High throughput sequencing and discovery of novel bat viruses: insights for biosurveillance in an important virus reservoir

by

Adrian Caroline Paskey

Dissertation submitted to the Faculty of the
Emerging Infectious Diseases Graduate Program
Uniformed Services University of the Health Sciences

Distribution Statement

Distribution A: Public Release.

The views presented here are those of the author and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences, the Department of Defense or the U.S. Government.

UNIFORMED SERVICES UNIVERSITY OF THE HEALTH SCIENCES
SCHOOL OF MEDICINE GRADUATE PROGRAMS
Graduate Education Office (A 1045), 4301 Jones Bridge Road, Bethesda, MD 20814

APPROVAL OF THE DOCTORAL DISSERTATION IN THE EMERGING INFECTIOUS DISEASES GRADUATE PROGRAM

Title of Dissertation: "High throughput sequencing and discovery of novel bat viruses: insights for biosurveillance in an important virus reservoir"

Name of Candidate: Adrian Paskey
Doctor of Philosophy Degree
February 27, 2020

DISSERTATION AND ABSTRACT APPROVED:

DATE:

[redacted] 2/27/20

Dr. D. Scott Merrell
DEPARTMENT OF MICROBIOLOGY & IMMUNOLOGY
Committee Chairperson

[redacted] 3/11/2020

Dr. Kimberly A. Bishop-Lilly
DEPARTMENT OF MICROBIOLOGY & IMMUNOLOGY
Dissertation Advisor

[redacted] 2-27-2020

Dr. Christopher C. Broder
DEPARTMENT OF MICROBIOLOGY & IMMUNOLOGY
Committee Member

[redacted] 2/27/20

Dr. Clifton L. Dalgard
DEPARTMENT OF ANATOMY, PHYSIOLOGY & GENETICS
Committee Member

UNIFORMED SERVICES UNIVERSITY OF THE HEALTH SCIENCES

SCHOOL OF MEDICINE GRADUATE PROGRAMS

Graduate Education Office (A 1045), 4301 Jones Bridge Road, Bethesda, MD 20814

FINAL EXAMINATION/PRIVATE DEFENSE FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN THE EMERGING INFECTIOUS DISEASES GRADUATE PROGRAM

Name of Student: Adrian Paskey

Date of Examination: February 27, 2020
Time: 11:00 AM
Place: Room B4004

DECISION OF EXAMINATION COMMITTEE MEMBERS:

	PASS	FAIL
[redacted] Dr. D. Scott Merrell DEPARTMENT OF MICROBIOLOGY & IMMUNOLOGY Committee Chairperson	X	
[redacted] Dr. Kimberly A. Bishop-Lilly DEPARTMENT OF MICROBIOLOGY & IMMUNOLOGY Dissertation Advisor	X	
[redacted] Dr. Christopher C. Broder DEPARTMENT OF MICROBIOLOGY & IMMUNOLOGY Committee Member	✓	
[redacted] Dr. Clifton L. Dalgard DEPARTMENT OF ANATOMY, PHYSIOLOGY & GENETICS Committee Member	x	

ACKNOWLEDGMENTS

Thank you to my advisor, Dr. Kimberly Bishop-Lilly, who enthusiastically mentored me through this project. I have grown thanks to her excellent guidance and benefitted enormously from the environment she cultivates in the Genomics & Bioinformatics Department. Thank you to my committee for their advice and participation in this work. I am especially grateful that both my advisor and committee encouraged me to always follow my curiosity during the process. I am thankful that my classmates at USUHS and colleagues at NMRC-Frederick were always open to hearing about my latest research obsession. Their questions, answers and encouragement influenced me greatly. I am grateful to former mentors and professors who encouraged me to pursue my passion for research.

I am thankful for my family, who supported me during this journey in immeasurable ways. I am particularly thankful for the example they have set in helping others through work in all realms from service to science. Many thanks to my parents and friends who repeatedly answered the phone over the past four years, fully knowing that they would be speaking to an anxious friend or daughter. These individuals repeatedly convinced me that I was enough. Thank you to my husband, Alex Kowaleski, for embracing my choice to move to a new state and for making sacrifices at every turn to ensure our success as a team. I am grateful for the faith and running communities that welcomed me in Montgomery and Frederick Counties and made my everydayness so meaningful. Lastly, I am grateful for the opportunities I have enjoyed at NSA Bethesda and Fort Detrick, as well as for Joni Mitchell’s Blue album, which kept me sane as I commuted between them.

DEDICATION

To the health professionals who treat undiagnosable illnesses and kept me alive during my own traumatic illness from infectious diseases: May this be a small but meaningful contribution to the field of infectious diseases.

COPYRIGHT STATEMENT

The author hereby certifies that the use of any copyrighted material in the dissertation manuscript entitled: "High throughput sequencing and discovery of novel bat viruses: insights for biosurveillance in an important virus reservoir" is appropriately acknowledged and, beyond brief excerpts, is with the permission of the copyright owner.

[Signature] [redacted]

Adrian Caroline Paskey

April 5, 2020

DISCLAIMER

The views presented here are those of the author and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences, the Department of Defense or the U.S. Government.

ABSTRACT

High throughput sequencing and discovery of novel bat viruses: insights for biosurveillance in an important virus reservoir

Adrian Caroline Paskey, Doctor of Philosophy, 2020

Thesis directed by: Kimberly A. Bishop-Lilly, Ph.D., Adjunct Assistant Professor, Department of Microbiology and Immunology

Bats are rich reservoirs of viruses, including viruses associated with several high-consequence zoonoses. High-throughput sequencing and hybridization-based target enrichment sequencing were used to characterize the virome of a captive colony of fruit nectar bats, lesser dawn bats (*Eonycteris spelaea*) in Singapore through a longitudinal study, and a wild bat population of cyclops roundleaf bats (*Hipposideros cyclops*) in Uganda. Through the use of viral RNA extracted from bat swabs, we evaluated the utility of external and internal swab sites for biosurveillance and discovered novel viruses by shotgun and enrichment sequencing. Several viruses cataloged in this study are related to viruses that have previously crossed the species barrier from bats to humans, or other incidental intermediate hosts. To our knowledge, this is the first study that combined probe-based viral enrichment with high-throughput sequencing to create a viral profile from multiple swab sites on individual bats as a cohort. It was necessary to develop a new pipeline for the bioinformatic analysis of our samples, as well as a normalization technique to make comparisons among samples.

We hypothesized that some viruses may persist within the captive colony of bats long-term, as opposed to decreasing below the level of detection in the absence of migration and new, naive bats. This work demonstrated distinct temporal patterns of the lesser dawn bat virome and also led to the discovery of novel viruses in both lesser dawn bats and cyclops roundleaf bats. We found that noninvasive surveillance methods that target the body of bats not only detected viruses shed within the colony, but also represented viral populations dispersed throughout the entire colony. This new knowledge of persistent viral families should inform future directions for biosurveillance of viruses that have the potential to cross the species barrier from bats to humans or other amplifying hosts.

Perhaps most immediately relevant, the knowledge that a rubella-like virus circulates in equatorial African bats should be used to inform decisions with regard to the World Health Organization's plan to eliminate human rubella virus. Through this work, we evaluated and developed new tools for use in wet-lab and computational components of biosurveillance, and implemented them to generate a framework for future public health-related efforts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

ONE HEALTH

Numerous emerging infectious diseases have been linked to bats, including henipaviruses, filoviruses (i.e., Ebola and Marburg viruses), lyssaviruses, such as rabies, and coronaviruses, such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (20; 181; 212). While most of these viruses can cause severe morbidity in humans and result in robust immunity for survivors, bats are seemingly unaffected reservoir hosts (158). In fact, it is unknown whether these viruses persist long term in bats or transiently circulate in wild bat populations (147); this knowledge gap is increasingly being studied through the recently developed movement of 'One Health.'

One Health is the concept that the health of humans, animals and the environment are intrinsically linked. This concept of One Health was first described in 2000 (34), and has since become the basis of journals and supporting organizations, as well as an impetus for interdisciplinary collaboration. The One Health movement has bridged interdisciplinary communication by encouraging collaboration among veterinarians, public health officials, analytical scientists and field scientists (33). An important aspect of One Health-framed research is the consideration of the human impact on emerging infectious diseases. This framework incorporates molecular research of pathogens as well as human behaviors and ecological disruptions that may influence the emergence of zoonotic disease.

By applying the One Health lens to study viruses circulating in animal reservoirs, the scientific community can obtain a more complete understanding of the risk of cross-species transmission of viruses from bats to other hosts. One Health research has benefitted from the advent of metagenomic sequencing, which increased laboratory capabilities to detect genomic evidence of viruses (53). An increased awareness of the ecological role of viruses, beyond as being a cause of emerging diseases, has led to numerous metagenomic studies that found viruses to be a necessary part of the global ecosystem. The nascent idea that some viruses may play a nonpathogenic role in plant and animal health is reflective of a greater understanding of the virosphere as a part of the global ecosystem. This is supported by research on the association of functional redundancy and biodiversity with healthy ecosystems (53). The work presented in this thesis follows an ecosystems perspective by including analyses of both viruses of known zoonotic relevance and viruses that are currently regarded as "environmental." The impact of interdisciplinary collaboration among biologists through the One Health perspective has the potential to lead to more comprehensive and effective approaches for the mitigation of viral spillover.

BATS ARE A RICH, NATURAL RESERVOIR FOR MAMMALIAN VIRUSES

Bats and the ecosystem

There are more than 1,300 known bat species (order Chiroptera), classified as Megachiroptera (megabats) and Microchiroptera (microbats) (201). Megabats include all Pteropid species and are commonly called fruit bats (20). Microbats, such as the cyclops roundleaf bats studied in this work, can echolocate and are known to roost in internal shelters such as caves, tree cavities and man-made structures (96). Bats are crucial to the global ecosystem because they consume arthropods and play a role in pollination and seed dispersal (112).

Loss of biodiversity and ecosystem disruption are increasingly recognized as threats to global bat populations (143). As human behavior influences the ecosystem, population dynamics in bats continue to evolve. The geographic distribution of hosts and their intra- and interspecific interactions influence viral abundance, as well as the potential for transmission of viruses from one bat species to another or into a new mammalian host altogether (53). Bats are remarkably speciose and are known to share more viruses among species than any other mammal (112). While considerable effort has been invested to identify risks to human health by identifying and classifying bat-borne viruses, there is a need to marry this information with serological and human behavioral studies to properly anticipate risk to human or livestock populations in locations such as South-eastern Asia (143). A One Health approach requires the identification of human behaviors that influence bats, in addition to cataloging viruses that are carried by bat species (143). Further considerations that should be taken into account to understand risk of a virus crossing species barriers include receptor tropism and range of receptor distribution among possible spillover hosts, conservation and physical location of receptors in different host tissues, as well as the potential for cross-reactive antibodies that could protect from or enhance disease.

Rather than valuable ecological pollinators and seed dispersers, bats have been deemed nuisances and even villains by numerous cultures. Portrayal of bats in such a negative light can have dire consequences when reactive human behaviors (i.e., culling) influence the bat virome. Culling of bat populations increases human contact with wild bats and has been shown to ultimately increase viral prevalence within the targeted population (5). Effective interventions, such as vaccines, that can be autonomously transferred among bat colonies have been presented as viable alternatives to persecuting bat populations (8). Administration of vaccines that require minimal contact between bats and humans is a safer approach than interventions that ultimately kill or displace bat populations.

Potential intersections of virus transmission are illustrated in Figure 1. Scientists have continued to accumulate knowledge about bat-borne zoonotic threats by sampling bats in the wild. Bats do not show symptoms of infection when carrying most zoonotic viruses (one exception includes certain lyssaviruses), yet are capable of infecting humans or domestic livestock (Figure 1D) (143). Additionally, bats are known to shed virus at intermittent levels based on seasonal and birth pulse influences, further obscuring detection capabilities (145). Despite growing awareness of the ecological role of bats and despite One Health approaches, negative perceptions of Old World bats perpetuate in the wake of the detection of zoonotic-related viruses and human outbreaks. Such examples include Hendra virus, Nipah virus and Severe acute respiratory syndrome coronavirus, as well as agricultural outbreaks of viruses like Porcine epidemic diarrheal virus (10; 36; 43; 110; 202; 210). It has been hypothesized that viruses that coevolved with bats may use conserved receptors, increasing the likelihood of spillover from bats to other mammals (20).

Anthropogenic and innate factors influence the risk of spillover from bats

Bats are necessary components of ecological health as providers of pollination, seed dispersal, insect consumption and guano fertilizer production (95). Disruption of bat habitat and declining natural food resources result in an increased probability of human-bat interactions, increasing the risk of viral spillover from bats to humans (97). Bats adapt to urban human environments, a process termed synanthropic adaptation, by living in or near human dwellings. One example is the adaptation of flying foxes (*Pteropus alecto*) to favor urban flower gardens in response to declining wild food sources in East Australia (71). Not only does a dietary change such as this impact the physical health and immunity of the bats, but it has also been shown to increase the risk of viral emergence from flying fox populations via increasingly common equine or human interfaces (146). Individual bat behavior also plays a role in the likelihood of virus spillover from bats to other mammals. For example, territorial behavior or diet (i.e., ability to seek out adequate nutrition in an urban environment) may impact which bats shed or transmit viruses among a colony as compared to other individuals (53).

