# Title: Host specificity and geographic dispersion shape virome diversity in *Rhinolophus* bats

**Authors:** Daxi Wang#1,2, Linmiao Li#3, Zirui Ren#1,2, Yepin Yu3, Zhipeng Zhang1,2, Jiabin Zhou3, Hailong Zhao1,2, Zhiwen Zhao3, Peibo Shi1,2,4, Xinrui Mi1,2,4, Xin Jin5, Ziqing Deng1,2, Junhua Li\*2,5, Jinping Chen\*3

**Affiliation:**

1BGI Research, Beijing 102601, China

2Shenzhen Key Laboratory of Unknown Pathogen Identification, BGI Research, Shenzhen 518083, China

3Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Institute of Zoology, Guangdong Academy of Sciences, Guangzhou, Guangdong, 510260, China

4College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

5BGI Research, Shenzhen 518083, China

#These authors contributed equally.

\*Correspondence: chenjp@giz.gd.cn; lijunhua@genomics.cn

# ABSTRACT

*Rhinolophus* bats have been identified as natural reservoirs for viruses with global health implications, including severe acute respiratory syndrome–related coronaviruses (SARSr-CoV) and swine acute diarrhea syndrome-related coronavirus (SADSr-CoV). In this study, we characterized the individual viromes of 603 bats to systematically investigate the diversity, abundance, and geographic distribution of viral communities within *R. affinis,* *R. sinicus*, and 11 related bat species. The massive metatranscriptomic data revealed substantial viral genome resources of 133 vertebrate-infecting viral clusters, which contain occasional cross-species transmission across mammalian orders and specially across bat families. Notably, those viruses included nine clusters closely related to human and/or livestock pathogens, such as SARS-CoVs and SADS-CoVs. The investigation also highlighted distinct features of viral diversity between and within bat colonies, which appear to be influenced by the distinct host population genetics of *R. affinis* and *R. sinicus* species. The comparison of SARSr-CoVs further showed varied impact of host specificity along genome-wide diversification and modular viral evolution among *Rhinolophus* species. Overall, the findings point to a complex interaction between host genetic diversity, and the way viruses spread and structure within natural populations, calling for continued surveillance efforts to understand factors driving viral transmission and emergence in human populations. These results present the underestimated spillover risk of bat viruses, highlighting the importance of enhancing preparedness and surveillance for emerging zoonotic viruses.

# KEYWORDS

Virome; *Rhinolophus* bats; diversity; evolution; cross species transmission

# INTRODUCTION

Bats are the natural reservoirs for many zoonotic viruses. Among bats, the Rhinolophidae were identified carrying viruses with global health implications, severe acute respiratory syndrome–related coronaviruses (SARSr-CoVs) (Zhou et al., 2020b) and the swine acute diarrhea syndrome-related coronavirus (SADSr-CoV) (Zhou et al., 2018). Examples include the close relative of SARS-CoV-1 in *Rhinolophus sinicus* (SL-CoV-WIV1) (Ge et al., 2013), and the close relatives (RaTG13 and many others) of SARS-CoV-2 from *R. affinis*(Zhou et al., 2020b) and *R. malayanus* (Zhou et al., 2020a), raising continuous concerns of spillover risk. Understanding the diversity and evolution of the viral communities in *Rhinolophus* bats becomes an essential step if our aims are to enhance preparedness for emerging zoonotic viruses. Recent investigations have provided deep evolutionary insights into the major pathogens associated with *Rhinolophus* bats, showing significant viral diversity and signs of frequent host-switching related to SARS-CoV-1 and SARS-CoV-2 (Latinne et al., 2020). Recently, a comprehensive survey of bat sarbecoviruses revealed a major recombination of SARS-CoV-related and SARS-CoV-2-related lineages (Wu et al., 2023). Most pathogen surveillance in bats has traditionally relied on targeted pathogen screening approach. Despite some recent metatranscriptomic studies of the Rhinolophus bats (Wang et al., 2023), the lack of sufficient sample size limits the knowledge of total viral diversity within and among Rhinolophus species, leaving underestimated spillover risks.

During the development of viral diversity, host specificity has been recognized as a primary factor driving the isolation and adaptation of viral populations (Ahlquist et al., 2003). In contrast, occasional host-switching contributes to the development of genetic plasticity and spillover risks (Munderloh et al., 2015). The viral sharing potential tends to be associated with evolutionary relatedness and ecological overlap of hosts (Albery et al., 2020). However, disentangling the evolutionary connectivity among viral communities of closely related host taxa remains a challenge due to the lack of ecological consideration and the biological difference among viral families. Geographic distribution of viral populations is another key factor in shaping the genetic diversity and evolutionary plasticity of viral populations (Garcia-Arenal et al., 2001). Geographic co-occurrence provides opportunities for genomic exchange among genetically distinct viral populations (Wang et al., 2023). For example, the building blocks of SARS-CoV genome, including the highly variable S gene, ORF3 and ORF8, were found in different SARSr-CoV viruses from a single bat cave (Hu et al., 2017). Therefore, estimating local population structuring and tracing recombination signals between distinct viral lineages, are essential for zoonotic virus surveillance, especially for viral taxa with major spillover risk.

With the advances in sequencing technologies, metatranscriptomic sequencing has become a valuable tool in virome research, enabling comprehensive characterization of viral diversity worldwide. Particularly, individual level metatranscriptome provides co-infection signals and high-quality viral genomes (Wang et al., 2023). The high resolution offered by deep metatranscriptomic sequencing facilitates systematic comparisons of evolutionary and ecological patterns across diverse viral groups.

In this study, we systematically investigated the richness, abundance, and diversity of viral communities within two major horseshoe bat species, *R. affinis* and *R. sinicus*, and other bat species with varying degrees of relatedness. The comparison of virome diversity and viral populations within co-habitat bat species across times not only provides baseline data for estimating viral sharing and host-switching across bat species, but also revealed geographic transmission of viruses related to the emergence of human pathogens. These results present the underestimated spillover risk of bat viruses, highlighting the importance of enhancing preparedness and surveillance efforts aimed at emerging zoonotic viruses.

### **MATERIALS AND METHODS**

### **2.1 Sampling and sequencing**

The bat samples were collected across 13 locations of Guangdong and Hainan Province in southern China from 2008 to 2021, with the approval by the Committee on the Ethics of Animal Experiments of the Institute of Zoology of the Guangdong Academy of Sciences. The sampling site include cave, sewer, mine, bomb shelter and abandoned residential buildings. Anal and oral swabs were sampled from bat individuals and suspended in tubes with viral transport medium (VTM), composed of Hanks’ balanced salt solution at pH 7.4 containing bovine serum albumin (BSA) (1%), amphotericin B (15 mg/mL), and 1×penicillin-streptomycin solution . All the samples were stored at −80 °C until further use. For each sample, 140 mL of the supernatant was collected after vortexing for 1 min and centrifuging at 3,000 rpm and 4°C for 1 min. Nucleic acid was extracted with a QIAamp viral RNA minikit (QIAGEN, Hilden, Germany) according to the manufacturer’s instructions. The RNA libraries were prepared using the MGIEasy RNA Library Prep Kit V3.0 following the manufacturer’s recommendations. The DNA nanoball (DNB) based libraries into which circular cDNA was rolling-circle replicated were sequenced on the DNBSEQ series platform (MGI, Shenzhen, China) to generate 150-bp paired-end reads. To reduce biosafety risks, sampling personnel need to take good safety precautions; sample processing, nucleic acid extraction, sequencing, etc. are all carried out in BSL-2, and operators strictly follow relevant regulations for experimental operations.

### **2.2 Data filtering**

Raw paired-end reads were firstly filtered to remove most ribosomal RNA (rRNA) reads using URMAP (v1.0.1480) (Edgar, 2020). After excluding adapters and low-quality reads with fastp(v.0.23.1) (Chen et al., 2018), the duplicates and low complexity reads were filtered using PRINSEQ++ (v.1.2.4), and the remaining rRNA reads were further removed using SortMeRNA (v4.3.2) (Kopylova et al., 2012) based on the SILVA database to obtain clean paired-end reads (Kopylova et al., 2012).

### **2.3 Identification of bat species**

A mitochondrial database was constructed by integrating sequences of mitochondrial marker genes (cytB, cox1 and nad1) from GenBank and BOLD database (Ratnasingham and Hebert, 2007). The read data for each sample were mapped to the mitochondrial database using bowtie2 (v.2.5.1) (Langmead and Salzberg, 2012) to define the abundance of each reference marker sequence. The candidate host species for each mitochondrial marker were annotated using megablast according to the first matched reference sequence with cut-off of 98% nucleotide identity and covered length longer than 400 bp, following the abundance order of reference marker sequences.

