INTRODUCTION

METHODS

**Dataset**

Astrovirus infections were identified from a dataset of viruses detected in samples from a long-term longitudinal sampling of fruit bats across Madagascar. Details on bat sampling and sequencing protocol can be found in previous work (CITE); here we will only give a brief overview.

Between 2018 – 2019, monthly bat captures were taken out at four species-specific locations: Ambakoana roost (-18.513 S, 48.167 E, *Pteropus rufus)*; Angavobe cave (-18.944 S, 47.949 E, *Eidolon dupreanum)*; Angavokely cave (-18.933 S, 47.758 E, *Eidolon dupreanum)*; Maromizaha cave (-18.9623 S, 48.4525 E, *Rousettus madagascariensis).* Each bat was identified by species, sex, and age (adult vs juvenile), and throat, fecal, and urine samples were taken.

In total, RNA from 285 fecal, 143 throat, and 196 urine swab samples was prepped into libraries and submitted for Illumina sequencing (see supplemental table). Raw reads from Illumina were filtered and assembled on the CZID bioinformatics platform (v3.10 NR/NT 2019-12-01). Samples were marked positive for astrovirus infection if at least two contigs with an average read depth >2 reads/nucleotide were assembled that showed significant nucleotide or protein BLAST alignment to astroviruses present in NCBI NR/NT database (v12-01-2019).

**Genome Annotation and BLAST**

We annotated all novel Astrovirus sequences by aligning them to annotated homologs published on NCBI and visually identifying open reading frames, using Geneious Prime (v08-18-2022).

**Statistical Analyses**

Done in r.

**Phylogenetic Analysis Overview**

To perform phylogenetic analysis, we combined our Astrovirus sequences with published NCBI sequences. We carried out three major phylogenetic analyses, building (a) a full-genome *Mamastrovirus* maximum (ML) likelihood phylogeny, (b) a *Mamastrovirus* ML phylogeny corresponding to a conserved 410 bp fragment of the RNA-dependent RNA polymerase (RdRp) gene encapsulated in the AstV ORF1b, (c) a time-resolved Bayesian phylogeny corresponding to a selection of full genome *Mamastrovirus* sequences in NCBI virus. Detailed methods for the construction of each phylogeny are available at <https://github.com/brooklabteam/Mada-Bat-AstV/>.

**Sequence Compilation**

Our full genome ML phylogeny consisted of one novel full length *Mamastrovirus* sequence, 41 unique NCBI *Mamastrovirus* sequences, and one full length *Avastrovirus* sequence as an outgroup, for a total of 43 sequences. We compiled the NCBI sequences through three queries: 1) all complete RefSeq Genomes under Virus: *Mamastrovirus* (taxid:249588) and Virus: *unclassified Mamastrovirus* (taxid:526119) greater than 6,000 bp (36), 2) *Mamastrovirus* nucleotide genomes under Virus: *Astroviridae* (taxid:39733) and Virus: *unclassified Astroviridae* (taxid:352926) with Host: *Chiroptera (bats)* (taxid:9397) over 6,000 bp (2), and 3) manual searching of nucleotide genomes >6,000 bp identified in the literature (3).

Our *Mamastrovirus* RdRp ML phylogeny consisted of the RdRp region of our novel full length *Mamastrovirus* sequence, 122 unique NCBI *Mamastrovirus* sequences, and one Avastrovirus RdRp fragment as an outgroup, for a total of 124 sequences. NCBI sequences were restricted to those from bat hosts sampled in the Southwest Indian Ocean (SWIO) region.

consisted of an overlapping subset of a 410 bp fragment in the center of the RdRp gene. They were compiled through one query: Virus: Astroviridae (taxid:39733), Host: Chiroptera (taxid:9397), and Geographic Region: Madagascar (64), Mozambique (31), and Reunion (27). Sequences were confirmed to be RdRp fragments via alignment and review of source literature.

**Alignment and Substitution Model**

Following dataset compilation, sequences were aligned using MAFFT (v7.450) (CITE) in Geneious Prime (v 2022-08-18) using default parameters. Alignments were visually examined and trimmed to conserved regions. We then used Modeltest-NG (v0.1.7) (cite) to determine the best fit nucleotide substitution for each alignment. All sequences, subsets, and alignments are available on GitHub: <https://github.com/brooklabteam/Mada-Bat-AstV/>.

**Tree-Building**

Both the full genome and RdRp ML trees were build using RAxML-NG (v1.1.0) (CITE), using the best nucleotide substation model from Modeltest-NG. For each tree, twenty ML inferences were made, followed by bootstrap replicate trees inferred using Felsenstein’s method (CITE). The MRE-based bootstrapping test was performed every 50 replicates, and bootstrapping was terminated when the diagnostic result was below the threshold value. Support values were compiled onto the best-scoring tree.

The Bayesian timetree was built using BEAST2 (v2.6.7) (cite). Following Lebarchenon 2018 which also looks at Astroviruses in Madagascan bats, we used the Coalescent Constant Population Model (CITE), with a strict lognormal clock rate with mean 0.0001 (CITE CARA, JENKINS), and all other priors as default. Sampling date for each sequence was inferred from NCBI ‘Collection Date’ or though reading source literature; if day and month were not available the sampling date was set to July 15th. Markov Chain Monte Carlo (MCMC) chains were run for X iterations, and terminated when we identified convergence using TRACER (v1.7), with 10% burn-in. We used TreeAnnotator (v2.6.3) to examine mean posterior densities at each node.