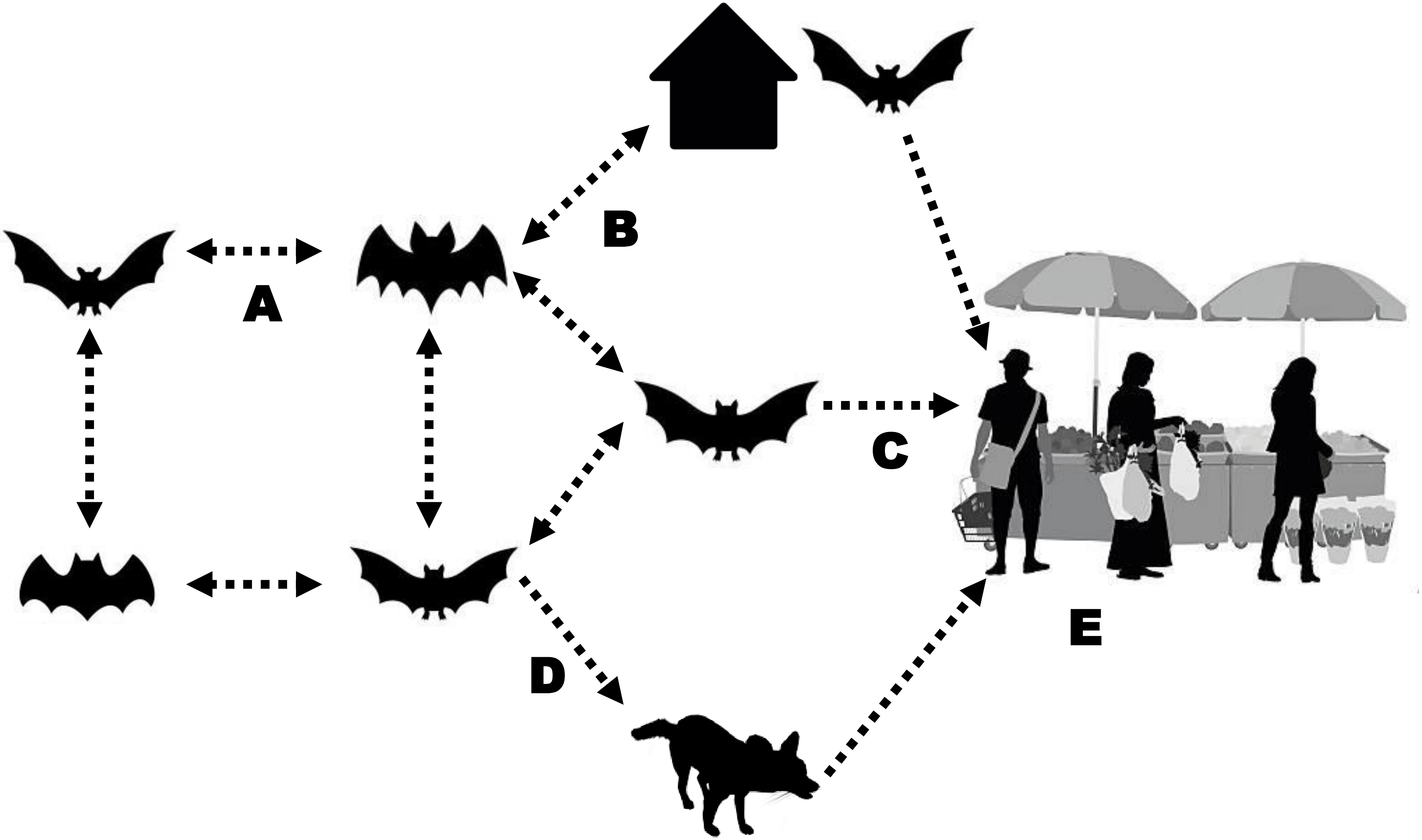
Unique immunological features of bats

Bats are the most widely distributed land animals and the only flying mammals (20; 188). It is believed that bat-associated viruses evolved in ancient bats using receptors that are conserved in later-evolving mammals (20). Unique immunological features and body regulation patterns, such as torpor (hibernation) and a heightened body temperature that occurs during flight, distinguish bats from other mammals. Interestingly, not all bat species are alike. Some bats do not exhibit the reduced body temperature and metabolic rate of torpor, migrate, or remain in warm climates year-round (73), which is the case for the lesser dawn bat and cyclops roundleaf bat species discussed in this thesis. In addition to a heightened body temperature, flight also causes the production of oxygen-free radicals, which damage DNA and result in inflammation. In response to this unique physiology, bats have evolved mechanisms to avoid excess inflammation, and as a result, can asymptomatically tolerate viral infections (26). The tightened control of immune responses with constitutive expression of interferon and interferon-stimulated genes is tempered by multiple mechanisms to avoid the over-induction of inflammatory genes (26; 165; 169; 211). While bats do not display symptoms when infected with zoonotic viruses, evidence of prior infection can be detected by serology (147).

Bats specific to this study

This study includes megabat lesser dawn bats (*Eonycteris spelaea*) that were caught in the wild and sampled in captivity and microbat cyclops roundleaf bats (*Hipposideros cyclops*) that were caught and sampled in the wild. Lesser dawn bats are a common nectar-feeding species in Southeast Asia that produce one pup annually and breed throughout the year. This

Figure 1. Possible intersections of viral transmission.
The schematic represents possible avenues for viral transmission among bats, intermediate amplifying hosts and humans. Intersection A represents co-roosting and interspecies transmission of viruses that have long been considered to evolve in bats. Examples of factors that may influence transmission include migration and birth pulses. Intersection B illustrates synanthropic adaptation of bats to human residencies, increasing the probability of bat-human interaction. Intersection C represents possible direct or indirect contact between wild bats and humans. This avenue could occur through the preparation or consumption of bat meat or indirectly through consumption of contaminated produce such as fruit. Intersection D illustrates transmission of viruses from bats to an intermediate amplifying host such as a palm civet, which was the intermediate host for SARS-CoV. The final intersection of transmission, E, is via human-to-human contact. In some primary cases of zoonotic spillover, viruses may not spread efficiently among human hosts unless the virus spills into a human population with behaviors that enhance viral transmission.


B
A
C
E
D

understudied species exhibits docile behavior and typically roosts in colonies of thousands of individuals, often with other bat species (74; 121). Constant gregarious behavior exhibited by lesser dawn bats is thought to increase the transmission of viruses between animals. This species does not hibernate, as the seasons in their hot tropical geographical territory are wet and dry. As nectar-feeding mammals, lesser dawn bats consume primarily nectar from flowers and orchard crops (160).

Lesser dawn bats are significant to One Health as a known host of zoonotic-related viruses. The number of known zoonotic-related viruses detected in lesser dawn bats has recently grown, perhaps as a result of increased surveillance of the species following predictive modeling studies that named the species a potential host of filoviruses (70). Lesser dawn bat exposure to filoviruses was also discovered in Singapore via serological methods (98). Furthermore, a metagenomic study published in 2017 expanded the diversity of known viruses shed by wild lesser dawn bats (124), and, as of January 2020, the Database of Bat-Associated Viruses listed only 93 entries for viruses detected in this species. For comparison, 146 viruses are listed for Leschenault's rousette (*Rousettus leschenaultii*) (25). The continuing study of lesser dawn bat behavior and their virome will help to determine the risk for spillover to humans and livestock and, if necessary, effective public health interventions to prevent it from occurring.

Cyclops roundleaf bats, are microbats found in equatorial Africa and have an insectivorous diet. They are named for their distinctive leaf-shaped nose and forehead opening (cyclops) (38). This species roosts in colonies of up to twelve, or singly, and breeds once a year in December. It is known to co-roost in hollow trees with other bat species and flying squirrels (38). Cyclops roundleaf bats are not included in the Database of Bat-Associated viruses (25), but have been evaluated as a potential host for filoviruses (70; 142). This bat species is difficult to study due to its roosting habits in deciduous forests and low roosting numbers, perhaps accounting for the dearth of publications that include cyclops roundleaf bats. This bat species is discussed further in Chapter Five and is significant to One Health as a host of viruses of pandemic potential.

OUR AWARENESS OF THE VIROSPHERE IS INCREASING

Known and previously unknown viruses

The cataloged, or known, virosphere has rapidly expanded due to the metagenomic revolution, clarifying the diversity of viral families and disrupting conventional viral taxonomy classifications (208). Modern metagenomic research has been biased toward mammalian, avian and invertebrate hosts (208). Other areas that have progressed in stride with metagenomic advances include the fields of plant virology and bacteriophage biology, which have improved human understanding of microbial landscapes (72). Plant viral metagenomics is a major focus of virology research and has been a part of the developing ecological network through which viruses are increasingly contextualized (104).

This work focused on RNA viruses and methods were optimized to capture RNA genomic material because most emerging zoonotic viruses are RNA viruses (196). The sequence for viral RNA-dependent RNA polymerase (RdRp) is typically used to define relatedness of viruses at the genomic level because it is the only gene shared by all known RNA viruses (93). Examining virus relatedness through the lens of genomics has caused a disruption in conventional virus classification because observable replication strategies, genomic architecture, such as segmentation, etc. do not translate to evolutionary relatedness as much as was previously thought (189; 208). For example, human Rubella virus was classified in the *Togaviridae* viral family following its discovery in 1814 (50; 117), but a proposal was accepted in 2018 by the International Committee for the Taxonomy of Viruses (ICTV) to classify Rubella virus within its own viral family, *Motonaviridae* (152). In fact, the evolutionary relationship of the conserved catalytic domain of RdRp of *Motonaviridae* viral sequences is evolutionarily closest to *Hepeviridae* and *Benyviridae* families rather than *Togaviride* (189). As the virosphere continues to be sequenced and better understood, reclassification of viral families is expected to continue (159).

To date, the discovery of viruses has occurred through viral isolation, by sampling from suspected viral hosts and subsequent sequencing of specifically amplified PCR products or via unbiased metagenomic sequencing. There has been substantial effort by the scientific community to identify areas of inherent bias in sampling and sequencing approaches (62; 118). Major areas of recent progress include an increasing diversity of applied methodological approaches, expansion of geographic regions of study, and breadth of organisms targeted for sampling. For example, a massive effort to better define phage communities across multiple oceans was recently published. Not only did this work demonstrate the power of metagenomics to discover previously unknown viruses, but it also revealed unexpected population differences across latitudes (67). Appreciation for geospatial elements impacting the virosphere will grow as more studies are conducted globally and data are made publicly available.

As sequencing technology has improved over the past decade, read lengths and throughput have increased. This resulted in improved data quality and expanded the number of specimen types that could be readily sequenced using metagenomic sampling. One growing problem now faced by the field of viral genomics is the inability to classify viral dark matter (92; 208). Despite the abundance of viral genomic data that has been generated, viruses are still underrepresented in publically available sequence databases. The remaining, unclassifiable sequences constitute “dark matter” and can only be putatively labeled as derived from viral domains based on negative evidence. Dark matter is defined as sequence data that do not align to a known organism (92). In order to perform classification of viral sequences, database-reliant methods such as alignments are typically performed using public databases. The most commonly-used approach to classify viral sequence data is the Blast suite of methods, which relies on homology to known sequences (3; 72; 92). Classification by this approach, particularly for amino acid sequences, is computationally burdensome, but has been made easier by use of a computational indexing system called DIAMOND, which was utilized in this thesis (18). While it is cumbersome to perform these analyses at the amino acid level, it is often necessary to do so because primary nucleotide sequence alignments alone are not enough to classify viral sequence. We must take into special consideration the significance of conserved amino acid sequences in regions that may not be conserved at the nucleotide level among viruses because they mutate faster than cellular organisms.

The emphasis of this work is on mammalian viruses that are phylogenetic neighbors to human pathogens, and therefore carry a concern in the realm of biosurveillance. Animals that are sampled through biosurveillance programs are typically targeted based on extrapolated risk for spillover, resulting in a bias toward previously known vectors and reservoir hosts such as bats and rodents. Furthermore, animal surveillance is often biased toward unhealthy, stranded or deceased animals (60). As high throughput sequencing (HTS) and computational power improve, the ability to identify and characterize novel viruses will expand and ought to encompass a more diverse set of animals. Efforts to sample more diverse insects, fish and aquatic mammals are already underway (60; 208). For example, an unmanned aerial vehicle was used to sample the previously inaccessible virome of the Eastern Australian humpback whale by capturing exhaled breath (blow), extracting RNA from collected water samples, and analyzing HTS results (60). The better we understand the virome of a healthy ecosystem, the more clearly we can discern imbalances in a disturbed ecosystem that may lead to viral emergence.

Publically available databases

The majority of existing viruses have not yet been identified or are not represented in publically available sequence databases. One databank sponsored by the United States government is the National Center for Biotechnology Information (NCBI), which provides access to genomic information through databases of published genomes, known genes and amino acid sequences. Authors who conduct metagenomic studies are encouraged to deposit raw sequence data in NCBI's Sequence Read Archive (SRA) (106). This resource is the most accessible source for sequence data, and typically authors are required to submit their sequence data for public availability through NCBI prior to publication in reputable journals. While NCBI provides an indispensable breadth of data diversity, there is little incentive to provide accurate and thorough metadata descriptors. Thus, corresponding metadata is fraught with errors and in some cases sequencing data are misclassified (156).