### **2.4 Virus identification and annotation**

The clean reads of each sample were *de novo* assembled by MEGAHIT (v.1.2.9) with default settings under the pair-end mode (Li et al., 2015). Assembled contigs with length above 1000bp were retained for virus identification. For each contig, potential open reading frames for protein translation were predicted using getorf from the EMBOSS software kit (v.6.5.7.0) (Rice et al., 2000). To identify viral contigs, the translated protein sequences were compared against curated HMM profiles of replication-associated proteins (RAPs) for each viral family using hmmsearch of the HMMER software (v.3.3.2) (Finn et al., 2011). For RNA viruses, the HMM profiles of the RdRp core motif were derived from the multiple sequence alignments of the RdRp databases from RdRp-scan (Charon et al., 2022). For DNA viruses, the HMM profiles of RAPs were curated from corresponding Pfam profiles (*Adenoviridae*: PF03175; *Circoviridae*: PF02407; *Herpesviridae*: PF00136; *Papillomaviridae*: PF00519; *Parvoviridae*: PF01057; *Polyomaviridae*: PF06431). All the HMM profiles were iteratively compared and and trimmed using SeqKit (v.2.4.0) (Shen et al., 2016). To remove non-viral sequences, the contigs encoding RAPs were compared against the non-redundant (NR) protein database (available as of May 2, 2022) using BLASTX (e-value < 10-5) of DIAMOND (v.2.1.8.162) (Buchfink et al., 2021) and NT database using megablast of BLASTN(Ye et al., 2006). Contigs that were not best matched to Eukaryotic or Prokaryotic entries were retained and further compared to the reference genomes under the Chiroptera order to remove host sequences using BLAT(Kent, 2002). The retained contigs were defined as putative viral sequences. The family of each viral sequence was assigned according to the taxonomy of the nearest known viral sequence from NR database. To annotate vertebrate-infecting viruses, all viral sequences were compared against proteins of the Virus-Host Database (<http://www.genome.jp/virushostdb/>, release 217) using BLASTX (e-value < 10-5) of DIAMOND (v2.1.8.162) (Buchfink et al., 2021). Only the viral sequences that were most similar to a vertebrate-infecting virus and/or belong to vertebrate-specific families were selected for further analyses. The completeness of each viral sequence was assessed using CheckV (v.1.0.1) (Nayfach et al., 2021). The viral sequences were then clustered using MMseqs2 (v.14.7e284) at 80% (vANI80) nucleotide identity (vANI80) with --cov-mode 1 parameter (Steinegger and Soding, 2017). Within each clustering level, viral sequence with the longest length was selected as representative sequence of each vANI80 cluster (Li and Godzik, 2006). For coronaviruses, viral contigs from the same sample were scaffolded under the guidance of closely related reference genomes when applicable. Gaps in the draft genomes were filled using the TaKaRa Taq™ Version 2.0 plus dye (TaKaRa, China) by reverse transcription PCR (RT-PCR). Primers were designed based on the assembled draft viral sequences (**Table S1**). PCR products of the expected size were gel purified and subjected to sanger sequencing.

### **2.5 Phylogeny of bat virome**

Within each viral family, the replication-associated proteins of each vANI80 cluster were aligned to the family's representative marker proteins using MAFFT (v.7.520) (Katoh et al., 2002) and then refined with TrimAI (v. 1.4) (Capella-Gutierrez et al., 2009). Sequences in the alignment shorter than 50 non-gap amino acids were excluded. Maximum likelihood phylogenetic trees were then generated using IQ-TREE multicore (v.2.2.0.3) (Nguyen et al., 2015), employing 1000 bootstrap replicates with default configuration. Phylogenetic trees of RNA viruses closely related to human or livestock pathogens were estimated using a maximum likelihood method based on the nucleotide sequences of the RdRp or S1-gene. The trees were midpoint-rooted for clarity. Phylogenic trees were visualized by ggtree (v.3.7.2) (Yu et al., 2017) in R.

### **2.6 Virome quantification**

Clean reads of each sample were mapped to all representative viral genomes of vANI80 clusters using Bowtie2 (v.2.5.1) (Langmead and Salzberg, 2012) with the --very-sensitive-local parameter. Viral genome coverage was calculated from the mapped BAM files using coverage program of SAMtools (v.1.17) (Danecek et al., 2021). To reduce false-positive, only the mapping records with more than 30% genome coverage, 1000 bp covered length and 0.5 RPM were considered viral positive.

### **2.7 Host switching pattern**

The virus-host associations were derived from the VIRION database (Carlson et al., 2022b). Viral sequences recorded on the VIRION database were compared against the representative vANI80 genomes using BLASTN (e-value < 10-5) (Ye et al., 2006). Protein sequences of the matched viral sequences were subjected to the detection of RAP HMM profiles (as described in the virus identification section), which were then added to the multiple sequence alignments of viral family phylogeny using the hmmalign mode of the HMMER software (v.3.3.2) (Finn et al., 2011). The multiple sequence alignment was subjected to the calculation of amino acid identity (AAI) between every pair of vANI95 representatives from the same viral family using customed Python script. Sequences with fewer than 100 amino acids within the alignment were removed from the comparison.

### **2.8 Virus recombination**

For SARSr-CoVs and SADSr-CoVs, the complete viral genomes identified in the present study were compared to the viral genomes from GenBank and aligned using MAFFT (version 7.520) (Katoh et al., 2002) with default settings. Recombination tests were carried out using RDP, GENECONV, MaxChi, BootScan, SiScan, and 3Seq within the RDP4 (Martin et al., 2015) framework with window size greater than 300bp and step size greater than 30bp. A recombination event was valid if the identified recombination segment was longer than 300bp and confirmed by at least four different recombination tests. The visualization of genomic similarity was conducted using Simplot (version 3.5.1) with a window size of 600 bp and a step size of 20 bp (Lole et al., 1999). Within vANI80 cluster with sufficient sample size and genomic variability, viral lineages and regions of recent recombination within the alignment of consensus genomes were defined using fastGEAR (Mostowy et al., 2017) with default settings and visualized using the plotRecombinations program of fastGEAR.

### **2.9 Variant detection**

For each sample, the clean reads were mapped to the representative genomes of vANI80 clusters using Bowtie2 (v.2.5.1) (Langmead and Salzberg, 2012). The consensus single nucleotide variants (cSNVs) among different hosts were identified from the resulting BAM files using Freebayes (<https://github.com/freebayes/freebayes>, v.1.3.6) on haploid mode (-p 1) with standard filters. For each vANI80 cluster, samples with over 20% missing data across cSNV sites were excluded and cSNV sites detected in less than 75% of samples were excluded using vcftools (v.0.1.16) (Danecek et al., 2011). The remaining cSNVs from each sample were then converted into consensus sequences and aligned with those from all corresponding samples. The intra-host single nucleotide variants (iSNVs) were detected using the variant caller LoFreq (v.2.1.5) (Wilm et al., 2012), applying a threshold of 5% minor allele frequency for samples with at least 30-fold viral read coverage. Samples with at least two iSNVs per kb were considered with the co-existence of genetically distinct viruses of the same vANI80 cluster. The identiﬁed variants were annotated using the SnpEff (v.5.1) with default settings (Cingolani et al., 2012). Data visualization was performed using the R package ggplot2 (v.3.4.2) (Ginestet, 2011).

### **2.10 Population genetics of bat hosts**

To investigate the geographic structuring of host populations, the single nucleotide variants (SNVs) were detected within the coding regions of single-copy orthologouses (SCOs) shared across *R. sinicus* and *R. affinis* using Freebayes (<https://github.com/freebayes/freebayes>, v.1.3.6) with standard filters. We used BUSCO (v.5.4.7) (Simao et al., 2015) to detect SCOs in *R. sinicus* on its reference genome (RefSeq accession GCF\_001888835.1). Due to the absence of a complete genome assembly for *R. affinis*, SCOs were detected among the assembled transcriptomic data in this species. The 1,345 SCOs shared between *R. sinicus* and *R. affinis*, with over 70% of the samples showing read coverage in more than 50% of the coding regions were used for the following analysis. Following the filtering of missing data and SNV singletons performed by vcftools (v.0.1.16), PCA analysis was conducted using plink (v1.90b6.21) (Purcell et al., 2007) on the filtered VCF files of two bat species. Nucleotide diversity (Pi) was calculated with window-size at 1000 nucleotides and fixation index were calculated between bat individuals from two different sites using vcftools (v.0.1.16).