**Nucleotide Sequence Accession Numbers**

RESULTS

**Astrovirus Detection**

Astrovirus was detected in 4/285 (1.4%) fecal samples and 7/196 urine samples (3.57%). None of the 143 throat swabs assayed demonstrated evidence of AstV infection.

1/44 (2.27%) *Pteropus rufus*, 8/145 (5.52%) *Eidolon dupreanum*, and 2/96 *Rousettus madagascariensis* (2.08%) positives detected. Juvenile vs adult prevalence was 1/15 (6.66%) vs. 0/29 (0%) for *P. rufus*, 1/13 (7.69%) vs. 7/132 (5.30%) for *E. dupreanum*, and 1/13 (7.69%) vs. 1/83 (1.20%) for *R. madagascariensis* (Figure 1). Prevalence did not vary significant across seasons.

Diagram

Description automatically generated

Figure X: Map of sampling sites for *P. rufus*, *E. dupreanum*, and *R. madagascariensis* in the districts of Moramanga and Manjakandriana, Madagascar (*P. rufus:* Ambakoana roost; *E. dupreanum*: Angavobe/Angavokely caves; *R. madagascariensis*: Maromizaha cave). Pie charts correspond to astrovirus prevalence in juveniles vs. adults across all three species: 1/15 (6.66%) vs. 0/29 (0%) for *P. rufus*, 1/13 (7.69%) vs. 7/132 (5.30%) for *E. dupreanum*, and 1/13 (7.69%) vs. 1/83 (1.20%) for *R. madagascariensis*. Pie circle size corresponds to sample size on a log-10 scale.

**Genome Annotation and Similarity Analysis**

One near-full length Astrovirus contig was recovered from CZID, derived from one male, a juvenile *R. madagascariensis,* 6,593 bp in length. This sequence was identified in a male, juvenile *R. madagascariensis* sampled from Maromizaha cave. By aligning this sequence to annotated full genome *Mamastroviruses* from NCBI, we successfully annotated ORF1a, ORF1b, which contains the RdRp region, and ORF2, which contains the spike protein.

BLAST analysis of the near-full genome against all available partial and full astrovirus sequences in NCBI indicated that this *R. madagascariensis* sequence is highly divergent, demonstrating only 75.91% - 79.63%. All top blast associations were in other chiropteran species (see supplement).

This divergence is supported by a nucleotide similarity plot comparing this novel sequence to all full genome bat Astrovirus sequences from NCBI (Figure X). Using the novel near-full sequence as a query (F\_MIZ141\_1), we see low similarity across the entire genome, with all other bat sequences cluster together at a range of average similarity from 28.18% - 40.28%.

Graphical user interface, application

Description automatically generated

Figure X: Nucleotide similarity plot comparing all chiropteran full genome astrovirus sequences to the novel Madagascar bat astrovirus sequences. Query sequence is F\_MIZ141\_1, comparison sequences are MG693176 *(E. helvum)*, MZ218054 *(M. daubentonii 1),* MZ218053 *(M. daubentoniid 2),* MN832787 *(M. daubentoniid 3),* MT734809 *(Myotis yumanensis).*

**Phylogenetic Analysis**

Diagram

Description automatically generated

Figure X: Maximum likelihood phylogeny of full genome *Mamastrovirus* sequences (RAxML-NG, TVM+I+G4). Bootstrap values computed using Felsenstein’s method (CITE) are visualized on tree branches. Tip labels include NCBI taxon ID, strain, host species, location of collection, and year of collection. Tip points are colored by order, and shaped by novelty, with Madagascar sequences labeled with triangles. The novel astrovirus sequence presented in this paper is highlighted in yellow. Tree is rooted by a turkey *Avastrovirus* (NC\_002470) and a divergent Porcine *Mamastrovirus* (NC\_023636) (CITE). Branch lengths are scaled by nucleotide substitutions per site, corresponding to the scalebar.

Chart

Description automatically generated with low confidence

Fig X: Bayesian phylogeny of full genome *Mamastrovirus* sequences generated from 1.4 billion steps under a Coalescent Constant Population Model (TVM+I+G4). Node color represents mean posterior estimates averaging over 1.4 billion steps with 25% burn-in (see scale bar on left). Tip labels include NCBI taxon ID, strain, host species, sampling location, and year of sample, and are colored by strain.

Diagram, schematic

Description automatically generated

Diagram, schematic

Description automatically generated

Figure X: Maximum Likelihood phylogeny of a 410bp fragment of *Mamastrovirus* Orf1b (RdRp) from Southwest Indian Ocean (SWIO) bat hosts (RAxML-NG, TVM+I+G4). Bootstrap values computed using Felsenstein’s method (CITE) and those >30 are visualized on tree branches. Tip labels include NCBI taxon ID, host family, and host species. Tip points are colored by sub-order and shaped by country of origin. The novel astrovirus sequence presented in this paper is highlighted in yellow. Tree is rooted by a Turkey *Avastrovirus* (NCBI taxID: NC\_002470) and a divergent bat *Mamastrovirus* (NCBI taxID: MZ614426). Branch lengths are scaled by nucleotide substitutions per site, as indicated by the scale bar.

Discussion

1. Urine swab probably contamination