The European Bioinformatics Institute (EBI) sponsors the European Nucleotide Archive (ENA) sequence database through the European Molecular Biology Laboratory (EMBL) (163). Both NCBI and EMBL databases face problems with misclassified data submitted by the scientific community (177). These two publically available databases, in addition to the DNA Data Bank of Japan, are synchronized according to the International Nucleotide Sequence Database Collaboration. Effort has been made to curate known viral sequences into correctly-classified databases (i.e., ViPR (144)), which reduces computational time required to perform classifications, but can also limit the potential for viral discovery if the database is not comprehensive. Additionally, research facilities that maintain internal, unpublished databases limit continuity of research among the global scientific community.

Challenges and approaches to virome characterization

Reference genomes are essential for virome characterization because computational pipelines rely on comparison of the unknown query to previously described sequences. This can be accomplished by aligning (i.e., Burrows-Wheeler Alignment, BLAST) reads or contiguous sequences (contigs) that were generated by *de novo* assembly (contigs generated without the use of a reference genome) to a database of reference sequences (3; 107). Thresholds of homology are used to define the classification of the query to the reference sequence. The measurement of homology between two sequences is called identity. According to the International Committee for the Taxonomy of Viruses (ICTV), viruses with 80% amino acid sequence identity or higher are typically of the same species (75). Cutoffs for individual viral families vary and must take into consideration factors beyond genomic architecture, such as viral particle morphology. Therefore, with no consistent cutoff for sequence identity, the definition of a truly novel virus sequence can be ambiguous. While one challenge associated with novel virus identification is the cutoff for a given viral family, another difficult challenge is the altogether lack of sequenced near-phylogenetic neighbors (54). In instances where the sequence belongs to a novel virus, it may be necessary to join separately assembled contigs together that may belong to the same genome to create a longer, single contiguous sequence (called scaffolding). Another option is to use clustering to identify original sequences among a complex metagenomic dataset (109).

Cross-assembly among large datasets has previously been used as an approach in viral classification and enabled the discovery of crAssphage among sequenced dark matter (46; 47). In this approach, multiple sequence datasets from separate samples are utilized in *de novo* assembly and compared to identify common sequences that compose longer, more broadly covered contigs. Once a representative genome is obtained through cross-assembly, read mapping can be utilized to determine if variants are present in certain samples. This approach was utilized for this thesis in the initial discovery of Ruhugu virus and bat mumps-like virus. Virus discovery is a difficult process because it is computationally burdensome, typically involves an unwieldly amount of data, and lacks a standardized approach for the classification of dark sequence matter. As more innovative bioinformatic approaches emerge, computational biologists can continue to classify the previously unclassifiable sequences that are accumulating in publically available databases.

Perhaps an underappreciated yet crucial element of the metagenomic revolution is the accessibility of high-powered computing clusters. Metagenomic projects not only require intense computational power to perform initial analyses, but also require terabytes of storage capacity for metadata, raw data, intermediate products and finished analyses (129). The provision of relevant metadata is crucial to include in publically available database entries, and in many cases samples that lack methodological details cannot be used by the public without proper descriptors. A study published in 2016, titled “Optimizing viral discovery in bats,” excluded results from HTS experiments from a summary of known viruses detected in bats because sufficient details associated with the data were not available (206). Careful cataloging of samples and their associated metadata is necessary for metagenomic projects to fully benefit the wider scientific community and should not be overlooked.

As experimental approaches shift from viral isolation to conventional or quantitative techniques such as PCR, and most recently to HTS, sharing of pipelines for analysis and databases will be as crucial as having access to validated tools of the wet lab (i.e., primers). Endeavors to expand publically available databases will provide scientists with access to a greater body of information and further promote scientific curiosity. This will lead to new discoveries and leaps in our understanding of basic biology, as well as of the threat of newly emerging viral diseases (93).

ADVANTAGES AND LIMITATIONS OF VIRUS DETECTION METHODOLOGIES

Conventional biosurveillance methods

Biosurveillance has benefitted from the advancement of new methods that can be applied to the characterization of viromes (defined as all of the viruses that exist within or on a specific organism or niche – in the case of this study, a bat). Comprehensive biosurveillance methods are listed in Table 1 and discussed in detail in this section. Typical surveillance involves the study of known viruses through serological assays, conventional PCR or quantitative PCR (qPCR). Assays must be designed and validated using known specimens in advance of screening environmental samples to set a threshold by which to discern a "true positive." While the aforementioned assays are essential for biosurveillance and increasingly made more mobile for field work, they are each met by limitations. Serological assays, for example, can be performed in a high-throughput manner by using new technology such as MagPix, yet this assay does not address current infection and there is also a known risk of cross-neutralization across targets. Conventional PCR and qPCR suffer from primer drift, which can lead to false negatives. Additionally, limited sample volumes can preclude the possibility of optimizing and validating new (q)PCR assays for previously unknown viruses. Culture remains the gold standard in

Table 1. Uses & limitations of biosurveillance methods

Method	Sample type example	Computational burden	Approximate time requirement	Limitation	Product
Culture	Urine or homogenized tissue	None	Days	Requires known permissive cell line	Virus isolate
Serology	Blood	Low to moderate (throughput dependent)	Hours	Does not address current infection; potential cross-neutralization	Evidence of prior virus exposure
PCR	Variable (blood, urine, tissue, swab, etc.)	Low	Hours	Primer drift; unrealistic to optimize with limited sample volumes	Nucleic acid detection
qPCR		Low	Hours		Nucleic acid detection and semi-quantitation
Amplicon sequencing		Moderate	Days	Primer drift	Nucleic acid detection up to whole genome characterization
Targeted enrichment sequencing		Moderate to high	Days	Limited to near-neighbors or solely known viruses	
Shotgun sequencing		High	Days	Not quantitative; large amount of data produced	

pathogen detection; however, this approach can fail due to sample and growth condition limitations (54).

Serology and disease dynamics

Serological data provides a framework for understanding prior exposure to viruses, but are insufficient to discern the transmission patterns of viruses among bats in the wild. This problem has been discussed in detail by R. Plowright, who defined dynamics of virus transmission in the wild as potentially following three patterns: susceptible-infectious-latent-infectious (SILI), susceptible-infectious-recovered (SIR) or susceptible-infectious-recovered-susceptible (SIRS) patterns (147). SIR refers to the notion that infection is not persistent within the colony of bats and immunity is long-lasting. Periods with detectable viral infection would follow births or migration of new bats that are naïve to infection. Meanwhile, the second 'susceptible' component of SIR**S** implies that immunity is not long-lasting due to antibody decay in aging hosts that were previously exposed to the virus. The SILI pattern of infection implies that the virus persists within individuals and colonies long-term, even if shedding of the virus is below the limit of detection by biosurveillance assays. It is possible that each of these dynamics exists within a single community of bats in the context of different viruses (147).

The aforementioned concepts of disease dynamics are important to consider when addressing approaches to mitigate viral spillover. Longitudinal studies of bat populations have been proposed as a solution to understand pathogen dynamics in bat (147). By assessing viral abundance and the nature of genetic variation at different time points for evidence of foreign reintroduction of new strains and/or evolutionary mutations, the distinction between viral reintroduction through migration versus long-term maintenance within a bat colony can be discerned (Figure 2). If we do not completely understand viral dynamics within host populations, it will be difficult to appropriately determine protective measures. Intermittent shedding of virus is a major hurdle for detection (expressed in Line 2, Figure 2), which complicates any attempts to address the knowledge gap through noninvasive measures (147; 206). Serology and sequencing data in complement could address more problems together than when either is applied alone. Longitudinal HTS studies of bat populations, such as the studies presented in this thesis, can help close the knowledge gap; however, the vast amount of sampling and data processing required to fully understand disease dynamics will require extensive global contributions.

High throughput sequencing

Sequencing is becoming an increasingly portable and affordable approach for biosurveillance. The footprint of a high-throughput sequencing platform such as Illumina's MiSeq occupies approximately the same space as a desktop computer. Technological advances have resulted in smaller options such as Illumina's iSeq, which is about the size of a toaster oven. Yet smaller, Oxford Nanopore Technology's MinION is approximately the size of a cell phone. Despite concerns with regard to accuracy and throughput of newer devices, continued advances in sequencing technology will continue to make HTS more mobile and affordable. New methods are continually being developed that leverage the recent accessibility of sequencing platforms to interrogate difficult-to-sequence samples or unculturable microbes, making sequencing an adaptable approach that does not require *a priori* knowledge of the microbial content within a sample.

Targeted sequencing

Targeted sequencing describes both amplicon sequencing and hybridization-based target enrichment. While shotgun sequencing results in oversampling of the most abundant species, targeted sequencing mitigates that problem by enriching for known viruses of biosurveillance

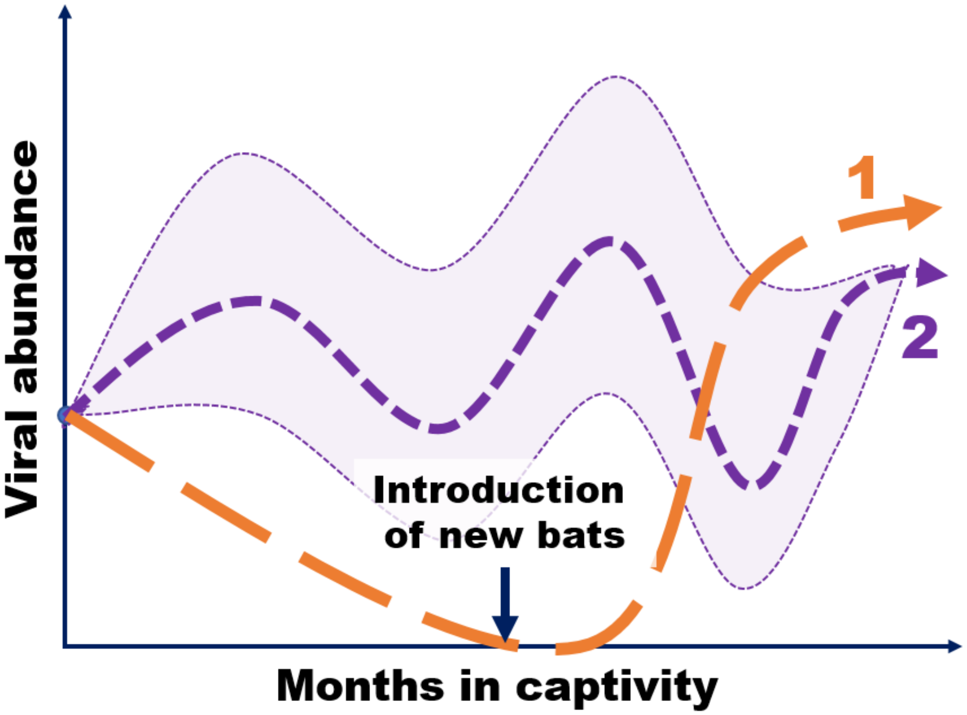


Figure 2. Surveillance of wild-caught bats does not distinguish between two proposed mechanisms of pathogen dynamics. 1) Viruses become extinct within a reservoir host population (colony) but are reintroduced through migration and interaction with new bats and 2) Viruses are maintained within a colony long-term (147). Definitive evidence that bats are persistently infected with emerging viruses could come from longitudinal studies of individual bats that are isolated from re-exposure.

concern. Amplicon sequencing utilizes primers designed against a genome of interest, followed by as many as 40 cycles of PCR. The target can be one conserved region of a genome at the species or family level or can include numerous regions that span the entire genome. This is a highly sensitive approach if the target sequence is conserved, but present at low levels in the sample; however, this method is vulnerable to false negatives due to the potential for primer drift that may hinder specific enrichment (54).

While amplicon sequencing is a useful approach for the detection of known viruses that are present at low titer, one potential problem with this approach can arise if the target virus is prone to high genetic variation. The use of conserved sequences across the genome, or specific conserved sequences such as a viral untranslated region (UTR) or RdRp, can aid the characterization of targeted viral species. For example, amplicon sequencing has been used to discern Dengue virus subtypes and avoid viral passage that could potentially introduce laboratory adaptions (30). This approach was highly sensitive; however, it failed to produce amplicons that spanned complete genomes due to the genetic variability of samples. The need for complete genomes depends on the research question and may not be necessary if the goal is simply virus detection or strain identification.

Hybridization-based target enrichment involves the use of 50-120mer oligonucleotide probes to pull down cDNA that corresponds to sequences of interest. Variations of this method can be performed pre- or post-sequencing library preparation. This targeted sequencing approach is more tolerant of primer mismatches due to the length of the oligonucleotide probes (16; 116; 127). We have demonstrated a tolerance for mismatches that is as low as 60-70% identity for several species (140). Probes are designed to tile the entire genome of target viruses and can be highly multiplexed. This method is highly effective and has been quickly implemented in recent biosurveillance efforts, including the recent outbreak of Ebola in West Africa (16; 116) and of Zika in the United States (68). In cases of metagenomic samples that may have low input DNA, one approach is to pool as many as 12 samples (beyond the recommended four) (140).

