiency of sampling, increase the number of samples, and enhance the sequencing depth. Furthermore, classifying vOTUs at the species and family levels and determining viral lifestyles remain challenging due to limitations in current viral databases and analysis tools, as detailed in recent studies [59, 117, 118]. Optimizing bioinformatic approaches for viral genome recovery can improve genome completeness, enabling more accurate taxonomic identification, gene annotation (e.g., AMGs and Acr proteins), provirus detection, and host prediction. Additionally, analyzing bulk metagenomes without viral enrichment may have introduced vOTUs not representative of free viruses in air samples. Incorporating enrichment techniques could improve specificity and more accurately profile of airborne viral communities. Including airborne RNA viruses would also provide a more holistic understanding of airborne viral communities in BEs.

Beyond methodological challenges, the biogeographic patterns of global airborne viruses and the diversity and dynamics of viruses in BEs require closer examination. This could be achieved by improving the evenness of sample sizes between sampling sites, expanding the sampling scope to include additional sites and time points in global cities, and investigating environmental factors to identify potential drivers of viral diversity and dynamics. Culture-based experiments in controlled settings could further validate the field observations regarding viral lifestyles, including the switching between temperate and virulent viruses, as well as their interactions with hosts and variations in abundance within airborne environments. Finally, studying the conditions that trigger virulent or temperate infections and the expression of phage-encoded AMGs in host genomes would enhance our understanding of virus–host dynamics in BEs.

In summary, this study revealed the distinct biogeographical characteristics of airborne viruses in oligotrophic BEs and identified previously unknown viral clades. The findings indicate significant global variations in viral diversity, composition, functional potential, abundance, and host interactions within airborne viromes in BEs. The presence of many poorly classified viruses underscores the need for further investigations to understand these viruses' novel biological characteristics and taxonomies and to explore their potential impact on airborne microbial communities. The study also highlighted that airborne viruses predominantly adopt a virulent lifestyle and that they are closely associated with host abundance, suggesting the presence of coevolutionary strategies and coexistence mechanisms in oligotrophic BEs. These findings enhance our understanding of viromes in the air of BEs, highlighting their unique characteristics and survival strategies compared with those found in more nutrient-rich ecosystems.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-025-02173-z.

Supplementary Material 1: Supplementary Figures S1–S13.

Supplementary Material 2: Supplementary Tables S1–S13.

Acknowledgements

The authors are grateful to all those who supported the fieldwork in all cities.

Authors' contributions

H.L. performed data analysis and data interpretation, and wrote the manuscript. S.D., M.H.Y.L., and X.T. provided advice on data interpretation and bioinformatic support. K.O.B. processed all the samples. D.C.D., D.B., and K.R. supervised the sequencing and performed quality control on the initial data. W.L.C., D.B., D.J.B., D.C.D., D.C.G., M.T.H., F.J.K., A.G.L., C.M., M.N.C., K.R., B.T., K.I.U., and B.G.Y. contributed samples and provided advice on data interpretation. C.E.M. and M.D. conceived and designed the study. P.K.H.L. conceived the study and supervised the research. All authors have read and approved the final manuscript.

Funding

This research was supported by the Hong Kong Research Grants Council through the Research Impact Fund (R1016-20F) and the General Research Fund (11214721 and 11206224). This research was also supported by Igor Tulchinsky and the WorldQuant Foundation, the US National Institutes of Health (R01AI151059 and U01DA053941), the Bill and Melinda Gates Foundation (OPP1151054), and the Stockholm Health Authority (SLL 20160933).

Data availability

The raw DNA-sequencing data used in this study have been deposited in the NCBI Sequence Read Archive under BioProject accession numbers PRJNA1129830 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA1129830?reviewer=1g7d3ssamqkhspagcpf2cenq6t) and PRJNA1132165 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA1132165?reviewer=uu8vu4rqqn6vlfp122spfju3g0). The viral genome sequences are available at https://figshare.com/articles/dataset/5346_highquality_airborne_virus/27628491. The supporting code can be found at the following GitHub page: https://github.com/HuaxinLEI-CityU/Global-Built-Environments-Airborne-Viromes.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

C.E.M. is a co-Founder of Biotia, Inc. The other authors declare that they have no competing interests.