### **2.11 Genome comparison among SARSr-CoVs**

The genomes of SARSr-CoVs were selected according the SARSr-CoV (L1) lineage of sarbecoviruses defined in a recent survey (Wu et al., 2023). Only the viral genomes with clear Rhinolophus host species from China were used for the following analysis. Redundant viral genomes were filtered using CD-HIT-EST (Li and Godzik, 2006) with 99.9% nucleotide identity and then aligned using MAFFT (v.7.520) (Katoh et al., 2002). The recombination events were detected using RDP4 (Martin et al., 2015) as described in the above section. Recombination events among viral genomes were converted into network format and visualised using cytoscape (Shannon et al., 2003), with edges connecting major parent, minor parents and recombinant genome of each recombination event. Within the network, the observed number of recombination linkage among viral genomes from different host species were compared to the null distribution of edges under a Bernoulli process within a random graph model. To compare the recombination frequency among host taxa or among geography, viral genomes were randomly subsampled (n=4) from each group (host taxa/province) for 100 times. In each time, recombination linkages among the subsampled viral genomes were counted for each node within the network. The FST was calculated among viral populations following a previous study (Hudson et al., 1992) () using a customed script. The multiple sequence alignment for each gene was extracted based on the GenBank annotation of the reference virus (NC\_004718.3) using SeqKit (v.2.4.0) (Shen et al., 2016). Pairwise nucleotide identity, nucleotide diversity, and minor allele frequency (MAF) for each nucleotide difference were calculated using customed script.

### **2.12 Pseudovirus preparation and infection**

The S genes of SADS-CoV and 200604 cloned into pcDNA3.1 by Wuhan GeneCreate Biological Engineering Co., Ltd. were used for pseudovirus construction. In brief, 5 μg pLenti-C-GFP Lentiviral Gene Expression Vector plasmid, 5μg psPAX2 plasmid and the 2.5μg S-protein-expressing plasmid or pCMV-VSV-G (control) were co-transfected into HEK293T cells using Lipofectamine 2000 (Thermo Fisher Scientific). Supernatants were collected 48h or 72h after transfection and clarified by centrifugation at 3,000×g , then passed through a 0.45-μm filter. The filtered supernatants were packaged and stored at −80 °C until use. Pseudoviruses prepared above together with polybrene (1: 1,000) were added to PK-15 cells. The unabsorbed viruses were removed and replaced with fresh medium including nuclear staining reagent Hoechst33342 (1: 1,000) at 48 h or 72 h after infection. The infection was monitored by microscopic (EVOS) to observation of GFP fluorescence intensity.

### **2.13 Statistic tests**

The Mann-Whitney U test was used for pairwise comparisons and the Kruskal-Wallis test for multiple group comparisons. The non-parametric tests were chosen due to the skewed nature of abundance data. For correlation analyses, Spearman's rank test was used to assess correlation. All tests were two-tailed, and a p-value of less than 0.05 was considered statistically significant.

# RESULTS

## 3.1 Sampling information of 603 bats in Southern China

In this study, a total of 603 anal swabs from distinct bat individuals were collected from southern China (**Figure 1A, Table S2**). The samples represented a diverse set of 13 bat species spanning seven genera across four families (**Figure 1B**). The samples were collected from 13 geographic sites affiliated with Guangzhou, Foshan, Huizhou, Shenzhen city in Guangdong Province and Qiongzhong city in Hainan Province (**Figure 1C**) from 2008 to 2021 (**Figure 1D**). Particularly, the majority of *Rhinolophus* bat individuals (299 out of 351) were collected from two caves (Site6\_HZ and Site1\_GZ) in Huizhou and Guangzhou from 2012 to 2021, respectively (**Figure 1E**). Within Site6\_HZ and Site1\_GZ, 154 anal swabs of *R. affinis* and 145 anal swabs of *R. sinicus* were collected, respectively.

## 3.2 Virome diversity and their host-switching patterns related to *Rhinolophus* bats

Metatranscriptomic sequencing of the 603 individuals generated 18.4 Tb raw data, with an average data size of 30.6 Gb per individual (**Table S2**). The filtering of low quality, ribosomal RNA (rRNA) and duplicated reads yielded an average of 10.3 Gb clean data per individual. The following de novo assembly and viral genome annotation of the 603 metatranscriptomes revealed 133 vertebrate-infecting viral clusters with 80% average nucleotide identity (hereafter vANI80). The 133 vANI80 clusters comprised 104 RNA vANI80 clusters and 29 DNA vANI80 clusters from 15 families (**Table S3**), including 74 vANI80 clusters identified in *R. affinis and R. sinicus* (**Figure 2A, Table S4**). Interestingly, despite the viruses shared between *Rhinolophus* species, none of the vANI80 clusters were shared between *Rhinolophus* bats and other bat genera (**Figure S1A**). The vANI80 representative sequences (n=133) included 93 medium (completeness>50%) and 65 complete/high quality (completeness>90%) genomes (**Figure 2B**).

Using the replication-associated proteins (RAPs) as evolutionary markers, we inferred phylogenies of major viral families. Only 21.8% of the vANI80 representatives were closely related to (>90% average amino-acid identity, RAP-AAI) the known viruses (**Table S3**), reflecting that most viral diversity in bats remains to be described. For *R. affinis and R. sinicus*, the comparison against public records from the VIRION database (Carlson, 2022) showed that only 21/74 of vANI80 clusters were identified previously in these two bat species (**Figure S1B**), reflecting the strong specificity of viral distribution. The novelty of the 133 vANI80 representatives also varied across viral families. Among major viral families with more than five vANI80 representatives, the median RAP-AAI against the known viruses was relatively high in *Coronaviridae* (99.51%) (**Figure 2C**). In contrast, the median RAP-AAI of the rest major viral families (n>=5) were all less than 80%. Several viral families (*Astroviridae*, *Picornaviridae*, *Parvoviridae*) comprised phylogenetic clades with at least five novel vANI80 clusters. Notably, a group of 46 vANI80 clusters of *Mamastrovirus* was newly identified in *Astroviridae* (**Figure 2A**). Particularly, all the astroviruses in *R. affinis and R. sinicus* were newly identified (**Figure S1B**). We also identified two complete viral genomes with high novelty (< 30% RAP-AAI) to known viruses (**Figure S2**). In *Hepeviridae*, we identified a virus (HepV-1) in *R. sinicus* with only 29.7% RAP-AAI to known hepevirus (**Figure S2A**). The infection of hepevirus causes hepatitis in mammals and birds. The deep phylogenetic divergence between HepV-1 and other hepeviruses suggests that bats may serve as important reservoirs of diverse hepeviruses. In *Parvoviridae*, we identified a virus (ParV-14/CH035) encoding four open reading frames (ORFs), which is unusual compared to the typical nonstructural and structural ORFs found in classic parvoviruses (**Figure S2B**). Taken together, the undescribed viral diversity here reflected an imbalanced effort across viral groups of previous surveillance.

Using the RAP-AAI between the vANI80 representatives and their nearest neighbor from other host taxa, our results showed that host-switching tend to occur at different phylogenetic depth across viral families (**Figure 2B**), which might relate to varied extent of host specificity. Most (42/74) of the *Rhinolophus* vANI80s showed close relatedness (>75% RAP-AAI) to the viruses of other mammalian genera. Among them, 21 *Rhinolophus* vANI80s showed the closest relatives from the bat genus *Hipposideros*, mainly distributed in *Astroviridae*, *Parvoviridae* and *Picornaviridae*, reflecting virus-host co-divergence during long-term evolution (**Table S4**)*.* Nonetheless, it seems that host-switching does not always occur between mostly related host-taxa. Among bats, 9 *Rhinolophus* vANI80s showed close relatedness to viruses from the bat genera (*Myotis*, *Miniopterus*, *Eptesicus*, *Scotophilus*) of a distinct suborder (Yangochiroptera). Furthermore, 11 *Rhinolophus* vANI80s were closely related to the viruses of humans (CoV-4), pigs (CoV-8), dogs (CV-8), non-human primates (ParV-2,4,5,12), and Beluga whale (PicoV-2,4,8,13). These observations underscore the frequent switching across host taxa with distant evolutionary distance.