While the initial approach for hybridization-based enrichment was developed as a technique for human DNA exome sequencing, numerous variations for microbial sequencing protocols have been developed during the past five years. Freely available software has been developed by the Broad Institute that allow for the design of probes to suit unique research questions (126). In fact, this approach has been proposed by researchers at Washington University as a clinical diagnostic approach to supplement PCR assays (ViroCap) (200). Protocols vary by duration of probe hybridization and enrichment pre- or post-library preparation. Disadvantages include that PCR amplification is usually required for this approach, and protocols extend beyond one day of work.

Shotgun sequencing

Shotgun sequencing is unbiased and does not involve any form of targeted enrichment or specific primers for PCR during the preparation of a sequencing library. Unbiased HTS as used in this thesis was performed using Illumina MiSeq and NextSeq platforms. This approach allows for the detection of all domains of life, giving a full profile of microbiota. Output is typically 10-15 gigabytes of data from a MiSeq run and 120-160 gigabases (Gb) from a NextSeq run. Read lengths range from 150 to 600 base pairs. Illumina short read sequencing is the gold standard for HTS and recent improvements in read length have improved accuracy and usefulness for downstream analysis. The depth of sequencing provided by a NextSeq or HiSeq platform is accompanied by increased computational burden.

Data generated may be influenced by the choice of the sequencing platform and the library preparation method, and thus should be carefully selected based on the question to be answered by the study (161). Even for experiments that employ the state-of-the-art techniques, microbial "noise" in complex samples continues to be a major challenge in viral sequencing (54). An additional complication in viral sequencing is genome length. Preparation of samples for sequencing requires fragmentation of nucleic acid that is performed mechanically, enzymatically, or chemically to generate an average fragment size that is appropriate for the selected DNA sequencing method. In the case of metagenomic samples, fragmentation leads to an overrepresentation of nucleic acid fragments that originate from large genomes (i.e., from bacteria) compared to fragments that originate from small genomes (i.e., from viruses) that are then randomly sampled for sequencing (54). Viral genomes are typically thousands of base pairs (bp) long, while much larger bacterial genomes range from .5 to 10 Mbp and are easily overrepresented in the set of randomly sequenced nucleic acid fragments. In turn, smaller viral genomes are typically underrepresented in number of reads because the genomes are represented by fewer overall nucleic acid fragments (78). This problem can be accounted for by careful bioinformatic normalization to give "semi-quantitative" results. Careful normalization during bioinformatic analysis is essential to account for this preparation bias and results in output data that can be compared across batches and time (i.e., data homogenization).

GOALS AND SPECIFIC AIMS

With the ongoing concern for emerging and re-emerging infectious diseases, the overarching goal of this thesis was to determine the population persistence of viruses in bats that are related to human pathogens, and thereby inform future biosurveillance efforts. The specific aims of this study were to evaluate the efficacy of probe-based target enrichment sequencing in environmental samples, to characterize the virome of lesser dawn bats in Singapore and to characterize the virome of wild-caught cyclops roundleaf bats in Uganda. Three current insufficiencies in the field were identified and addressed through this thesis: 1) a lack of validation of hybridization-based target enrichment of environmental samples (addressed by methods characterization); 2) the unknown diversity of the virome of an understudied nectar-feeding bat (addressed by virome characterization); and 3) the population persistence of viruses related to human pathogens within a closed population of bats (addressed by colony-level trend evaluation).

To test the performance of a target enrichment method for viruses of biosurveillance concern, both mock spiked samples and environmental samples were studied. As will be described in detail in Chapter Two, contrived mixtures containing combinations of viral strains in a background of bat guano or human serum were evaluated by comparing results of target enrichment and shotgun sequencing processed from aliquots of the same samples. Sensitivity of the hybridization-based target enrichment was evaluated by calculating the fold-enrichment of spiked-in virus as compared to shotgun sequencing. Limit of detection was evaluated by testing a range of spiked-in genome equivalents of influenza A virus. Strain discrimination was evaluated by spiking multiple serotypes of dengue virus into a background of human serum at varying estimated genome equivalents. As described in Chapter Three, the utility of probes for the characterization of unknown viruses was evaluated using environmental samples (bat swabs). Shotgun and target enrichment data were generated using RNA extracted from head, body, oral and rectal swabs. The efficacy of target enrichment in real world samples was compared by evaluating the number of detected zoonotic-related viruses in enriched data as compared to shotgun data.

To characterize the virome of lesser dawn bats in Singapore (described in Chapter Three), all detectable viruses were cataloged and semi-quantitatively evaluated for changes in abundance within a captive colony of lesser dawn bats sampled at six time points over the course of 18 months (illustrated in Figure 3). Lesser dawn bats in South-eastern Asia were previously known to host henipaviruses and were also predicted as a host for filoviruses (70). A deeper investigation of one virus discovered in the colony (described in Chapter Four), an unusual cross-family recombinant coronavirus, elucidated that there was very little genetic variation within that particular virus. Orthogonal computational approaches were utilized to determine the zoonotic-related taxonomic members of each sample. Internal (oral and rectal) as well as external (head and body) swabs were collected for this study. External swab sites are illustrated in Figure 4. Semi-quantitative calculations using the normalized abundance of viral reads based on classification by VirusSeeker (209), a virus discovery pipeline, were used for analyses in Chapter Three. This approach not only allowed for the confident cataloging of known zoonotic-related viruses, but also revealed longitudinal patterns within the colony. Furthermore, these data provided support for the utility of external swab collection for biosurveillance purposes.

To determine the capacity for population persistence of viruses related to human pathogens, and thus guide future biosurveillance efforts, the normalized abundance of viruses detected in the captive colony was compared among time points. Multiple zoonotic-related viruses, including several that have been associated with spillover from bats to humans, were detected long-term within the colony. Unsupervised clustering was used to further investigate common properties of the most frequently detected and longest-persisting zoonotic-related

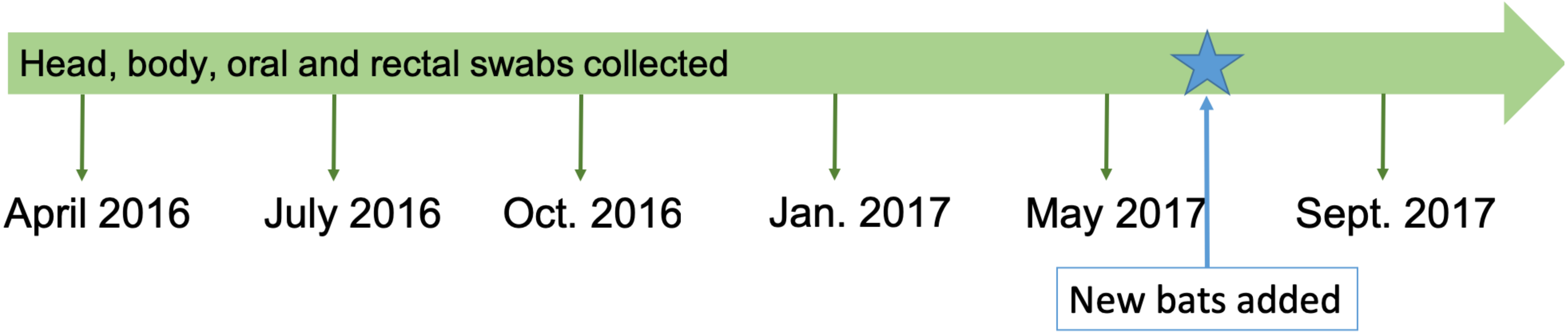


Figure 3. Study timeline. Head, body, oral and rectal swabs were collected every three or four months beginning on April 2016, when wild-caught lesser dawn bats were brought into captivity in Singapore. Newly-caught bats that were captured from the same original location were added to cages in close proximity to the colony in summer 2017.

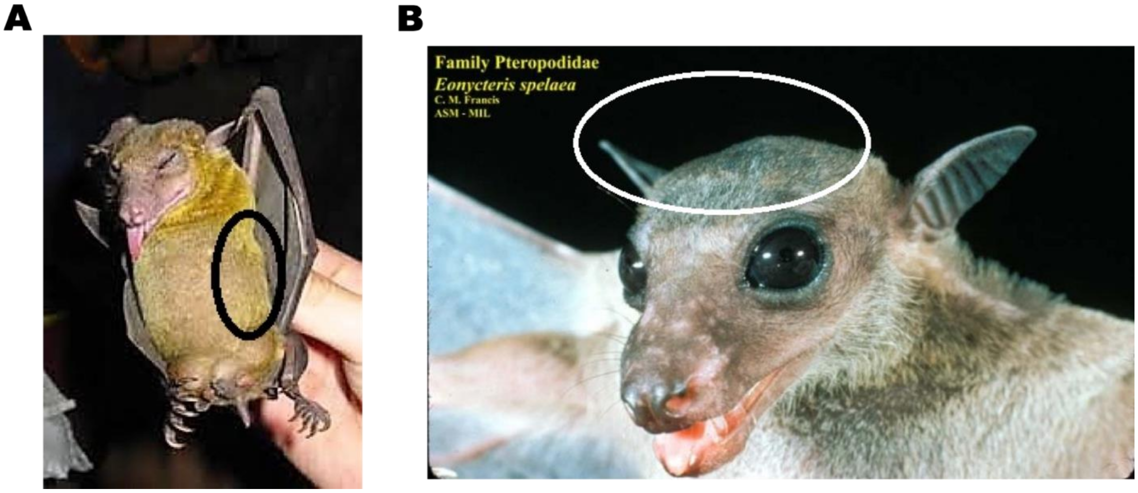


Figure 4. External swab sites. Panel A shows the swab site for body swabs and panel B shows the swab site for head swabs.

viruses. These observations can be used to prioritize biosurveillance methods by optimizing assays toward efficient sample collection that targets relevant viral species. Chapter Five discusses the discovery of a novel Rubella-like virus, Ruhugu virus. This discovery is a valuable addition to the body of knowledge surrounding Rubella viruses due to potential interference with global eradication efforts of Rubella virus. The high prevalence of Ruhugu virus in apparently healthy cyclops roundleaf bats suggests that bats are reservoir hosts of Ruhugu virus and raises the possibly that bats could be hosts of other ancient rubiviruses such as the progenitor of human Rubella virus.

This thesis shows that HTS is an efficient method for detecting and characterizing both known and previously unknown viruses. The work explored customizable approaches that can be adapted based on biosurveillance concern or research question. It also catalogued 53 new zoonotic agent-related viruses, including multiple with significant relevance to U.S. Department of Defense public health laboratories due to the implications for disease prevention and control.

CHAPTER 2: Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples

This work published as: **Paskey, A.C., Frey, K.G., Schroth, G., Gross, S., Hamilton, T. and Bishop-Lilly, K.A., 2019.** “Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples.” *BMC genomics*, *20*(1), p.155.

The work presented here is the sole work of A.C.P. with the following exceptions: G.S. and S.G. designed the probe set. K.G.F. and K.A.B.-L. selected viral targets and assisted with experimental design.

ABSTRACT

Background:

Sequencing-based detection and characterization of viruses in complex samples can suffer from lack of sensitivity due to a variety of factors including, but not limited to, low titer, small genome size, and contribution of host or environmental nucleic acids. Hybridization-based target enrichment is one potential method for increasing the sensitivity of viral detection via high-throughput sequencing.

Results:

This study expands upon two previously developed panels of virus enrichment probes (for filoviruses and for respiratory viruses) to include other viruses of biodefense and/or biosurveillance concern to the U.S. Department of Defense and various international public health agencies. The newly expanded and combined panel is tested using carefully constructed synthetic metagenomic samples that contain clinically

relevant amounts of viral genetic material. Target enrichment results in a dramatic increase in sensitivity for virus detection as compared to shotgun sequencing, yielding full, deeply covered viral genomes from materials with Ct values suggesting that amplicon sequencing would be likely to fail. Increased pooling to improve cost- and time-effectiveness does not negatively affect the ability to obtain full-length viral genomes, even in the case of co-infections, although as expected, it does decrease depth of coverage.

Conclusions:

Hybridization-based target enrichment is an effective solution to obtain full-length viral genomes for samples from which virus detection would fail via unbiased, shotgun sequencing or even via amplicon sequencing. As the development and testing of probe sets for viral target enrichment expands and continues, the application of this technique, in conjunction with deeper pooling strategies, could make high-throughput sequencing more economical for routine use in biosurveillance, biodefense and outbreak investigations.

BACKGROUND

High-Throughput Sequencing (HTS), also known as Next-Generation Sequencing (NGS), has many advantages for pathogen detection as compared to traditional methods such as Polymerase Chain Reaction (PCR), serological assays, and/or culture-based methods. Metagenomic sequencing is the high-throughput sequencing of nucleic acid from complex samples rather than from purified microorganisms. Metagenomic sequencing is much less biased than other methods and allows for the detection of fastidious or nonculturable organisms as well as multiple unrelated pathogens within a single sample (54). Moreover, detection via HTS is much less susceptible to false-negative results caused by antigenic drift or signature erosion. Despite these advantages, one of the technical challenges encountered with respect to metagenomic sequencing is obtaining adequate depth and breadth of coverage from pathogens like RNA viruses that i) typically have small genomes, and ii) are typically present at low titers amidst the background 'noise' of the host and commensals (54). Genome size directly affects sensitivity of detection by HTS because the sampling of sequence fragments within a sample depends on the prevalence of those fragments and organisms with larger genomes typically contribute more fragments, therefore being sampled more often than organisms with smaller genomes. In other words, organisms with larger genomes have the potential to contribute a larger proportion to the overall number of sequencing reads even when the plaque-forming units (PFU), or colony-forming units (CFU) in the case of bacteria, are equivalent to that of an organism with a smaller sized genome.