Author details

[1]School of Energy and Environment, City University of Hong Kong, Kowloon, Hong Kong SAR, China. [2]Department of Biological Sciences, School of Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, P.R. China. [3]Total Defence Division, Norwegian Defence Research Establishment FFI, Kjeller, Norway. [4]Department of Analytical, Environmental and Forensic Sciences, King's College London, London, UK. [5]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [6]Environmental Research Group, MRC Centre for Environment and Health, Imperial College London, London, UK. [7]NIHR HPRU in Environmental Exposures and Health, Imperial College London, London, UK. [8]Environmental Engineering Program, College of Engineering and Applied Science, University of Colorado, Boulder, CO, USA. [9]Department of Biological Sciences, University of Idaho, Moscow, ID, USA. [10]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden. [11]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [12]The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA. [13]The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. [14]School of Energy and Environment and State Key Laboratory of Marine Pollution, City University of Hong Kong, Kowloon, Hong Kong SAR, China. [15]Low-Carbon and Climate Impact Research Centre, City University of Hong Kong, Kowloon, Hong Kong SAR, China.

Received: 21 May 2025 Accepted: 13 June 2025
Published online: 29 August 2025

References

1. Gilbert JA, Stephens B. Microbiology of the built environment. Nat Rev Microbiol. 2018;16(11):661–70.
2. Prussin AJ 2nd, Garcia EB, Marr LC. Total virus and bacteria concentrations in indoor and outdoor air. Environ Sci Technol Lett. 2015;2(4):84–8.
3. Whon TW, Kim MS, Roh SW, Shin NR, Lee HW, Bae JW. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. J Virol. 2012;86(15):8221–31.
4. Prussin AJ 2nd, Belser JA, Bischoff W, Kelley ST, Lin K, Lindsley WG, et al. Viruses in the Built Environment (VIBE) meeting report. Microbiome. 2020;8(1):1.
5. Rosario K, Fierer N, Miller S, Luongo J, Breitbart M. Diversity of DNA and RNA viruses in indoor air as assessed via metagenomic sequencing. Environ Sci Technol. 2018;52(3):1014–27.
6. Prussin AJ 2nd, Torres PJ, Shimashita J, Head SR, Bibby KJ, Kelley ST, et al. Seasonal dynamics of DNA and RNA viral bioaerosol communities in a daycare center. Microbiome. 2019;7(1):53.
7. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobian-Guemes AG, et al. Lytic to temperate switching of viral communities. Nature. 2016;531(7595):466–70.
8. Brown TL, Charity OJ, Adriaenssens EM. Ecological and functional roles of bacteriophages in contrasting environments: marine, terrestrial and human gut. Curr Opin Microbiol. 2022;70:102229.
9. Chen T, Liu T, Wu Z, Wang B, Chen Q, Zhang M, et al. Virus-pathogen interactions improve water quality along the Middle Route of the South-to-North Water Diversion Canal. ISME J. 2023;17(10):1719–32.
10. Zimmerman AE, Howard-Varona C, Needham DM, John SG, Worden AZ, Sullivan MB, et al. Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. Nat Rev Microbiol. 2020;18(1):21–34.
11. Hurwitz BL, U'Ren JM. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol. 2016;31:161–8.
12. Chevallereau A, Pons BJ, van Houte S, Westra ER. Interactions between bacterial and phage communities in natural environments. Nat Rev Microbiol. 2022;20(1):49–62.

13. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327(5962):167–70.
14. Li Y, Bondy-Denomy J. Anti-CRISPRs go viral: The infection biology of CRISPR-Cas inhibitors. Cell Host Microbe. 2021;29(5):704–14.
15. Nelson AR, Narrowe AB, Rhoades CC, Fegel TS, Daly RA, Roth HK, et al. Wildfire-dependent changes in soil microbiome diversity and function. Nat Microbiol. 2022;7(9):1419–30.
16. Johansen J, Atarashi K, Arai Y, Hirose N, Sorensen SJ, Vatanen T, et al. Centenarians have a diverse gut virome with the potential to modulate metabolism and promote healthy lifespan. Nat Microbiol. 2023;8(6):1064–78.
17. Shi LD, Dong X, Liu Z, Yang Y, Lin JG, Li M, et al. A mixed blessing of viruses in wastewater treatment plants. Water Res. 2022;215:118237.
18. Munch PC, Franzosa EA, Stecher B, McHardy AC, Huttenhower C. Identification of natural CRISPR systems and targets in the human microbiome. Cell Host Microbe. 2021;29(1):94-106 e4.
19. Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. Nat Commun. 2023;14(1):7318.
20. Yuan L, Ju F. Potential auxiliary metabolic capabilities and activities reveal biochemical impacts of viruses in municipal wastewater treatment plants. Environ Sci Technol. 2023;57(13):5485–98.
21. Zhang J, Tang A, Jin T, Sun D, Guo F, Lei H, et al. A panoramic view of the virosphere in three wastewater treatment plants by integrating viral-like particle-concentrated and traditional non-concentrated metagenomic approaches. iMeta. 2024;3(3):e188.
22. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat Microbiol. 2020;5(2):265–71.
23. Nayfach S, Paez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6(7):960–70.
24. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun. 2017;8:15955.
25. Yunha Hwang JR, Dirk Schulze-Makuch, Michael Schloter, Alexander J. Probst diverse viruses carrying genes for microbial extremotolerance in the Atacama Desert hyperarid soil. mSystems. 2021;6(3):e00385–21.
26. Danko D, Bezdan D, Afshin EE, Ahsanuddin S, Bhattacharya C, Butler DJ, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell. 2021;184(13):3376-93 e17.
27. Boifot KO, Gohli J, Moen LV, Dybwad M. Performance evaluation of a new custom, multi-component DNA isolation method optimized for use in shotgun metagenomic sequencing-based aerosol microbiome research. Environ Microbiome. 2020;15(1):1.
28. Leung MHY, Tong X, Boifot KO, Bezdan D, Butler DJ, Danko DC, et al. Characterization of the public transit air microbiome and resistome reveals geographical specificity. Microbiome. 2021;9(1):112.
29. Krueger F. TrimGalore: a wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. Babraham Institute. 2015. Available from: https://github.com/FelixKrueger/TrimGalore. [accessed 21 March 2025].
30. Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, et al. The DNA sequence and biology of human chromosome 19. Nature. 2004;428(6982):529–35.
31. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6(1):158.
32. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.
33. Lu J, Breitwieser FP, Thielen P, and Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science. 2017;3:e104.
34. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6(1):226.
35. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
36. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
37. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
38. Nissen JN, Johansen J, Allesoe RL, Sonderby CK, Armenteros JJA, Gronbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol. 2021;39(5):555–60.
39. Johansen J, Plichta DR, Nissen JN, Jespersen ML, Shah SA, Deng L, et al. Genome binning of viral entities from bulk metagenomics data. Nat Commun. 2022;13(1):965.
40. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020;8(1):64–77.
41. Du S, Tong X, Lai ACK, Chan CK, Mason CE, Lee PKH. Highly host-linked viromes in the built environment possess habitat-dependent diversity and functions for potential virus-host coevolution. Nat Commun. 2023;14(1):2676.
42. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9(1):37.
43. Fan X, Ji M, Mu D, Zeng X, Tian Z, Sun K, et al. Global diversity and biogeography of DNA viral communities in activated sludge systems. Microbiome. 2023;11(1):234.
44. Kieft K, Adams A, Salamzade R, Kalan L, Anantharaman K. vRhyme enables binning of viral genomes from metagenomes. Nucleic Acids Res. 2022;50(14):e83.
45. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39(5):578–85.
46. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8(1):90.
47. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
48. Camargo AP, Nayfach S, Chen IA, Palaniappan K, Ratner A, Chu K, 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. 2023;51(D1):D733–43.
49. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18(4):366–8.
50. Nishijima S, Nagata N, Kiguchi Y, Kojima Y, Miyoshi-Akiyama T, Kimura M, et al. Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. Nat Commun. 2022;13(1):5252.
51. Shen W, Le S, Li Y, Hu F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11(10):e0163962.
52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
53. Liao H, Li H, Duan CS, Zhou XY, Luo QP, An XL, et al. Response of soil viral communities to land use changes. Nat Commun. 2022;13(1):6027.
54. Langwig MV, De Anda V, Dombrowski N, Seitz KW, Rambo IM, Greening C, et al. Large-scale protein level comparison of *Deltaproteobacteria* reveals cohesive metabolic groups. ISME J. 2022;16(1):307–20.
55. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11(12):2864–8.
56. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022;38(23):5315–6.
57. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 2022;50(D1):D785–94.
58. Zhao Y, Liu S, Jiang B, Feng Y, Zhu T, Tao H, et al. Genome-centered metagenomics analysis reveals the symbiotic organisms possessing ability to cross-feed with anammox bacteria in anammox consortia. Environ Sci Technol. 2018;52(19):11285–96.