## 3.3 Viruses related to human or livestock infection in *Rhinolophus* bats

Among the identified viruses, the ones with at least 90% AAI to the known viruses were detected in a considerable proportion of the viral positive *Rhinolophus* bats (64/100 for *R. affinis* and 58/79 for *R. sinicus*), reflecting their broad distribution among natural hosts (**Table S5**). Among them, nine vANI80 clusters were closely related to human or livestock pathogens (**Table S6**).

For SARSr-CoV (CoV-4), our study revealed 10.1% (35/351) positive rate of SARSr-CoVs among *Rhinolophus* individuals (**Table S5**). The signature of two key deletions(Letko et al., 2020) within the receptor binding domain (RBD) suggests non-ACE2 usage of the identified SARSr-CoVs (**Figure S3A**). The RdRp (nsp12) phylogeny showed that the identified SARSr-CoVs grouped with the public SARS-CoVs from Guangdong (**Figure 3A**). The genomes of SARSr-CoVs exhibited 93.5% average nucleotide identity between the members of Guangzhou and Huizhou. This contrasted with the 97.5% average nucleotide identity among members of the same bat roosts, suggesting spatial heterogeneity of viral diversity within the same province ( **Figure 3B**).

Another viral group with major spillover risk concern is the SADSr-CoV (CoV-8), which was detected in both *R. affinis* and *R. sinicus*. The RdRp phylogeny of SADSr-CoVs was structured corresponding to their host species. In Huizhou, one SADSr-CoV carried by *R. affinis* (strain SADSr-CoV/HZ/200604) showed the closest genetic relationship with SADS-CoV isolated from swine (**Figure 3B**). The S-gene of SADSr-CoV/HZ/200604 showed 96.6% nucleotide similarity to that of SADS-CoV from piglet, with completely identical amino acid sequence in the CTD region (**Figure S4A**). To verify the infectivity, we constructed the pseudovirus encoding the S-gene of SADSr-CoV/HZ/200604. The results showed that the pseudovirus could infect porcine kidney (PK-15) cells (**Figure 3C**), indicating the spillover potential for domestic pig infections. Further recombination analyses showed that SADSr-CoVs have undergone recombination with SADSr-CoV/HZ/200604 and SADSr-CoV/162140 strain (**Figure S4B, Table S7**), with the recombination hotspot surrounding the S-gene (**Figure S4C**). Remarkably, genome comparisons of SADSr-CoVs (CoV-8) revealed exchange of genomic segments between viruses from distinct host species, suggesting ongoing host sharing during host-specific viral diversification among bat taxa (**Figure S4D**). Nonetheless, there is no evidence of futher recombinations between SADS-CoVs from pigs and SADSr-CoVs from bats, considering that isolates from pigs across different farms were monophyletic.

Some other viruses from *Rhinolophus* batsshowed evidence of bat-like origin of human pathogens. In *Picornaviridae*, one vANI80 cluster (PicoV-12) from both *R. affinis* and *R. sinicus* showed close relatedness to (92.7% RAP-AAI) the Aichivirus A (AiV-A) (**Figure 3D**), a foodborne pathogen associated with gastrointestinal illnesses. The clade of AiV-A comprised viruses of a wide host range, including kobuviruses in humans (AiV-A1), canines (AiV-A2), sewage (AiV-A3), felines (AiV-A4), birds (AiV-A5), and rats (AiV-A6). Particularly, the viral phylogeny here showed the first-time detection of AiV-A in *Rhinolophus* bats, expanding the host range among bat taxa. However, it is still hard to infer the direct animal origin of AiV-A1 in humans.

In *Caliciviridae*, two vANI80 clusters (CalV-1 and CalV-3) from *R. affinis* and *R. sinicus* showed close relatedness to (75.6% RAP-AAI) the human Norovirus (**Figure 3D**), a foodborne viral agent causing gastroenteritis and diarrhea outbreaks worldwide. For the norovirus-related bat viruses, CalV-1 and CalV-3 and other *Rhinolophus* virusesformed distinct clades corresponding to their host species, suggesting strict specificity among bat species. The phylogeny of diverse norovirus-related *Rhinolophus* viruses surrounding the human and canine noroviruses suggested the possible bat origin of the Norovirus.

In *Parvoviridae,* four vANI80 clusters (ParV-2, ParV-12, ParV-5, ParV-4) were mostly related to Adeno−associated viruses (AAVs, genus: *Dependoparvovirus*) from humans with the RAP-AAI ranging from 79.8% to 77.7%. The phylogenetic relationship between Bat AAVs and other mammalian AAVs suggests that bats are the origin of mammalian AAVs (**Figure 3D**). The four clusters of Bat AAVs detected in the present study, and the previously reported Bat AAVs, appeared to form an ancestral branch for all known mammalian AAVs, including Adeno-associated virus A and Adeno-associated virus B, which consist of all known human AAVs. Due to the deficiency in replication, Adeno-associated viruses need to co-infect with a helper virus, including adenoviruses or other DNA viruses (Meier et al., 2020). The higher diversity across bat species suggests a bat origin of *Adeno-associated viruses.* Interestingly, our data showed co-occurrence of Adeno−associated viruses and adenoviruses. Most (7/8) AAV positive samples also carried *Adenovirus* in *Rhinolophus bats* (**Figure 3E**), supporting the biological association between the two viral groups.

## 3.4 Distinct viral community dynamics between *R. affinis* and *R. sinicus*

Systematic viral surveillance requires comprehensive understanding of viral community structures. We elected to compare the core viral communities (vANI80s occurred in at least two samples) of *R. affinis* (n=154) and *R. sinicus* (n=145) in Site1\_GZ and Site6\_HZ, considering the sufficient sample size in these two regions. The two *Rhinolophus* species showed some seemingly similar viral community features, including viral family composition (**Figure 4A**), viral positive rate (**Figure S5A**) and high occurrence of coronaviruses (**Figure 4A**). Nonetheless, *R. affinis* and *R. sinicus* were distinct in the composition of vANI80 clusters and the diversity of their viral communities. The two *Rhinolophus* species only shared 9/51 of the core vANI80s (**Figure 4A**), reflecting a maintained host specificity regardless of their ecological overlaps. In Site1\_GZ and Site6\_HZ, 36 core vANI80s were detected in *R. affinis*, while only 24 were detected in *R. sinicus*. The higher viral diversity in *R. affinis* was mainly contributed by the high prevalence (**Figure 4B**) and higher diversity (**Figure 4A**) of *Astroviridae* and *Picornaviridae* (**Table S8**). For example, 16 vANI80 clusters of astroviruses were specifically identified in *R. affinis* while only two astroviruses were specifically identified in *R. sinicus*, reflecting the distinct viral family competence between *R. affinis* and *R. sinicus* despite their close evolutionary relatedness. We compared the number of vANI80s for each bat individual. At individual level, *R. affinis* showed significantly higher number of vANI80s than *R. sinicus* per individual in Site6\_HZ (**Figure 4C**). Among bats, *R. affinis* has a higher percentage (35.06%, 54/154) of individuals carried more than one core vANI80s compared to *R. sinicus* bats (24.83%, 36/145). Particularly, eight *R. affinis* individuals showed co-infection of different CoV vANI80s, while only two *R. sinicus* individuals showed CoV co-infection. To demonstrate the complexity of viral communities, we further compared viral nucleotide diversity between *R. affinis* and *R. sinicus* using thesingle nucleotide variant data per vANI80 cluster. *Rhinolophus affinis* showed a higher nucleotide diversity of intra-specific viral populations compared to *R. sinicus* (**Figure 4D, Table S9**), suggesting a more complex origin of viral populations in *R. affinis*.Using the intra-host single nucleotide variants (iSNVs), we found out that *R. affinis* has a higher proportion of individuals with sufficient (>30X) viral coverage (10/49) carrying genetically distinct viruses of the same vANI80 cluster compared to *R. sinicus* (4/35), suggesting higher potential for viral genetic exchange in *R. affinis* (**Figure S5B**).

For geographic differences among viral communities, *R. affinis* showed higher percentage (61.11% vs. 41.67%) of core vANI80 clusters shared across Site1\_GZ and Site6\_HZ, suggesting more frequent viral transmission across geography (**Figure S5C**). In general, *R. affinis* showed higher proportion of individuals carrying viruses across geography than that of *R. sinicus*, supporting the distinct capacity of viral transmission between host taxa(**Figure 4E**)**.** The vANI80s of *R. sinicus* also showed significantly higher relative genetic differentiation between Site1\_GZ and Site6\_HZ compared to that of *R. affinis*, suggesting a more geographically structured viral population(**Figure 4F**).