Although conventional shotgun sequencing allows for the detection of all domains of life, it rarely returns robust coverage of a small viral genome when taken from a very complex sample. A variety of possible strategies exist to enhance the sensitivity of HTS for virus detection and characterization, including purification of specific viral fractions by physical methods such as filtration and ultracentrifugation (170), amplicon-based target enrichment, and hybridization-based target enrichment. Purification of viral fractions is ideal in some cases, although it can be laborious, and for certain sized samples (for instance clinical samples of very limited volume) it may not be realistic. The use of hybridization-based target enrichment could be preferable to the aforementioned technologies because it has the potential to yield sequence data covering the entire genome of multiple viruses with just one sequencing reaction by using genome-wide probes designed against multiple viruses to specifically select for viral cDNA prior to sequencing. Amplicon sequencing of viral genomes is a technique that has been widely used, but it has some disadvantages, which were articulated by Metsky *et al.* in a recent study of Zika virus (ZIKV) (125). First, traditional amplicon sequencing typically requires technically challenging normalization and pooling of individual amplicons to cover the entire genome of one specific virus. However, recently a protocol was published for efficient amplicon primer design and multiplex amplicon generation in a single tube for sequencing in the MinION or Illumina platforms (150). Although this method obviates the amplicon normalization and pooling steps and is effective for producing whole genome sequence data from a low titer ZIKV sample, this method has not been demonstrated for production of whole genome sequence data for multiple diverse viruses from a single complex sample. Additionally, amplicon sequencing typically requires as much as 40 cycles of PCR amplification (30; 105; 150), which can introduce sequence errors. Furthermore, amplicon-based sequencing is vulnerable to false negative results caused by mutations in primer binding sites, as was recently demonstrated for Dengue virus (DENV) (30) as well as false positive variant results possibly caused by low and/or uneven coverage (155). By contrast, the use of probes tiled along the entire length of a viral genome to hybridize and select for virus-specific fragments has the potential to produce less false-negative pathogen detection results by virtue of many more potential binding sites along an individual genome and resulting more uniform coverage.

Ebola virus (EBOV) is one specific example of a pathogen for which false-negative PCR results can have devastating consequences and for which available PCR-based assay effectiveness has been shown to be affected by drift (162). Therefore, a panel of 80-mer oligonucleotide probes designed against eight Filovirus genomes was recently used for post-sequencing library enrichment in a HTS-based study of a recent EBOV outbreak in West Africa and in an investigation of potential genetic variation of EBOV in experimentally infected nonhuman primates (16; 113; 116). In this protocol, viral enrichment is coupled with the RNA Access kit, developed by Illumina, Inc. The technical advancements of the RNA Access kit had already enabled the sequencing of previously unsequencable materials such as those of low concentration and formalin-fixed paraffin-embedded (FFPE) tissue (83), and now this protocol has been employed not only for the detection and characterization of EBOV from clinical samples but also for detection and characterization of respiratory viruses in clinical samples (133; 204). In general, hybridization-based viral target enrichment has been successfully employed to characterize viruses found within both contrived samples and clinical samples (16; 17; 32; 39; 116; 125; 127; 133; 134; 200). The performance of the Respiratory Virus Panel (RVP) version of this method (204), which uses probes for 34 common respiratory viruses in conjunction with the TruSeq RNA Access protocol, was recently investigated and it was demonstrated to work well overall when tested on human clinical samples (133). Specifically, the authors reported successful enrichment for 30 of 33 human clinical samples tested. Importantly, RT-PCR was conducted on those same samples and Ct values of respiratory viruses in those clinical samples ranged from 21 to 33 (133), which provides a framework for beginning to assess the limits of detection of hybridization-based enrichment sequencing. Herein, we extend this approach by i) expanding this viral probe panel to include viruses of biosurveillance and biodefense concern and ii) employing carefully constructed mock clinical samples to systematically assess this technique's performance in a variety of conditions, such as deeper multiplexing for cost effectiveness as well as more extensive co-infection scenarios. We demonstrate the sensitivity and reproducibility of hybridization-based viral enrichment sequencing despite virus divergence and we show that this sensitivity is maintained even with extensive multiplexing of samples to decrease cost. Herein we demonstrate that within one reaction tube, this technique can even be used to detect and discriminate between multiple serotypes of a virus within a clinical sample or to detect and discriminate amongst multiple unrelated viruses that present similar clinical symptoms, and we demonstrate this performance at clinically relevant concentrations of virus.

METHODS

Preparation of contrived metagenomic samples and nucleic acid extraction and quality control

IFV (H1N1) particles (A/Swine/Iowa/15/30; ATCC, Manassas, VA), MERS-CoV RNA (Jordan-N3/2012; NAMRU-3), HAdV nucleic acid extract from particles (RI-67 and Bom; ATCC, Manassas, VA), ZIKV RNA extract from particles (MR766 and R116265; ATCC, Manassas, VA), CHIKV RNA (gift from LTC Richard Jarman, Walter Reed Army Institute of Research), DENV-1 RNA extract from particles (TH-SMAN; ATCC, Manassas, VA) and DENV-2 RNA extract from particles (New Guinea C; ATCC, Manassas, VA) were spiked into relevant matrices to construct contrived metagenomic samples for testing.

To prepare guano samples five-gram quantities of commercial Jamaican bat guano (Planet Natural; Bozeman, MT) were placed in 50 mL of sterile-filtered Hank's Balanced Salt Solution (HBSS), vortexed to mix, centrifuged at 3,100 x g for 10 min, and filtered sequentially through 0.45 µm and 0.22 µm filters prior to spiking with IFV particles. Post-addition of IFV, samples were centrifuged at 39,000 x g for three hours at 10°C to concentrate spiked and native virus particles, supernatant was removed, and total RNA was extracted from the pellet using the QIAampViral RNA Isolation Kit (QIAGEN; Valencia, CA). After elution in 30 µL buffer AVE, a second elution using 20 µL of the eluate was performed.

To prepare cell culture matrix samples, a nucleic acid spiking approach was used. In this case, decreasing amounts of IFV RNA were spiked into a constant mass of total RNA that had been extracted from Vero cells infected with MERS-CoV. Aliquots of the same Vero cell culture were used for each sample. Genome equivalents of IFV spiked into samples were calculated based on RNA mass extracted from virus particles and genome size.

For serum samples, RNA was extracted from Human Serum (BioIVT, Westbury, NY) and mixed with viral RNA.Total nucleic acid was extracted from adenovirus particles using the QIAGEN QiAMP MinElute Virus Spin Kit, omitting carrier RNA. The samples were eluted in 24 µl buffer AVE. Viral RNA was extracted from virus particles and human serum using the QIAampViral RNA Isolation Kit as described above. The Qubit double stranded DNA Broad-Range Assay Kit and the Qubit RNA Broad-Range Assay Kit (Thermo Fisher Scientific; Waltham, MA) were used to assay extracts.

Quantitative Reverse Transcription PCR

For quantitative reverse transcription PCR, SuperScriptIII RT/Platinum *Taq* Mix (Thermo Fisher Scientific; Waltham, MA), dNTP mix, $MgSO_4$, ROX Reference Dye (Thermo Fisher Scientific; Waltham, MA), and the primer and probes listed in Table 2 were used to assay in the CFX Connect Real-Time PCR Detection System (Bio-Rad; Hercules, CA) using the following conditions: 50˚C for 15 minutes, 95˚C for two minutes, and 40 cycles of 95˚C for 15 seconds and 60˚C for one minute. Standard curves for DENV-1, DENV-2, CHIKV and ZIKV were generated from titrated viral RNA extract.

Library preparation, virus enrichment, and sequencing

For virus enriched sequencing, TruSeq RNA Access libraries were created as per manufacturer's protocol (Illumina; San Diego, CA), with the following two modifications: i) rather than the standard CEX oligonucleotides that are designed for enrichment of human genes, a custom pool of oligonucleotides was used that includes probes along the entire genome length of 84 viruses (Table 3) as well as probes specific for several human house-keeping genes, and ii) in the second PCR amplification, 17 cycles were used rather than ten. Samples were probed singly or in pools of four or 12 and multiplexed for sequencing on the Illumina MiSeq platform using v3 chemistry, 2X75 bp read lengths.

For conventional HTS (shotgun) sequencing, TruSeq libraries were pooled and sequenced on the MiSeq platform using v3 chemistry, 2X300 bp read lengths. For the guano samples, these consisted of an aliquot of each of the TruSeq RNA Access libraries from the step prior to virus enrichment. For the MERS-CoV-Vero cell matrix samples, these consisted of shotgun libraries made with the TruSeq RiboZero Gold Library Preparation Kit (Illumina; San Diego, CA).

Virus enrichment probe design

A composite panel of 80-mer DNA probes was assembled using a previously described panel for respiratory viruses (39; 133; 204), a previously described panel for Filoviruses (16; 113; 116), plus an additional panel of newly designed probes for 41 viruses of biosurveillance and biodefense concern, for a total of 19,077 probes. The methods employed for capture oligo design were essentially as described in O'Flaherty at al (133), although the design varied somewhat across the target viral genomes. For instance, in the case of the previously described respiratory virus panel, the design was focused on coding regions (133). In general, genomes were tiled with capture oligos in a way so as to avoid low-complexity sequences and repetitive sequences. Probe spacing and overlap vary per virus due to attempts to design probes that cover multiple related virus strains resulting in overlapping tiled design around more variable regions, whereas regions more conserved among multiple strains resulted in probes more or less tiled end-to-end. All probes were biotinylated on the 5' end. Sequences of viral capture probes are provided in Additional File 5 of this published work.

Bioinformatic analyses

Quality control, *de novo* assembly, taxonomic classification, and reference-based analyses were conducted using EDGE Bioinformatic software v 2.0 (108) with default parameters and host removal of human reference GRCh38 and CLC Genomics Workbench v11 (QIAGEN Bioinformatics; Redwood City, CA). The reference mapping

Table 2. Primers and probes used for qRT-PCR

Virus	**Forward Primer Sequence**	**Reverse Primer Sequence**	**Probe Sequence**	**Reference**
CHIKV (181/Clone 25)	AGCTCCGCGTCCTTTACCA	GCCAAATTGTCCTGGTCTTCCT	Express One SYBR Green I (Life Technologies)	(176)
DENV-1 (TH-SMAN)	GACACCACACCCTTTGGACAA	CACCTGGCTGTCACCTCCAT	FAM-AGAGGGTGTTTAAAGAGAAAGTTGACACGCG-TAMRA	(21)
DENV-2 (New Guinea C)	ACAGGCTATGGCACTGTTACGAT	TGCAGCAACACCATCTCATTG	FAM-AGTGCTCTCCAAGAACGGGCCTCG-TAMRA	(153; 175)
IFV-A – HA (A/Swine/Iowa/15/30)	CCAGTCACAATAGGAGAGTG	AAACCGGCAATGGCTCCAAA	Express One SYBR Green I (Life Technologies)	(55)
ZIKV (MR766 and R116265)	AARTACACATACCARAACAAAGTGG**T***	TCCRCTCCCYCTYTGGTCTTG	Express One SYBR Green I (Life Technologies)	(49)

*boldface **T** was modified from originally published **R** parameters in CLC were modified from defalt settings to 0.8 length fraction and 0.8 similarity fraction with global alignment and random mapping of non-specific matches.

BLAST (3) was also used to further investigate specific datasets. The ggplot2 R package was used to generate depth of coverage plots (187).

RESULTS

Hybridization-based target enrichment enhances sensitivity of HTS for detection of virus in complex environmental samples.

Enhanced detection of viruses from various clinical sample types using filovirus- or respiratory virus-specific probes has recently been demonstrated (16; 116; 133). To expand the range of viruses that could be detected, in this study those two probe panels were combined with new probes for 41 additional viruses that are of biosurveillance and biodefense concern, for a full panel targeting 84 diverse viruses (Table 3). In order to test this newly expanded probe panel and to specifically assess the effect of hybridization-based viral enrichment on the sensitivity of HTS for detection of a single virus within a complex environmental sample, commercial bat guano was spiked with increasing concentrations of Influenza virus (IFV). Spiked samples were split into two parts each, with each part being processed in parallel with unbiased, shotgun sequencing versus target enrichment sequencing using an expanded panel of probes.