59. Li Z, Pan D, Wei G, Pi W, Zhang C, Wang JH, et al. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. ISME J. 2021;15(8):2366–78.
60. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JA, et al. A network-based integrated framework for predicting virus-prokaryote interactions. NAR Genom Bioinform. 2020;2(2):lqaa044.
61. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol. 2018;3(8):870–80.
62. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
63. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36(7):2251–2.
64. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003;31(1):371–3.
65. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9.
66. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536(7617):425–30.
67. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7(10):e1002195.
68. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.
69. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. 2023;51(D1):D690–9.
70. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11(1):12728.
71. Yin X, Jiang XT, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. Bioinformatics. 2018;34(13):2263–70.
72. Yang Y, Jiang X, Chai B, Ma L, Li B, Zhang A, et al. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. Bioinformatics. 2016;32(15):2346–51.
73. Jian H, Yi Y, Wang J, Hao Y, Zhang M, Wang S, et al. Diversity and distribution of viruses inhabiting the deepest ocean on Earth. ISME J. 2021;15(10):3094–110.
74. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. 2020;48(16):8883–900.
75. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845–58.
76. Russel J, Pinilla-Redondo R, Mayo-Munoz D, Shah SA, Sorensen SJ. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. CRISPR J. 2020;3(6):462–9.
77. Yang B, Zheng J, Yin Y. AcaFinder: genome mining for anti-CRISPR-associated genes. mSystems. 2022;7(6):e00817-22.
78. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, et al. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome. 2021;9(1):78.
79. Wang J, Dai W, Li J, Li Q, Xie R, Zhang Y, et al. AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins. Nucleic Acids Res. 2021;49(D1):D630–8.
80. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):e9490.
81. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2003;30:3059–66.
82. Teycheney PY, Geering ADW, Dasgupta I, Hull R, Kreuze JF, Lockhart B, et al. ICTV virus taxonomy profile: caulimoviridae. J Gen Virol. 2020;101(10):1025–6.
83. Moon K, Jeon JH, Kang I, Park KS, Lee K, Cha CJ, et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. Microbiome. 2020;8(1):75.
84. Chen ML, An XL, Liao H, Yang K, Su JQ, Zhu YG. Viral community and virus-associated antibiotic resistance genes in soils amended with organic fertilizers. Environ Sci Technol. 2021;55(20):13881–90.
85. Nang SC, Azad MAK, Velkov T, Zhou QT, Li J. Rescuing the last-line polymyxins: achievements and challenges. Pharmacol Rev. 2021;73(2):679–728.
86. Duerkop BA, Hooper LV. Resident viruses and their interactions with the immune system. Nat Immunol. 2013;14(7):654–9.
87. Deng A, Wang J, Li L, Shi R, Li X, Wen T. Synoptic variation drives genetic diversity and transmission mode of airborne DNA viruses in urban space. Adv Sci (Weinh). 2024;11(46):e2404512.
88. Leung MH, Lee PK. The roles of the outdoors and occupants in contributing to a potential pan-microbiome of the built environment: a review. Microbiome. 2016;4(1):21.
89. Jansson JK, Wu R. Soil viral diversity, ecology and climate change. Nat Rev Microbiol. 2023;21(5):296–311.
90. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. ISME J. 2017;11(1):237–47.
91. Billaud M, Lamy-Besnier Q, Lossouarn J, Moncaut E, Dion MB, Moineau S, et al. Analysis of viromes and microbiomes from pig fecal samples reveals that phages and prophages rarely carry antibiotic resistance genes. ISME Commun. 2021;1(1):55.
92. Zhanel GG, Ennis K, Vercaigne L, Walkty A, Gin AS, Embil J, et al. A critical review of the fluoroquinolones: focus on respiratory tract infections. Drugs. 2002;62:13–59.
93. Lu J, Guo J. Disinfection spreads antimicrobial resistance. Science. 2021;371(6528):474–574.
94. Li B, Yang Y, Ma L, Ju F, Guo F, Tiedje JM, et al. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. ISME J. 2015;9(11):2490–502.
95. Chui CSL, Cowling BJ, Lim WW, Hui CKM, Chan EW, Wong ICK, et al. Patterns of inpatient antibiotic use among public hospitals in Hong Kong from 2000 to 2015. Drug Saf. 2020;43(6):595–606.
96. Ellabaan MMH, Munck C, Porse A, Imamovic L, Sommer MOA. Forecasting the dissemination of antibiotic resistance genes across bacterial genomes. Nat Commun. 2021;12(1):2435.
97. De Smet J, Hendrix H, Blasdel BG, Danis-Wlodarczyk K, Lavigne R. Pseudomonas predators: understanding and exploiting phage-host interactions. Nat Rev Microbiol. 2017;15(9):517–30.
98. Silveira CB, Luque A, Rohwer F. The landscape of lysogeny across microbial community density, diversity and energetics. Environ Microbiol. 2021;23(8):4098–111.
99. Sakowski EG, Arora-Williams K, Tian F, Zayed AA, Zablocki O, Sullivan MB, et al. Interaction dynamics and virus-host range for estuarine actinophages captured by epicPCR. Nat Microbiol. 2021;6(5):630–42.
100. Khwaja HA. Atmospheric concentrations of carboxylic acids and related compounds at a semiurban site. Atmos Environ. 1995;29(1):127–39.
101. Paul JH. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? ISME J. 2008;2(6):579–89.
102. Chen T, Deng C, Wu Z, Liu T, Zhang Y, Xu X, et al. Metagenomic analysis unveils the underexplored roles of prokaryotic viruses in a full-scale landfill leachate treatment plant. Water Res. 2023;245:120611.
103. Cheong KH, Wen T, Benler S, Koh JM, Koonin EV. Alternating lysis and lysogeny is a winning strategy in bacteriophages due to Parrondo's paradox. Proc Natl Acad Sci USA. 2022;119(13):e2115145119.
104. Barrangou R. The roles of CRISPR-Cas systems in adaptive immunity and beyond. Curr Opin Immunol. 2015;32:36–41.
105. Hynes AP, Villion M, Moineau S. Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. Nat Commun. 2014;5:4399.
106. Landsberger M, Gandon S, Meaden S, Rollie C, Chevallereau A, Chabas H, et al. Anti-CRISPR phages cooperate to overcome CRISPR-Cas immunity. Cell. 2018;174(4):908-16 e12.