For the dynamics of viral populations, the bat species *R. affinis* showed an elevated intra-specific nucleotide difference of vANI80 clusters within the same site across years. However, we did not observe significant genetic difference across years in *R. sinicus*. For each viral family, the family of *Coronaviridae* and *Astroviridae* showed the highest proportion of vANI80s detected across years (**Figure S5D**). In Site6\_HZ, we could not compare the viral communities across time in *R. sinicus,* given that only four *R. sinicus* bats were viral positive in 2020 and 2021. For *R. affinis* in Site6\_HZ, most (21/24) core vANI80 clusters in 2013 were detected in 2020 or 2021, suggesting long-term stability of viral communities (**Figure 4G**). In contrast, we observed a small overlap of core vANI80 clusters between 2020 and 2021 of *R. affinis* in Guangzhou (5/10) and Huizhou (7/24), respectively, suggesting temporal changes of viral community.

## 3.5 Distinct host population patterns between *R. affinis* and *R. sinicus*

The activities of host population are expected to be informative when forecasting pathogen distribution. To investigate the geographic structuring of host populations, we detected the single nucleotide variants (SNVs) within the coding regions of single-copy orthologouses (SCOs) shared across *R. sinicus* and *R. affinis*. We identified 1,345 SCOs shared between *R. sinicus* and *R. affinis*, with over 70% of the samples showing read coverage in more than 50% of the coding regions. Following the filtering of missing data and SNV singletons, 13,302 SNVs from 149 *R. affinis* samples and 12,127 SNVs from 135 *R. sinicus* samples were used for subsequent comparison. Using principal component analysis (PCA), *R. sinicus* comprised two major ancestral populations, while *R. affinis* did not show geographic differentiation (**Figure 5A**). We also estimated FST among bat populations. The genetic differentiation of host populations between and within geographic sites support the major differentiation in *R. sinicus* (**Figure 5B**)*,* which may contribute to the relatively lower genetic diversity and higher geographic differentiation of viral communities within *R. sinicus*.

To investigate the host dynamics, we compared the nucleotide differences among samples within and between years. In general, *R. affinis* population and *R. sinicus* populationshowed similar levels of nucleotide diversity (mean: 0.00086 for *R. affinis* and mean 0.00088 for *R. sinicus*).For the comparison across time, *R. affinis* showed a decreased nucleotide diversity from 2013 to 2021, while *R. sinicus* showed an increased nucleotide diversity, suggesting a different trend of population dynamics within the two host taxa. Interestingly, for *R. sinicus* in Huizhou, the genetic differentiation between Guangzhou and Huizhou decreased from 2020 to 2021 (**Figure 5C**), suggesting that the increased diversity of *R. sinicus* was mainly contributed by the increased gene flow across geography, which may intensify the viral complexity of local bat community.

## 3.6 Genome comparison of SARSr-CoVs reveals modular viral evolution and dispersal among *Rhinolophus* species

Host specificity and geographic dispersal are expected to define the extent of genetic exchange among viral populations. Given the extensive availability of public sequences, we chose SARS-related CoVs (SARSr-CoVs) in China defined (L1 lineage) in a recent survey(Wu et al., 2023) as a representative viral group to compare patterns genetic exchange among *Rhinolophus* bat species and among provinces (**Figure S6, Table S10**).

After the removal of nearly identical genomes (>99.9% ANI), the SARSr-CoVs showed major inter-host genetic differentiation (FST) among host species, especially when compared to the differentiation across provinces (**Figure 6A**). This suggests a major impact of host specificity in structuring the viral population. We then estimated the recombination relationship among viral genomes (**Table S11**). Interestingly, edges of recombination were enriched among viruses from different host species (P-value = 3e-39), when compared to the null distribution of edges under a Bernoulli process within a random graph model (**Figure 6B**). Our findings here suggest the underestimated genetic exchange among viral communities across host taxa. Across host taxa, SARSr-CoVs of *R. pusillus* had the highest number of recombination connections and the highest nucleotide diversity of S-gene (**Figure 6C**). Those viral populations may therefore comprise higher genetic plasticity for inter-host recombination and the development of spillover risk.

To unravel how recombination shapes diversification across the viral genome, we compared FST value among genes. Here, we assume that host specificity represents a more significant force shaping population structure compared to geographic difference within the same host. Interestingly, we did not observe similar genome-wide profile of gene FST between inter-host populations and inter-geography populations (**Figure S7**), suggesting distinct patterns of viral diversification between short-term versus long-term isolation. For example, regions encoding RdRp-related nonstructural proteins (nsp8-14) showed the lowest FST among host taxa, while they were consistently elevated across provinces within *R. sinicus*. The RdRp-related nsps showed a higher proportion of synonymous mutations compared to other genes (**Figure S7**). The low FST observed in RdRp-related nsps might be driven by their high saturation of synonymous polymorphisms, potentially hindering fixation during long-term between-host isolation.

Across the viral genome, gene FST among host taxa negatively correlated with recombination density (**Figure 6D**). Indeed, genome-wide distribution of recombination frequency showed a major elevation surrounding the S-gene region of the genome (**Figure 6E**). The S-gene also showed the highest nucleotide diversity and highest proportion of non-synonymous mutations among genes (**Figure S7**), reflecting adaptive evolution. Based on pairwise comparison of sequences, we observed valley between peaks in the identity distribution of RdRp (95% ANI) and S-gene (92% ANI), which typically represents the boundary between evolutionary clusters (**Figure 6F, Table S12**) (Kim et al., 2014). Using the cut-off of 95% ANI on RdRp (nsp8-14), the viral genome pairs showed higher divergence of S-gene when their viral host species were different (**Figure 6G**). In contrast, S-genes with high similarity (>92% ANI) were shared across *Rhinolophus* species (**Figure 6H**). The result suggests that the elevated recombination may introduce gene flow and, thus, reduce genetic differentiation and maintains an interconnected gene pool across viral populations (Nikolaidis et al., 2022). In conclusion, our findings reveal a complex interplay between recombination and selective pressure, leading to diverse patterns of genetic differentiation across viral genome.

# DISCUSSION

*Rhinolophus* bats carry zoonotic pathogens with massive public health and socio-economic impact, including SADS-CoV, SARS-CoV-1 and SARS-CoV-2 (Drexler et al., 2010; Zhou et al., 2018; Latinne et al., 2020). Despite that, surveillance of total virome in *Rhinolophus* bats remains limited. In the present study, we performed a large-scale investigation of the evolution and distribution of viral and host communities among *Rhinolophus* batsand other related taxa from the Chinese subtropics, with substantial viral and host genomic resources identiﬁed. We contend that the characterization of individual metatranscriptome is an effective strategy for monitoring wildlife-borne viruses with potential spillover risks and for observing the host population structures that facilitate the transmission and spread of these viruses.

Our extensive metatranscriptomic sequencing has uncovered a significant amount of previously undescribed viral diversity in bats. A striking finding is the dominant presence of potentially novel vANI80 clusters, suggesting that a substantial portion of bat virome diversity in China is yet to be characterized. The assembled viral genomic resources in our study have significantly enhanced surveillance sensitivity. Notably, we identified 46 vANI80 clusters within the Mamastrovirus genus of the *Astroviridae* family, known for its association with diarrheal illnesses. This high number of potential new viruses indicates a disparity in research focus across different viral taxa, evident from the variation in novel vANI80 clusters among the viral families. Additionally, our analysis of newly discovered bat viromes has not only confirmed various degrees of host restriction in most viral communities but has also facilitated the identification of historical host-switching events in viruses associated with *Rhinolophus* bats. This provides crucial insights into the evolutionary adaptability and spillover potential of viruses in previously unexplored viral groups.

Our investigation into viral diversity reveals strict host specificity among most vANI80s, along with diverse host-switching preferences across different viral families. We found that most viruses from *Rhinolophus* bats exhibit close phylogenetic relationships to viruses from Hipposideros, a genus of Old World leaf-nosed bats closely related to *Rhinolophus*. Rhinolophid and Hipposiderid bats diverged in Africa around 42 million years ago, according to fossil analyses (Foley et al., 2015). In *Coronaviridae*, hibecoviruses in *Hipposideros* bats are likely the evolutionary result of host-switching from sarbecoviruses in *Rhinolophus* bats (Latinne et al., 2020). Within the *Rhinolophus* genus, we detected a small proportion of vANI80s shared among *Rhinolophus* species, with signatures of ongoing viral diversification among these closely related host species. These findings suggest a deep co-divergence relationship between vertebrate-infecting virome and their host taxa (Longdon et al., 2014). Indeed, ecological contacts, such as sharing food sources or roosting sites, likely amplify the chances of host-sharing among bat species within the same geographic region (Leopardi et al., 2018). Our viral comparison within the same bat roosting sites highlights the maintained host specificity for most bat viruses despite the ecological proximity between host taxa.