As expected, a dose-dependent effect in the proportion of sequencing reads derived from IFV was observed as the number of spiked genome copies increased (Figure 5A and Table 4), in both the unbiased shotgun sequence data as well as the virus enriched sequence data. However, in this context, hybridization-based target enrichment resulted in approximately 20- to 100-fold more sensitivity for detection of IFV as compared to

Table 3. Viruses included in target enrichment panel

Virus	Genome size (kb)	Genome type	NCBI accession(s) of reference used in probe design	Notes
Nipah virus	18,246	Negative sense ssRNA	NC_002728.1	New addition to probe panel
Bat Paramyxovirus	18,530	Negative sense ssRNA	NC_025256.1	New addition to probe panel
Cedar virus	18,162	Negative sense ssRNA	JQ001776.1	New addition to probe panel
Hendra virus	18,234	Negative sense ssRNA	NC_001906.3	New addition to probe panel
Tioman virus	15,522	Negative sense ssRNA	NC_004074.1	New addition to probe panel
Menangle virus	15,516	Negative sense ssRNA	NC_007620.1	New addition to probe panel
Middle East Respiratory Syndrome Coronavirus	30,094	Positive sense ssRNA	KJ614529.1	New addition to probe panel
Severe Acute Respiratory Syndrome virus	29,751	Positive sense ssRNA	NC_004718.3	New addition to probe panel
Lujo virus	10,352	Negative sense ssRNA	NC_012776.1, NC_012777.1	New addition to probe panel
Lassa fever virus	10,681	Negative sense ssRNA	NC_004296.1, NC_004297.1	New addition to probe panel
Machupo virus	10,635	Negative sense ssRNA	NC_005078.1, NC_005079.1	New addition to probe panel
Junin virus	10,525	Negative sense ssRNA	NC_005080.1, NC_005081.1	New addition to probe panel
Guanarito virus	10,424	Negative sense ssRNA	NC_005077.1, NC_005082.1	New addition to probe panel
Chapare virus	10,464	Negative sense ssRNA	NC_010562.1, NC_010563.1	New addition to probe panel
Sabia virus	10,499	Negative sense ssRNA	NC_006313.1, NC_006317.1	New addition to probe panel
Hantaan virus	11,845	Negative sense ssRNA	NC_005218.1, NC_005219.1, NC_005222.1	New addition to probe panel
Puumala virus	12,062	Negative sense ssRNA	NC_005223.1, NC_005224.1, NC_005225.1	New addition to probe panel

Sin nombre virus	12,317	Negative sense ssRNA	NC_005215.1, NC_005216.1, NC_005217.1	New addition to probe panel
Andes virus	12,104	Negative sense ssRNA	NC_003466.1, NC_003467.1, NC_003468.1	New addition to probe panel
Rift Valley fever virus	11,979	Negative sense ssRNA	NC_014395.1, NC_014396.1, NC_014397.1	New addition to probe panel
Crimean Congo hemorrhagic fever virus	19,146	Negative sense ssRNA	NC_005300.2, NC_005301.3, NC_005302.1	New addition to probe panel
Omsk hemorrhagic fever virus	10,787	Positive sense ssRNA	NC_005062.1	New addition to probe panel
Kyasanur forest disease virus	10,774	Positive sense ssRNA	JF416958.1	New addition to probe panel
Alkhurma hemorrhagic fever virus	10,685	Positive sense ssRNA	NC_004355.1	New addition to probe panel
Eastern equine encephalitis virus	11,703	Positive sense ssRNA	KJ469643.1	New addition to probe panel
Dengue type 1 virus	10,721	Positive sense ssRNA	AF309641.1	New addition to probe panel
Dengue type 2 virus	10,723	Positive sense ssRNA	EF051521.1	New addition to probe panel
Dengue type 3 virus	10,707	Positive sense ssRNA	AY662691	New addition to probe panel
Dengue type 4 virus	10,653	Positive sense ssRNA	AY618989	New addition to probe panel
Chikungunya virus	11,826	Positive sense ssRNA	NC_004162	New addition to probe panel
Bat coronavirus CDPHE15	28,035	Positive sense ssRNA	NC_022103.1	New addition to probe panel
Bat coronavirus 1A	28,326	Positive sense ssRNA	NC_010437.1	New addition to probe panel
Bat coronavirus 1B	28,476	Positive sense ssRNA	NC_010436.1	New addition to probe panel
Bat coronavirus HKU2	27,165	Positive sense ssRNA	NC_009988.1	New addition to probe panel
Bat SARS coronavirus HKU3-4	29,704	Positive sense ssRNA	GQ153539.1	New addition to probe panel
Bat coronavirus HKU4	30,286	Positive sense ssRNA	NC_009019	New addition to probe panel
Bat coronavirus HKU5-1	30,482	Positive sense ssRNA	NC_009020	New addition to probe panel

Bat coronavirus HKU8	28,773	Positive sense ssRNA	NC_010438.1	New addition to probe panel
Bat coronavirus HKU9-1	29,114	Positive sense ssRNA	NC_009021.1	New addition to probe panel
Bat coronavirus HKU10	28,494	Positive sense ssRNA	NC_018871.1	New addition to probe panel
Zika virus	10,794	Positive sense ssRNA	NC_012532	New addition to probe panel
Respiratory Syncytial virus B (S2)	15,190	Negative sense ssRNA	NC_001803.1	Previously used in (39; 133)
Respiratory Syncytial virus A	15,225	Negative sense ssRNA	AY353550	Previously used in (39; 133)
Influenza virus A (H9N2)	13,500	Negative sense ssRNA	NC_004905.2, NC_004906.1, NC_004907.1, NC_004908.1, NC_004909.1, NC_004910.1, NC_004911.1, NC_004912.1	Previously used in (39; 133)
Influenza virus A (H2N2)	13,460	Negative sense ssRNA	NC_007374.1, NC_007375.1, NC_007376.1, NC_007377.1, NC_007378.1, NC_007380.1, NC_007381.1, NC_007382.1	Previously used in (39; 133)
Influenza virus A (H3N2)	13,630	Negative sense ssRNA	NC_007366.1, NC_007367.1, NC_007368.1, NC_007369.1, NC_007370.1, NC_007371.1, NC_007372.1, NC_007373.1	Previously used in (39; 133)
Influenza virus A (H1N1)	13,590	Negative sense ssRNA	NC_002016.1, NC_002017.1, NC_002018.1, NC_002019.1, NC_002020.1, NC_002021.1, NC_002022.1, NC_002023.1,	Previously used in (39; 133)

Influenza virus A (H5N1)	13,590	Negative sense ssRNA	NC_007357.1, NC_007358.1, NC_007359.1, NC_007360.1, NC_007361.1, NC_007362.1, NC_007363.1, NC_007364.1	Previously used in (39; 133)
Influenza virus A (H7N9)	13,590	Negative sense ssRNA	KC885955, KC885956, KC885957, KC885958, KC885959, KC885960, KC885961, KC885962	Previously used in (39; 133)
Influenza virus B	14,450	Negative sense ssRNA	NC_002204.1, NC_002205.1, NC_002206.1, NC_002207.1, NC_002208.1, NC_002209.1, NC_002210.1, NC_002211.1	Previously used in (39; 133)
Parainfluenza virus 1	15,600	Negative sense ssRNA	NC_003461.1	Previously used in (39; 133)
Parainfluenza virus 2	15,650	Negative sense ssRNA	NC_003443.1	Previously used in (39; 133)
Parainfluenza virus 3	15,460	Negative sense ssRNA	NC_001796.2	Previously used in (39; 133)
Parainfluenza virus 4	17,050	Negative sense ssRNA	NC_021928.1	Previously used in (39; 133)
Human metapneumovirus	13,340	Negative sense ssRNA	NC_004148.2	Previously used in (39; 133)
Adenovirus C	35,937	dsDNA	NC_001405.1	Previously used in (39; 133)
Adenovirus B	35,343	dsDNA	NC_011203.1	Previously used in (39; 133)
Adenovirus E	35,994	dsDNA	NC_003266.2	Previously used in (39; 133)
Human Coronavirus HKU1	29,930	Positive sense ssRNA	NC_006577.2	Previously used in (39; 133)
Human Coronavirus NL63	27,550	Positive sense ssRNA	NC_005831.2	Previously used in (39; 133)
Human Coronavirus 229E	27,320	Positive sense ssRNA	NC_002645.1	Previously used in (39; 133)

Human Coronavirus OC43	30,738	Positive sense ssRNA	AY391777.1	Previously used in (39; 133)
Rhinovirus A	7,150	Positive sense ssRNA	NC_001617.1	Previously used in (39; 133)
Rhinovirus C	7,100	Positive sense ssRNA	NC_001490.1	Previously used in (39; 133)
Rhinovirus B14	7,210	Positive sense ssRNA	NC_001490.1	Previously used in (39; 133)
Human Bocavirus 1	5,299	ssDNA	NC_007455.1	Previously used in (39; 133)
Human Bocavirus 2	5,196	ssDNA	NC_012042.1	Previously used in (39; 133)
Human Bocavirus 3	5,242	ssDNA	NC_012564.1	Previously used in (39; 133)
Human Bocavirus 4	5,104	ssDNA	NC_012729.2	Previously used in (39; 133)
KI polyomavirus	5,040	dsDNA	NC_009238.1	Previously used in (39; 133)
WU polyomavirus	5,229	dsDNA	NC_009539.1	Previously used in (39; 133)
Human parechovirus type 1	7,296	Positive sense ssRNA	FM242866.1	Previously used in (39; 133)
Human parechovirus type 6	7,347	Positive sense ssRNA	AB252582.1	Previously used in (39; 133)
Human Enterovirus C104	7,408	Positive sense ssRNA	AB686524.1	Previously used in (39; 133)
Human Enterovirus C109	7,354	Positive sense ssRNA	GQ865517.1	Previously used in (39; 133)
Lloviu cuevavirus	18,927	Negative sense ssRNA	NC_016144	Previously used in (16; 113; 116)
Bundibugyo ebolavirus	18,940	Negative sense ssRNA	NC_014373	Previously used in (16; 113; 116)
Zaire ebolavirus	18,959	Negative sense ssRNA	NC_002549	Previously used in (16; 113; 116)
Reston ebolavirus	18,891	Negative sense ssRNA	NC_004161	Previously used in (16; 113; 116)
Sudan ebolavirus	18,875	Negative sense ssRNA	NC_006432	Previously used in (16; 113; 116)
Tai Forest ebolavirus	18,935	Negative sense ssRNA	NC_014372	Previously used in (16; 113; 116)
Marburg virus (isolate Marburg virus)	19,111	Negative sense ssRNA	NC_001608	Previously used in (16; 113; 116)
Marburg virus (isolate Ravn virus)	19,114	Negative sense ssRNA	NC_024781	Previously used in (16; 113; 116)

detection via unbiased, shotgun sequencing. At the lowest concentration tested (1,250 genome equivalents (GE) per mL), only 0.5% of sequencing reads produced by unbiased shotgun sequencing were derived from IFV (‘on target’ reads), whereas by stark contrast, the majority of reads produced by target enrichment sequencing (54.4%) were derived from IFV.

Given the dramatic increase in sensitivity observed when complex samples were spiked with an individual virus’s genetic material and subjected to target enrichment, we next sought to evaluate whether these effects would still be observed in the presence of an additional virus and at lower concentrations of IFV gRNA overall. Therefore, IFV gRNA was spiked into total RNA derived from Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV) cell culture lysate at an overall lower range of increasing concentrations than IFV was spiked in the prior experiment. As before, the samples were aliquoted into two parts that were processed by each method. In these synthetic co-infection samples, target enrichment sequencing resulted in a simultaneous increase in sensitivity for both viruses as compared to unbiased, shotgun sequencing (Figure 5B). As expected, a constant high proportion of reads mapping to MERS-CoV was observed and the proportion of reads mapping to IFV increased in a dose-dependent fashion with the number of genome equivalents spiked (Table 4).

Detection and discrimination of related viruses in clinical samples.

We next tested the sensitivity for detection of three clinically-relevant viruses that can co-circulate in tropical regions, can present with similar symptoms, and can be

Figure 5. Hybridization-based enrichment enhances sensitivity of HTS for viral detection from complex samples. Known concentrations of IFV genomic RNA (gRNA) were spiked into complex matrices. Samples were split into two parts and processed in parallel via unbiased, shotgun sequencing or target enrichment sequencing in pools of four. A IFV was spiked into bat guano at increasing concentrations to simulate environmental-type samples. Shown here is the percentage of IFV-specific reads. B Increasing concentrations of IFV gRNA were spiked into total RNA derived from MERS-CoV cell culture lysate. MERS-CoV genomic material was present at a constant, high level amongst all samples. The average percentage of IFV and MERS-CoV virus-specific reads derived from three biological replicates is shown. Black bars denote standard error of the mean for each sample.