107. Meaden S, Biswas A, Arkhipova K, Morales SE, Dutilh BE, Westra ER, et al. High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems. Curr Biol. 2022;32(1):220-27 e5.
108. Hynes AP, Rousseau GM, Lemay ML, Horvath P, Romero DA, Fremaux C, et al. An anti-CRISPR from a virulent streptococcal phage inhibits *Streptococcus pyogenes* Cas9. Nat Microbiol. 2017;2(10):1374–80.
109. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance 722 against viruses in prokaryotes. Science. 2007;315(5819):1709–12.
110. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J Bacteriol. 2008;190(4):1390–400.
111. Meaden S, Capria L, Alseth E, Gandon S, Biswas A, Lenzi L, et al. Phage gene expression and host responses lead to infection-dependent costs of CRISPR immunity. ISME J. 2021;15(2):534–44.
112. Luo XQ, Wang P, Li JL, Ahmad M, Duan L, Yin LZ, et al. Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. Microbiome. 2022;10(1):190.
113. Zgurskaya HI, Krishnamoorthy G, Ntreh A, Lu S. Mechanism and function of the outer membrane channel *TolC* in multidrug resistance and physiology of *Enterobacteria*. Front Microbiol. 2011;2:189.
114. Rowland JL, Niederweis M. A multicopper oxidase is required for copper resistance in *Mycobacterium tuberculosis*. J Bacteriol. 2013;195(16):3724–33.
115. Hartmann EM, Hickey R, Hsu T, Betancourt Roman CM, Chen J, Schwager R, et al. Antimicrobial chemicals are associated with elevated antibiotic resistance genes in the indoor dust microbiome. Environ Sci Technol. 2016;50(18):9807–15.
116. Hashemi SE, Fazlzadeh M, Ahmadi E, Parand M, Ramavandi B, Taghizadeh F, et al. Occurrence, potential sources, in vitro bioaccessibility and health risk assessment of heavy metal in indoor dust from different microenvironment of Bushehr. Iran Environ Geochem Health. 2020;42(11):3641–58.
117. An L, Liu X, Wang J, Xu J, Chen X, Liu X, et al. Global diversity and ecological functions of viruses inhabiting oil reservoirs. Nat Commun. 2024;15(1):6789.
118. Khan Mirzaei M, Xue J, Costa R, Ru J, Schulz S, Taranu ZE, et al. Challenges of studying the human virome-relevant emerging technologies. Trends Microbiol. 2021;29(2):171–81.

Publisher's Note