Our data reveal viruses bridging large evolutionary distances between host taxa. Those examples not only include evolutionary proximity between viruses from distinct bat hosts, but also include close relatedness between bat viruses and the pathogens with public health implications. Recent studies have highlighted frequent host-crossing events, as evidenced by major discrepancies between virus-host phylogenies (Geoghegan et al., 2017). Viruses with broad host range are usually associated with higher spillover risk (He et al., 2022). The Swine Acute Diarrhea Syndrome Coronavirus, for instance, caused large-scale fatal outbreaks in piglets in China during 2016 and 2018 and is believed to have originated from bats (Zhou et al., 2018). We observed a high prevalence and frequent recombination of SADSr-CoVs, including a strain with an S1-region closely resembling that of SADS-CoVs. Previous research provides both evolutionary and experimental evidence supporting the theory that SADS-CoV originated from HKU2-related coronaviruses in bats, underscoring the pivotal role of bats in the ecology and evolution of SADS-CoV (Zhou et al., 2018). However, the exact transmission pathway from bats to pigs remains unclear (Scarpa et al., 2021). Our study demonstrates the ability of a SADSr-CoV pseudovirus, containing the S-gene from a virus isolated from *Rhinolophus* bats, to infect pig cells. Genetic changes surrounding cell-binding related region drive adaptation to phylogenetically distant hosts, thereby expanding their ability to colonize new host species (Longdon et al., 2014). The recombination among the zoonotic pathogen related viruses underlines the significance of genomic surveillance in wildlife host groups.

Understanding how viral populations are organized in nature is a key step towards effective viral surveillance. Despite varied extent of dispersal limitations, our results reveal a structured distribution of bat viruses across geography. Viral populations tend to exhibit limited genetic diversity within individual bat roosts, accompanied by low rates of transmission between communities, as suggested by significant genetic differentiation among bat roosts. Further, our results reveal distinct viral community features between two *Rhinolophus* species. Those distinct features include the number of shared vANI80s, the differences of genetic diversity, prevalence of viral families, and the geographic dis-similarity among viral communities. Notably, our research also demonstrates a marked difference in the host population structures between two closely related bat taxa. Previous studies indicate that host population activities can impact viral assemblies (Sallinen et al., 2020), suggesting that host genetic structures can provide valuable insights for viral surveillance. The relatively lower viral genetic diversity and higher geographic differentiation within *R. sinicus* appeared to be result of more structured distribution of host population. It should be noted that the introduction of *R. sinicus* populations from Guangzhou to Huizhou in 2020 and 2021 led to increased host diversity in Huizhou, which may intensify the viral complexity of local bat community. Environmental factors, such as deforestation, have been identified as potential drivers that bring together bat populations from diverse genetic backgrounds (Carlson et al., 2022a), thus potentially increasing viral diversity through their influence on bat movement patterns.

The varied genetic differentiation across viral genomes during host specification illustrates the complex interplay between host specificity and gene flow in shaping the landscape of viral genomes. The association between recombination density and viral genetic differentiation among host taxa suggests that frequent recombination, especially around the Spike genes, maintains an interconnected gene pool across viral populations (Nikolaidis et al., 2022). This might explain the rapid emergence and adaptation of viruses capable of exploiting diverse host taxa. In future, areas with high host diversity should be key locations for viral surveillance. However, further research is required to dissect specific recombination patterns across gene regions and their role in driving host jumps and virulence evolution. Understanding these dynamics is crucial for predicting and mitigating future viral threats, especially as climate changes and human activities increasingly mix wildlife species from previously isolated regions (Carlson et al., 2022a).

Our study has revealed distinct host and viral distribution patterns for two *Rhinolophus* species, both within and among bat roosts. It is important to note, however, that our observations were primarily confined to two regions. To enhance the applicability of our findings, future research should aim to cover a broader geographic scope. Moreover, our results underscore the urgent need for increased global surveillance efforts. By expanding the range of surveillance, we can gain a better understanding of the ecology and dynamics of bat populations, which will contribute to more effective conservation strategies and a deeper insight into potential zoonotic disease transmission pathways.

In summary, our study provides valuable genomic resources of the vertebrate-infecting vANI80s in *Rhinolophus* bats and related bat hosts, which uncovered patterns of host-switching, viral diversification among host taxa. We identified viruses spanning evolutionarily diverse host taxa, including viruses closely related to human and livestock pathogens. The ongoing recombination and modular evolution in SARSr-CoVs and SADSr-CoVs also reflect their constant spillover risk. The large-scale metatranscriptome of *Rhinolophus* bats point to a complex interaction between host genetic diversity, and the way viruses spread and structure within natural populations, calling for continued and detailed surveillance efforts to understand factors driving viral transmission and emergence. Future efforts should focus on expanding geographic and host species scope to further elucidate the complex dynamics of viral ecosystems. Our study sets a research framework for the efforts aiming to enhance preparedness and response strategies against public health threats from wildlife.

# SUPPLEMENTARY MATERIAL

# Figure S1. Distribution of vANI80s. (A) Distribution of vANI80 clusters detected in at least two samples across host species. (B) Comparison of vANI80s clusters identified in the present study and public database (VIRION).

# Figure S2. Evolution of HepV-1 of Hepeviridae and ParV-14 of Parvoviridae. (A) Phylogenetic position, and genome structure of HepV-1. (B) Phylogenetic position, and genome structure of ParV-14.

# Figure S3. Genetic comparison of SARSr-CoVs. (A) Alignment in RBD region of severe acute respiratory syndrome-related coronaviruses (SARSr-CoVs). (B) Pairwise nucleotide identity between the identified SARSr-CoVs within and between the sampling location.

# Figure S4. Recombination of SADSr-CoVs. (A) Alignment in CTD region; (B) Recombination of SADSr-CoVs. (C) Distribution of recombination segments across genome. (D) Cluster of SADSr-CoVs using fastGEAR. Color represents the lineage of the genomic segment.

# Figure S5. Distribution of core vANI80s in *Rhinolophus* bats. (A) Viral positive rate of *Rhinolophus* bats across Site1\_GZ and Site6\_HZ from 2013 to 2021, with the numbers indicating viral positive rate. (B) Samples with intra-vANI80 coinfection. (C) Viral sharing count between Guangzhou and Huizhou among *Rhinolophus* bats. (D) The proportion of vANI80s shared across years among viral families.

# Figure S6. Recombination across provinces. Network of inter- and intra-province recombination frequency. Node represents viral genome. Edge represents the linkage among parents and recombination within each recombination event. Colour of nodes represents province. Size of nodes represents node degree.

# Figure S7. Modular evolution of SARSr-CoVs. Population statistics of SARSr-CoVs across genetic components, including fixation index of genetic components across host taxa, fixation index of genetic components across provinces within *R. sinicus*, minorallele frequency (MAF) of nonsynonymous mutations, MAF of synonymous mutations, proportion of nonsynonymous mutations and nucleotide diversity.

# Table S1. Primers for viral amplification and verification

# Table S2. Sampling information

# Table S3. Viral sequences of vANI80 clusters

# Table S4. Nearest AAI of vANI80 clusters within and between host genera

# Table S5. Viral quantification of vANI80 clusters

# Table S6. Viral sequences related to human and livestock pathogens

# Table S7. Recombination signals of SADS-CoV

# Table S8. Diversity of core virome in Rhinolophus bats

# Table S9. Population diversity of vANI80 clusters in *Rhinolophus* bats

# Table S10. Genome sequences of SARSr-CoV for comparison

# Table S11. Recombination signals among SARSr-CoVs

# Table S12. Pairwise nucleotide identity between gene sequences of SARSr-CoVs

# FIGURES

# Figure 1. Geography of sampling sites and host taxonomy. (A) Distribution of sampling location and host species of bat samples. Size of each circle represents the number of samples with distinct host individuals. Color within the circle represents host species according to the legend. Color of circles represents their geographic sites. For clarity, the number of individuals per host genus is summarized for each geographic site. (B) Evolutionary relationship among the sampled bat species (host phylogenies of host species were derived from the public subsets of mammalian phylogeny, http:// vertlife.org/phylosubsets). (C) Distribution of bat samples across cities. Color represents city of sampling locations. (D) Distribution of bat samples by sampling year. Color represents city of sampling locations. (E) Sampling distribution of *Rhinolophus* bats in bat roosts with at least 25 samples.