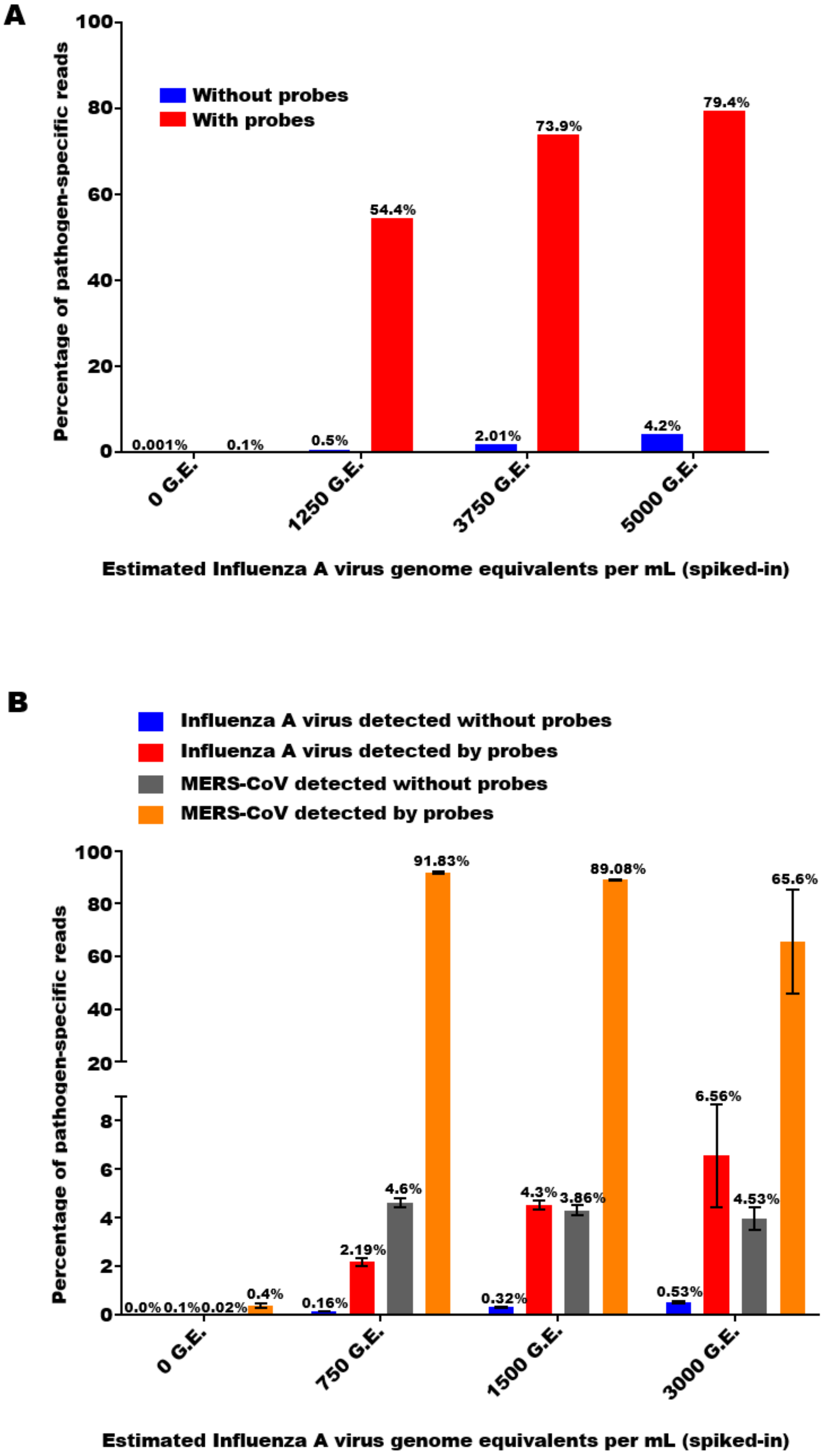

A
Percentage of pathogen-specific reads
100
80
60
40
20
0
Without probes
With probes
0.001%
0.1%
0.5%
54.4%
2.01%
73.9%
4.2%
79.4%
0 G.E.
1250 G.E.
3750 G.E.
5000 G.E.
Estimated Influenza A virus genome equivalents per mL (spiked-in)
B
Influenza A virus detected without probes
Influenza A virus detected by probes
MERS-CoV detected without probes
MERS-CoV detected by probes
Percentage of pathogen-specific reads
100
80
60
40
20
8
6
4
2
0
0.0%
0.1%
0.02%
0.4%
0.16%
2.19%
4.6%
91.83%
0.32%
4.3%
3.86%
89.08%
0.53%
6.56%
4.53%
65.6%
0 G.E.
750 G.E.
1500 G.E.
3000 G.E.
Estimated Influenza A virus genome equivalents per mL (spiked-in)

Table 4. Number of reads mapped to IFV

Spike-in level (IFV GE)	Enriched		Shotgun	
	Number of reads mapped to IFV (%)	Total number of reads	Number of reads mapped to IFV (%)	Total number of reads
0	1,551 (0.1)	950,334	88 (0.0)	7,408,474
1,250	2,004,766 (54.4)	3,683,304	31,783 (0.5)	6,295,476
3,750	2,828,992 (73.9)	3,825,462	160,727 (2.0)	7,964,390
5,000	9,431,355 (79.4)	11,879,278	161,078 (4.2)	3,820,210

Table 5. Number of reads mapped to MERS-CoV

Spike-in level (IFV GE)	Enriched				Shotgun			
	Replicate	Number of reads mapped to MERS-CoV (%)	Number of reads mapped to IFV (%)	Total number of reads	Replicate	Number of reads mapped to MERS-CoV (%)	Number of reads mapped to IFV (%)	Total number of reads
0	1	2,155 (0.6)	101 (0.0)	387,990	1	39 (0.0)	4 (0.0)	1,561,344
	2	174 (0.3)	6 (0.0)	63,176	2	371 (0.0)	26 (0.0)	6,064,582
	3	1,144 (0.4)	48 (0.0)	314,438	3	326 (0.0)	24 (0.0)	14,594,702
750	1	15,583,878 (91.3)	395,920 (2.3)	17,066,356	1	357,485 (4.9)	11,332 (0.2)	7,365,006
	2	2,827,898 (91.8)	72,904 (2.4)	3,079,278	2	27,662 (4.2)	1,125 (0.2)	653,402
	3	8,424,349 (92.4)	170,642 (1.9)	9,117,676	3	555,541 (4.7)	19,697 (0.2)	11,806,002
1,500	1	4,242,322 (89.4)	223,553 (4.7)	4,745,628	1	489,512 (4.1)	33,170 (0.3)	11,978,122
	2	2,128,591 (89.4)	111,979 (4.7)	2,381,744	2	330,164 (4.7)	25,037 (0.4)	7,001,470
	3	9,412,289 (88.5)	447,424 (4.2)	10,628,330	3	589,841 (4.1)	47,253 (0.3)	14,386,164
3,000	1	4,639,964 (85.4)	444,109 (8.2)	5,431,254	1	59,736 (3.5)	7,429 (0.4)	1,700,186
	2	8,963 (26.3)	801 (2.4)	34,048	2	236,784 (4.9)	28,929 (0.6)	4,876,586
	3	7,762,432 (85.0)	835,581 (9.2)	9,135,164	3	474,837 (3.5)	79,902 (0.6)	13,480,476

difficult to detect at low titers (137). Mock clinical samples were constructed containing combinations of ZIKV, CHIKV, and DENV at loads that correlate with real clinical loads from human specimens. Briefly, varying titers of ZIKV, Dengue virus 2 (DENV-2), and Chikungunya virus (CHIKV) were spiked into RNA extracted from human serum, in duplicate, to create synthetic co-infection samples. Negative control samples consisted solely of RNA extracted from human serum. Viruses were spiked-in at concentrations corresponding to Ct values from standard curves generated via RT-qPCR. The targeted spike-in values were chosen based on reports in the literature for clinical samples containing each virus to mimic a realistic co-infection scenario (29; 100; 138; 182). Given that in the literature there is at least one report of clinical samples being probed singly rather than pooled and probed (16) and it is not well known how pooling may affect virus detection levels, in this experiment we also sought to evaluate whether probing singly or within a pool would affect our ability to identify virus. Therefore, total RNA extracted from these samples was probed singly ("pool of 1") and also pooled in groups of four and 12 with singly-spiked and mock-spiked serum samples consisting of the other components of the pool. The resulting sequence reads were mapped to the reference genomes for each of these three viruses. In all cases, the three co-infecting viruses were able to be detected in each sample at relatively consistent proportions regardless of the number of samples within a pool (Figure 6A). Even DENV-2 was detectable within each sample it was spiked, despite the low concentration of viral RNA (estimated 100 genome equivalents per mL).

Although each targeted virus was represented by enough sequencing reads to be easily detected, there were differences in the depth and breadth of genome coverage observed. Whereas the full CHIKV genome was recovered from all spiked samples at a very high depth of coverage (Figure 6C), in the case of DENV-2, an average of 92.2% of the genome was recovered from co-infected samples (Figure 6D) and the average ZIKV linear genome coverage was lower, at 68.8% (Figure 6B). These patterns in coverage were similar for both replicates (Figure 7). Overall, the proportion of reads mapping to a given virus was consistent and reproducible regardless of whether a sample was probed singly or probed within a group of four or 12.

In addition to evaluating whether sensitivity and reproducibility are maintained despite multiplexing in pools of four and 12, we also sought to evaluate the probe panel's performance in the context of strain-level, and even species-level, genetic variation as well as differing concentrations of viral genetic material. Specifically, synthetic clinical samples were also constructed to contain a different strain of ZIKV than the strain the probes were designed against (strain R116265 rather than strain MR766, which is the strain whose reference genome was used for probe design; Figure 8A and 8B) and a different species of HAdV than the probes were designed to target (HAdV-51, a member of species D; as opposed to species C, B, and E, which the probes target specifically; Figure 8G and 8H). These samples were also constructed to include biological replicates and were probed singly or in pools of four or 12. In all cases, the spiked-in virus was detectable, although there was some variation in depth of coverage among multiplexed samples. As might be expected, samples that were multiplexed in sets of 12 yielded the lowest depth of coverage compared to samples that were multiplexed in sets of four or probed singly (Figure 8B, D, F). The target genomes were completely covered in the majority of on-target samples. The exception to this rich, consistent coverage included

Figure 6. Discrimination of ZIKV, CHIKV, and DENV in mock clinical samples at clinically relevant titers. Sequencing libraries made from serum samples spiked with ZIKV, DENV-2 and CHIKV were prepared in duplicate and either probed singly or within pools of four or 12. A The percentage of pathogen-specific reads detected within the synthetic co-infection samples is shown, along with the standard error of the mean for the two replicates. B-D Coverage plots demonstrating the number of reads that mapped to ZIKV, CHIKV, and DENV, respectively, as well as the distribution of those reads along the length of each genome.