# Figure 2. Evolutionary diversity of vertebrate-associated viruses. (A) Maximum likelihood phylogenetic trees of major viral families on replication-associated proteins (RAPs). The name of the viral family is shown above each tree. The solid black circles on each branch node represents bootstrap value above 50. The tip nodes on each tree represent vANI80 representatives identified in the present study, with host species annotated. (B) Completeness of representative genomes in each viral family, with the color indicating genome completeness. (C) Amino acid identity of replication-associated proteins (RAP-AAI) between the identified viruses and the known viruses. (D) Amino acid identity between vANI80s of *Rhinolophus* bats and their closest virus with different host genus across viral families.

# Figure 3. Intra-specific evolution of viruses related to human or livestock infection identified in *Rhinolophus* bats. (A) Phylogeny of severe acute respiratory syndrome-related coronavirus (SARSr-CoV, CoV-4) in RdRp and S1-gene. (B) Phylogeny of swine acute diarrhea syndrome-related coronavirus (SADSr-CoV, CoV-8) in RdRp and S1-gene. The phylogenetic tree was estimated using a maximum likelihood method based on the nucleotide of RdRp or S1-gene. The scale bar represents the number of nucleotide substitutions per site. Within each phylogeny, the names in black represented public viral genomes, whereas the viruses newly identified here are colored by geography. (C) Infection experiment of porcine kidney (PK-15) cells of pseudovirus encoding the S-gene of SADSr-CoV/HZ/200604. (D) Phylogeny of Adeno-associated virus (ParV-2, ParV-12, ParV-5, ParV-4), Aichivirus (PicoV-12) and Norovirus (CalV-3) in RdRp. The trees were midpoint-rooted for clarity only. (E) Heatmap of Adeno-associated viruses and adenovirus occurrence across *Rhinolophus* samples.

# Figure 4. Comparison across viral communities between *Rhinolophus* bats. (A) Geographic and host distribution of *Rhinolophus* virome across Site1\_GZ and Site6\_HZ. (B) Viral positive rate of *Rhinolophus* bats across Site1\_GZ and Site6\_HZ from 2013 to 2021. (C) Standard viral richness (viral records per individual) across viral families within *R. affinis* and *R. sinicus*. Nucleotide diversity of vANI80s within geographic site. (D) The number of samples carrying core vANI80 clusters shared across Site1\_GZ and Site6\_HZ. (E) Genetic difference of vANI80s between Site1\_GZ and Site6\_HZ. (F) Viral virome and genetic difference across time.

# Figure 5. Assessment of genetic variation among *Rhinolophus affinis* and *R. sinicus* samples across different years and bat roosts. (A) Principal component analysis (PCA) of host population genetics. The population structure within *R. affinis* (left) and *R. sinicus* (right) populations is visualized using PCA. The percentage of variation explained by each component is indicated in parentheses. Points are distinguished by collection site and year by shape and colors, respectively. (B) Boxplots of within-site genetic diversity. The boxplots represent the genetic diversity within *R. affinis* (left) and *R. sinicus* (right) samples, as measured across collection years 2013, 2020, and 2021. The y-axis shows genetic diversity on a log scale. Statistically significant differences in genetic diversity across years are denoted by p-values. (C) Fixation index (FST) among viral populations: *R. affinis* (left) and *R. sinicus* (right) are compared pairwise, with the FST values showing on the edge. The axes separate the data by collection site (GZ for Guangzhou, HZ for Huizhou) and year. The sample sizes for population are provided within the circles.

# Figure 6. Genome comparison of SARSr-CoVs. (A) Mean fixation index among viral populations. (B) Boxplot of recombination frequency across host taxa. (C) Network of within- and between-host recombination. Node represents viral genome. Edge represents the linkage among parents and recombinant within each recombination event. Colour of nodes represents host species. Size of nodes represents node degree. (D) Correlation between recombination density and gene fixation index among host taxa. (E) Distribution of recombination density across coronavirus genome. (F) Frequency of pairwise nucleotide identity in nsp8-14 and S-gene. (G) Nucleotide identity between the S-genes of SARSr-CoV genomes within and between host species after controlling for RdRp identity (ANI>95%). (H) S-genes sharing among SARSr-CoV genomes. Node represents viral genome. Edge represents the sharing of S-gene with at least 92% ANI. Colour of nodes represents host species.

# AUTHOR CONTRIBUTIONS

Conceptualization, DX Wang, LM Li, ZR Ren, JH Li, and JP Chen; Methodology, ZR Ren, ZP Zhang; Sample Collection and Processing, JP Chen, LM Li, JB Zhou; Data analysis, DX Wang, ZR Ren, LM Li, YP Yu, HL Zhao, PB Shi, ZP Zhang, XR Min, X Jin, ZQ Deng; Experimental verification, JB Zhou, ZW Zhao, LM Li, YP Yu; Writing – Original Draft, DX Wang; Writing – Review and Editing, LM Li, ZR Ren, JP Chen. Funding Acquisition, DX Wang, JH Li, and JP Chen; Resources (sampling), JP Chen; Resources (computational), DX Wang, JH Li; Supervision, DX Wang, JH Li, JP Chen.

# ACKNOWLEDGEMENTS

We gratefully acknowledge colleagues at BGI-Shenzhen and China National Genebank (CNGB) for RNA extraction, library construction and sequencing, and Libiao Zhang and Xiangyang He from the Institute of Zoology, Guangdong Academy of Sciences for their assistance in sample collection.This work was supported by grants from the National Key R&D Program of China (2021YFC2300900), GDAS Special Project of Science and Technology Development (2022GDASZH-2022010106), Planning Funds of Scientific and Technology of Guangdong Province (2023A1111110001) and Projects of the Guangdong Forestry Bureau [2024 Wildlife monitoring program].

# CONFLICT OF INTEREST STATEMENT

# The authors have no conflicts of interest.

# DATA AVAILABILITY STATEMENT

The host free sequencing data has been deposited in the CNGBdb with accession code CNP0005753. The assembled viral genome sequences have been deposited in the CNGBdb with the accession code CNP0005229. The assembled viral genome sequences have been deposited in the CNGBdb with the accession code N\_AAADWI010000000-N\_AAAECX010000000 (see **Table S3**). The original code and scripts has been deposited at Github (https://github.com/alexzrren/SouthernChina\_BatVirome). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

# REFERENCES

Ahlquist, P., Noueiry, A.O., Lee, W.M., Kushner, D.B., & Dye, B.T. (2003). Host factors in positive-strand RNA virus genome replication. *J Virol,* *77*(15), 8181-8186.

Albery, G.F., Eskew, E.A., Ross, N., & Olival, K.J. (2020). Predicting the global mammalian viral sharing network using phylogeography. *Nat Co*mmun, *11*(1), 2260.

Buchfink, B., Reuter, K., & Drost, H.G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods, 18*(4), 366-368.

Capella-Gutierrez, S., Silla-Martinez, J.M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972-1973.

Carlson, C.J., Albery, G.F., Merow, C., Trisos, C.H., Zipfel, C.M., Eskew, E.A., Olival, K.J., Ross, N., & Bansal, S. (2022a). Climate change increases cross-species viral transmission risk. *Nature, 607*(7919), 555-562.

Carlson, C.J., Gibb, R.J., Albery, G.F., Brierley, L., Connor, R.P., Dallas, T.A., Eskew, E.A., Fagre, A.C., Farrell, M.J., Frank, H.K., Muylaert, R.L., Poisot, T., Rasmussen, A.L., Ryan, S.J., & Seifert, S.N. (2022b). The Global Virome in One Network (VIRION): an Atlas of Vertebrate-Virus Associations. *mBio, 13*(2), e0298521.

Charon, J., Buchmann, J.P., Sadiq, S., & Holmes, E.C. (2022). RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol, 8*(2)*:veac082*.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., & Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin), 6*(2), 80-92.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., & DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience, 10*(2):giab008.

Drexler, J.F., Gloza-Rausch, F., Glende, J., Corman, V.M., Muth, D., Goettsche, M., Seebens, A., Niedrig, M., Pfefferle, S., Yordanov, S., Zhelyazkov, L., Hermanns, U., Vallo, P., Lukashev, A., Muller, M.A., Deng, H., Herrler, G., & Drosten, C. (2010). Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. *J Virol, 84*(21), 11336-11349.

Edgar, R. (2020). URMAP, an ultra-fast read mapper. *PeerJ, 8*, e9338.

Finn, R.D., Clements, J., & Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res, 39*, W29-37.

Foley, N.M., Thong, V.D., Soisook, P., Goodman, S.M., Armstrong, K.N., Jacobs, D.S., Puechmaille, S.J., & Teeling, E.C. (2015). How and Why Overcome the Impediments to Resolution: Lessons from rhinolophid and hipposiderid Bats. *Molecular Biology and Evolution, 32*(2), 313-333.

Garcia-Arenal, F., Fraile, A., & Malpica, J.M. (2001). Variability and genetic structure of plant virus populations. *Annu Rev Phytopathol, 39*, 157-186.

Ge, X.Y., Li, J.L., Yang, X.L., Chmura, A.A., Zhu, G., Epstein, J.H., Mazet, J.K., Hu, B., Zhang, W., Peng, C., Zhang, Y.J., Luo, C.M., Tan, B., Wang, N., Zhu, Y., Crameri, G., Zhang, S.Y., Wang, L.F., Daszak, P., & Shi, Z.L. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature, 503*(7477), 535-538.

Geoghegan, J.L., Duchene, S., & Holmes, E.C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog, 13*(2), e1006215.

Ginestet, C. (2011). ggplot2: Elegant Graphics for Data Analysis. *J Roy Stat Soc A ,174*, 245-245.

He, W.T., Hou, X., Zhao, J., Sun, J., He, H., Si, W., Wang, J., Jiang, Z., Yan, Z., Xing, G., Lu, M., Suchard, M.A., Ji, X., Gong, W., He, B., Li, J., Lemey, P., Guo, D., Tu, C., Holmes, E.C., Shi, M., & Su, S. (2022). Virome characterization of game animals in China reveals a spectrum of emerging pathogens. *Cell, 185* (7), 1117-1129 e1118.

Hu, B., Zeng, L.P., Yang, X.L., Ge, X.Y., Zhang, W., Li, B., Xie, J.Z., Shen, X.R., Zhang, Y.Z., Wang, N., Luo, D.S., Zheng, X.S., Wang, M.N., Daszak, P., Wang, L.F., Cui, J., & Shi, Z.L. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog, 13*(11), e1006698.

Hudson, R.R., Slatkin, M., & Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics, 132*(2*)*, 583-589.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res, 30*(14), 3059-3066.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664.

Kim, M., Oh, H.S., Park, S.C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol, 64*(Pt 2), 346-351.

Kopylova, E., Noe, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics, 28*(24), 3211-3217.

Langmead, B., & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359.

Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., Chmura, A.A., Field, H.E., Zambrana-Torrelio, C., Epstein, J.H., Li, B., Zhang, W., Wang, L.F., Shi, Z.L., & Daszak, P. (2020). Origin and cross-species transmission of bat coronaviruses in China. *Nat Commun, 11*(1), 4235.

Leopardi, S., Holmes, E.C., Gastaldelli, M., Tassoni, L., Priori, P., Scaravelli, D., Zamperin, G., & De Benedictis, P. (2018). Interplay between co-divergence and cross-species transmission in the evolutionary history of bat coronaviruses. *Infect Genet Evol, 58*, 279-289.

Letko, M., Marzi, A., & Munster, V. (2020). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol, 5*(4)*,* 562-569.

Li, D., Liu, C.M., Luo, R., Sadakane, K., & Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics, 31*(10), 1674-1676.

Li, W., Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*(13), 1658-1659.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., & Ray, S.C. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol, 73*(1), 152-160.

Longdon, B., Brockhurst, M.A., Russell, C.A., Welch, J.J., & Jiggins, F.M. (2014). The evolution and genetics of virus host shifts. *PLoS Pathog, 10*(11), e1004395.

Martin, D.P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol , 1*(1), vev003.

Meier, A.F., Fraefel, C., & Seyffert, M. (2020). The Interplay between Adeno-Associated Virus and its Helper Viruses. *Viruses, 12*(6), 662.

Mostowy, R., Croucher, N.J., Andam, C.P., Corander, J., Hanage, W.P., & Marttinen, P. (2017). Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol , 34*(5), 1167-1182.

Munderloh, U.G., Araujo, S.B.L., Braga, M.P., Brooks, D.R., Agosta, S.J., Hoberg, E.P., von Hartenthal, F.W., & Boeger, W.A. (2015). Understanding Host-Switching by Ecological Fitting. *Plos One, 10*(10), e0139225

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol, 39*(5), 578-585.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., & Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol, 32*(1), 268-274.

Nikolaidis, M., Markoulatos, P., Van de Peer, Y., Oliver, S.G., & Amoutzias, G.D. (2022). The Neighborhood of the Spike Gene Is a Hotspot for Modular Intertypic Homologous and Nonhomologous Recombination in Coronavirus Genomes. *Mol Biol Evol, 39*(1)*, msab292.*

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., & Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575.

Ratnasingham, S., & Hebert, P.D. (2007). bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol Ecol Notes, 7*(3), 355-364.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet, 16*(6), 276-277.

Sallinen, S., Norberg, A., Susi, H., & Laine, A.L. (2020). Intraspecific host variation plays a key role in virus community assembly. *Nat Commun, 11*(1), 5610.

Scarpa, F., Sanna, D., Azzena, I., Cossu, P., Giovanetti, M., Benvenuto, D., Coradduzza, E., Alexiev, I., Casu, M., Fiori, P.L., & Ciccozzi, M. (2021). Update on the Phylodynamics of SADS-CoV. *Life (Basel), 11*(8):820*.*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13*(11), 2498-2504.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One, 11*(10), e0163962.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., & Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212.

Steinegger, M., & Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol, 35*(11), 1026-1028.

Wang, J., Pan, Y.F., Yang, L.F., Yang, W.H., Lv, K., Luo, C.M., Wang, J., Kuang, G.P., Wu, W.C., Gou, Q.Y., Xin, G.Y., Li, B., Luo, H.L., Chen, S., Shu, Y.L., Guo, D., Gao, Z.H., Liang, G., Li, J., Chen, Y.Q., Holmes, E.C., Feng, Y., & Shi, M. (2023). Individual bat virome analysis reveals co-infection and spillover among bats and virus zoonotic potential. *Nat Commun, 14*(1), 4079.

Wilm, A., Aw, P.P., Bertrand, D., Yeo, G.H., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., & Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res, 40*(22), 11189-11201.

Wu, Z., Han, Y., Wang, Y., Liu, B., Zhao, L., Zhang, J., Su, H., Zhao, W., Liu, L., Bai, S., Dong, J., Sun, L., Zhu, Y., Zhou, S., Song, Y., Sui, H., Yang, J., Wang, J., Zhang, S., Qian, Z., & Jin, Q. (2023). A comprehensive survey of bat sarbecoviruses across China in relation to the origins of SARS-CoV and SARS-CoV-2. *Natl Sci Rev, 10*(6), nwac213.

Ye, J., McGinnis, S., & Madden, T.L. (2006). BLAST: improvements for better sequence analysis. *Nucleic Acids Res, 34*, W6-9.

Yu, G.C., Smith, D.K., Zhu, H.C., Guan, Y., & Lam, T.T.Y. (2017). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution, 8*, 28-36.

Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E.C., Hughes, A.C., Bi, Y., & Shi, W. (2020a). A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr Biol, 30*(11), 2196-2203 e2193.

Zhou, P., Fan, H., Lan, T., Yang, X.L., Shi, W.F., Zhang, W., Zhu, Y., Zhang, Y.W., Xie, Q.M., Mani, S., Zheng, X.S., Li, B., Li, J.M., Guo, H., Pei, G.Q., An, X.P., Chen, J.W., Zhou, L., Mai, K.J., Wu, Z.X., Li, D., Anderson, D.E., Zhang, L.B., Li, S.Y., Mi, Z.Q., He, T.T., Cong, F., Guo, P.J., Huang, R., Luo, Y., Liu, X.L., Chen, J., Huang, Y., Sun, Q., Zhang, X.L., Wang, Y.Y., Xing, S.Z., Chen, Y.S., Sun, Y., Li, J., Daszak, P., Wang, L.F., Shi, Z.L., Tong, Y.G., & Ma, J.Y. (2018). Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. *Nature, 556*(7700), 255-258.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., & Shi, Z.L. (2020b). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature, 579*(7798), 270-273.