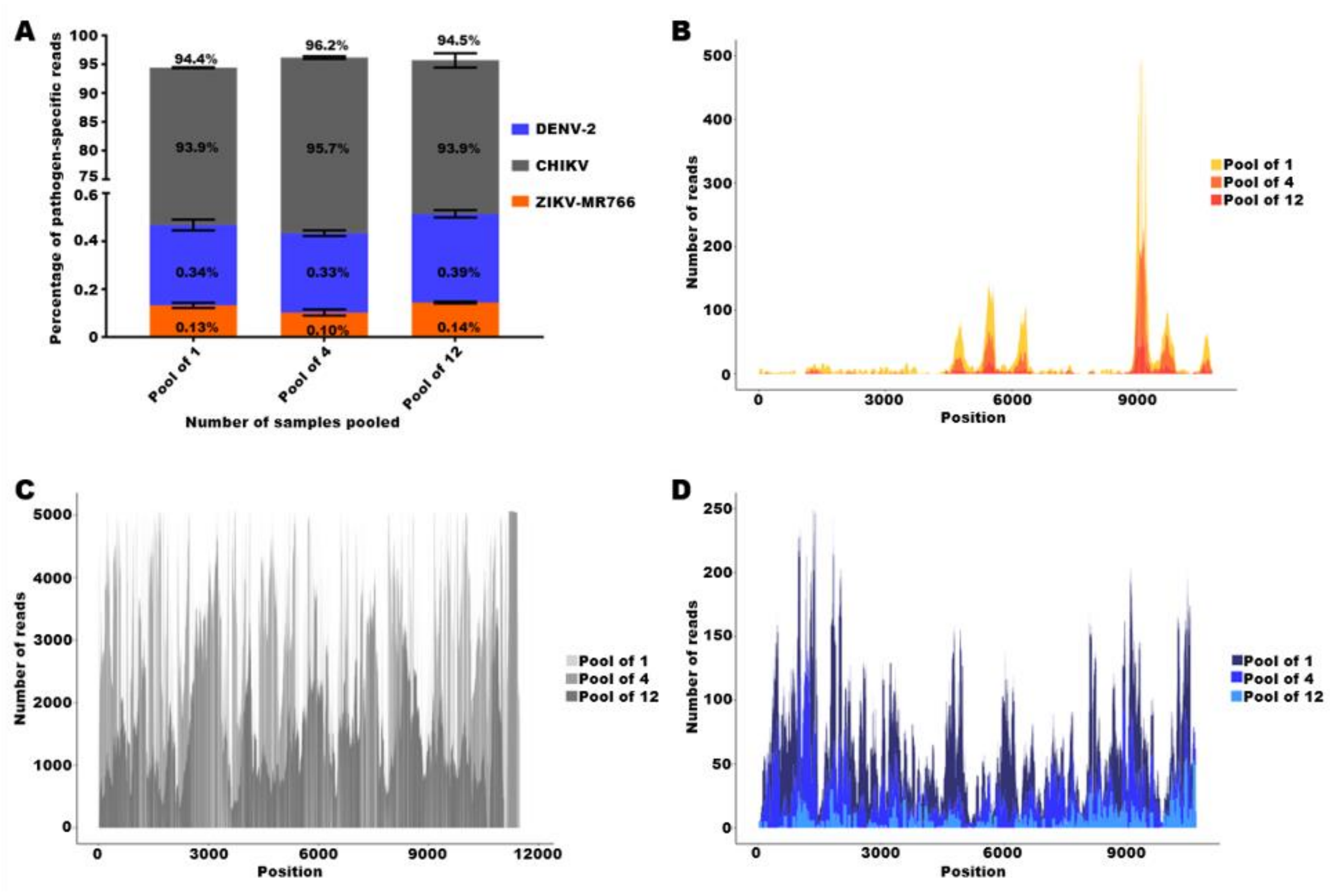
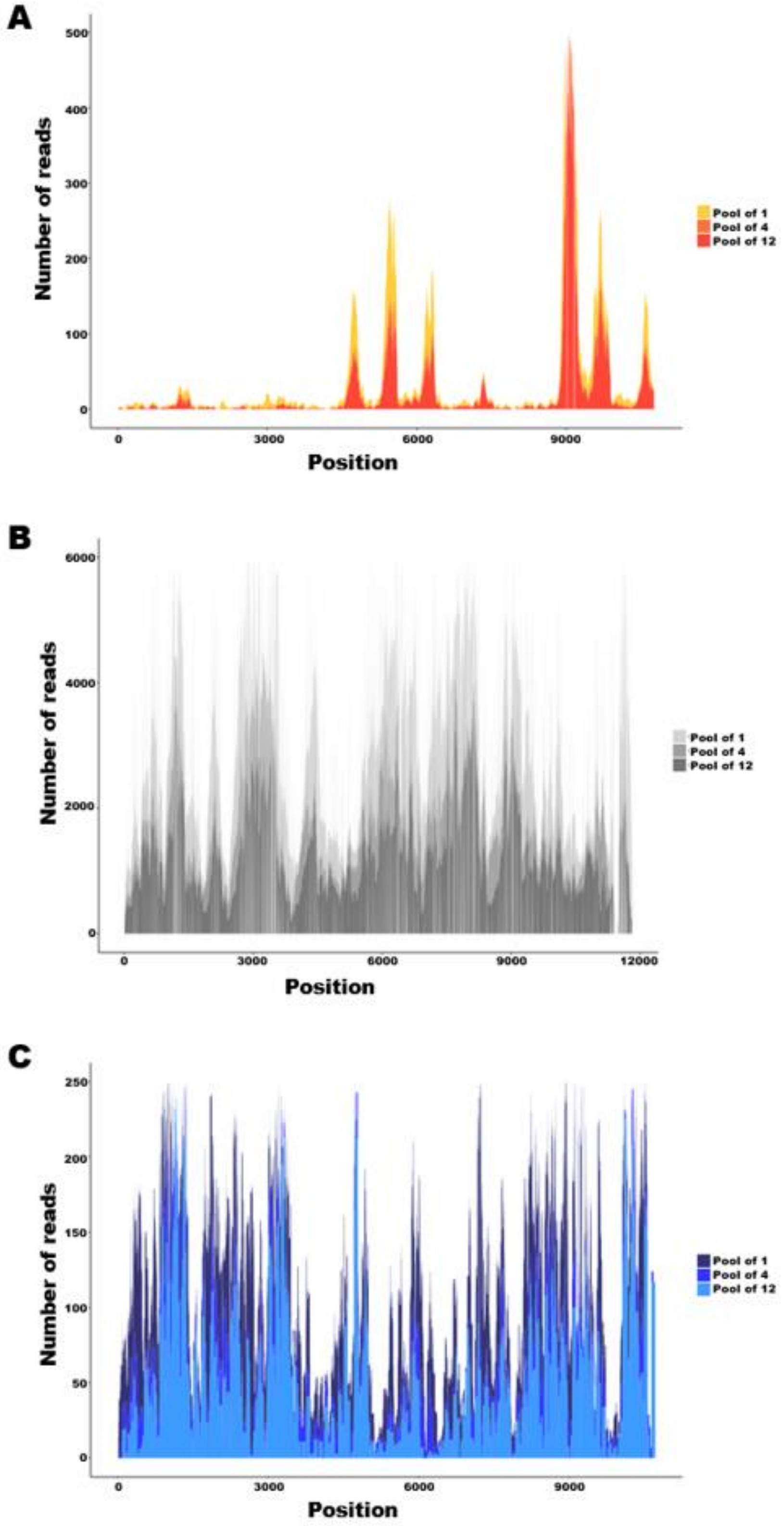

A
Percentage of pathogen-specific reads
100
95
90
85
80
75
0.6
0.4
0.2
0
94.4%
96.2%
94.5%
93.9%
95.7%
93.9%
0.34%
0.33%
0.39%
0.13%
0.10%
0.14%
DENV-2
CHIKV
ZIKV-MR766
Pool of 1
Pool of 4
Pool of 12
Number of samples pooled
B
Number of reads
500
400
300
200
100
0
0
3000
6000
9000
Position
Pool of 1
Pool of 4
Pool of 12
C
Number of reads
5000
4000
3000
2000
1000
0
0
3000
6000
9000
12000
Position
Pool of 1
Pool of 4
Pool of 12
D
Number of reads
250
200
150
100
50
0
0
3000
6000
9000
Position
Pool of 1
Pool of 4
Pool of 12

Figure 7. Replicates of ZIKV, CHIKV, and DENV in mock clinical samples at clinically relevant titers



both strains of ZIKV, which, although both were detectable, did not achieve 100% linear coverage and exhibited lower depth of coverage than the other viruses that were spiked in at similar levels (Figure 8B).

It should be noted that in this experiment, both purified RNA as well as total nucleic acid samples containing HAdV-4 and HAdV-51 genetic material were processed and sequenced (Figure 8G and 8H). The total nucleic acid samples were processed without a DNase step to allow for potential detection of both the DNA viral genome as well as viral transcripts. In the case of both HAdV-4 (species E) and HAdV-51 (species D, not targeted specifically by probes), the vast majority of the resulting sequencing reads were virus-specific and the reads were well-distributed along the length of the genome in coding regions, and in the case of the total nucleic acid samples, noncoding regions as well (Figure 8H), a phenomenon that was consistent between replicates (replicate coverage data is shown in Figure 9).

Strain-specific detection of DENV at titers below limit of detection by conventional shotgun HTS or amplicon-based sequencing.

It can be difficult to detect DENV-1 and DENV-2 in clinical samples when the Ct value crosses above 29 (30). Therefore, spiked samples were created using two serotypes of DENV with Ct values corresponding to low titer, and the samples were subjected to hybridization-based enrichment and sequencing. The resulting sequence reads were found to cover the entirety of each target genome, even for the samples corresponding to Ct value 32 (estimated 1000 genome equivalents per mL). A dose-dependent response was observed in the percentage of DENV-specific reads as the Ct value decreased (Figure 10A). For each serotype, the depth of coverage was greater than 50x even when the Ct

Figure 8. Detection of close relative viruses irrespective of extensive multiplexing. Sequencing libraries made from serum samples spiked with ZIKV, DENV, CHIKV and/or HAdV were prepared in duplicate and probed singly or probed in pools of four or 12. A, C, E, G The percentage of reads that map to each strain of spiked-in ZIKV, DENV, CHIKV, and HAdV, respectively. Each co-infected sample is denoted with an asterisk (*). Estimated genome equivalents per mL as extrapolated from RT-qPCR standard curves are listed along the top of each graph. The standard error of the mean of two replicates is shown. B, D, F, H Coverage plot for replicate one of each ZIKV-, DENV-2-, CHIKV-, and HAdV-containing sample, respectively.

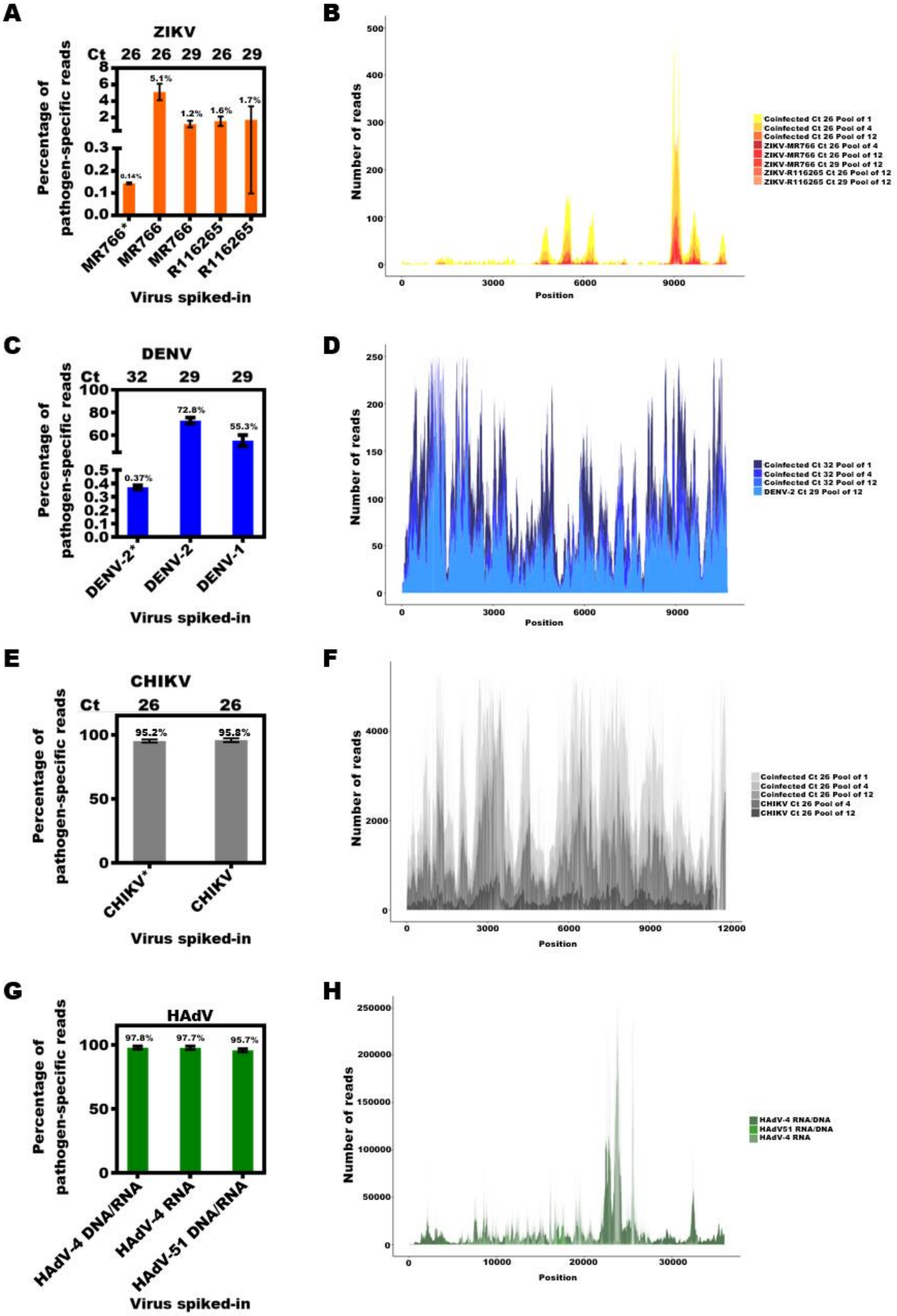

A
ZIKV
Ct 26 26 29 26 29
Percentage of pathogen-specific reads
8
6
4
2
0.3
0.2
0.1
0.0
0.14%
5.1%
1.2%
1.6%
1.7%
MR766*
MR766
MR766
R116265*
R116265
Virus spiked-in
B
Number of reads
500
400
300
200
100
0
0
3000
6000
9000
Position
Coinfected Ct 26 Pool of 1
Coinfected Ct 26 Pool of 4
Coinfected Ct 26 Pool of 12
ZIKV-MR766 Ct 26 Pool of 4
ZIKV-MR766 Ct 26 Pool of 12
ZIKV-MR766 Ct 29 Pool of 12
ZIKV-R116265 Ct 26 Pool of 12
ZIKV-R116265 Ct 29 Pool of 12
C
DENV
Ct 32 29 29
Percentage of pathogen-specific reads
100
80
60
0.5
0.4
0.3
0.2
0.1
0.0
0.37%
72.8%
55.3%
DENV-2*
DENV-2
DENV-1
Virus spiked-in
D
Number of reads
250
200
150
100
50
0
0
3000
6000
9000
Position
Coinfected Ct 32 Pool of 1
Coinfected Ct 32 Pool of 4
Coinfected Ct 32 Pool of 12
DENV-2 Ct 29 Pool of 12
E
CHIKV
Ct 26 26
Percentage of pathogen-specific reads
100
50
0
95.2%
95.8%
CHIKV*
CHIKV
Virus spiked-in
F
Number of reads
4000
2000
0
0
3000
6000
9000
12000
Position
Coinfected Ct 26 Pool of 1
Coinfected Ct 26 Pool of 4
Coinfected Ct 26 Pool of 12
CHIKV Ct 26 Pool of 4
CHIKV Ct 26 Pool of 12
G
HAdV
Percentage of pathogen-specific reads
100
50
0
97.8%
97.7%
95.7%
HAdV-4 DNA/RNA
HAdV-4 RNA
HAdV-51 DNA/RNA
Virus spiked-in
H
Number of reads
250000
200000
150000
100000
50000
0
0
10000
20000
30000
Position
HAdV-4 RNA/DNA
HAdV51 RNA/DNA
HAdV-4 RNA

Figure 9. Replicate detection of close relative viruses irrespective of extensive multiplexing

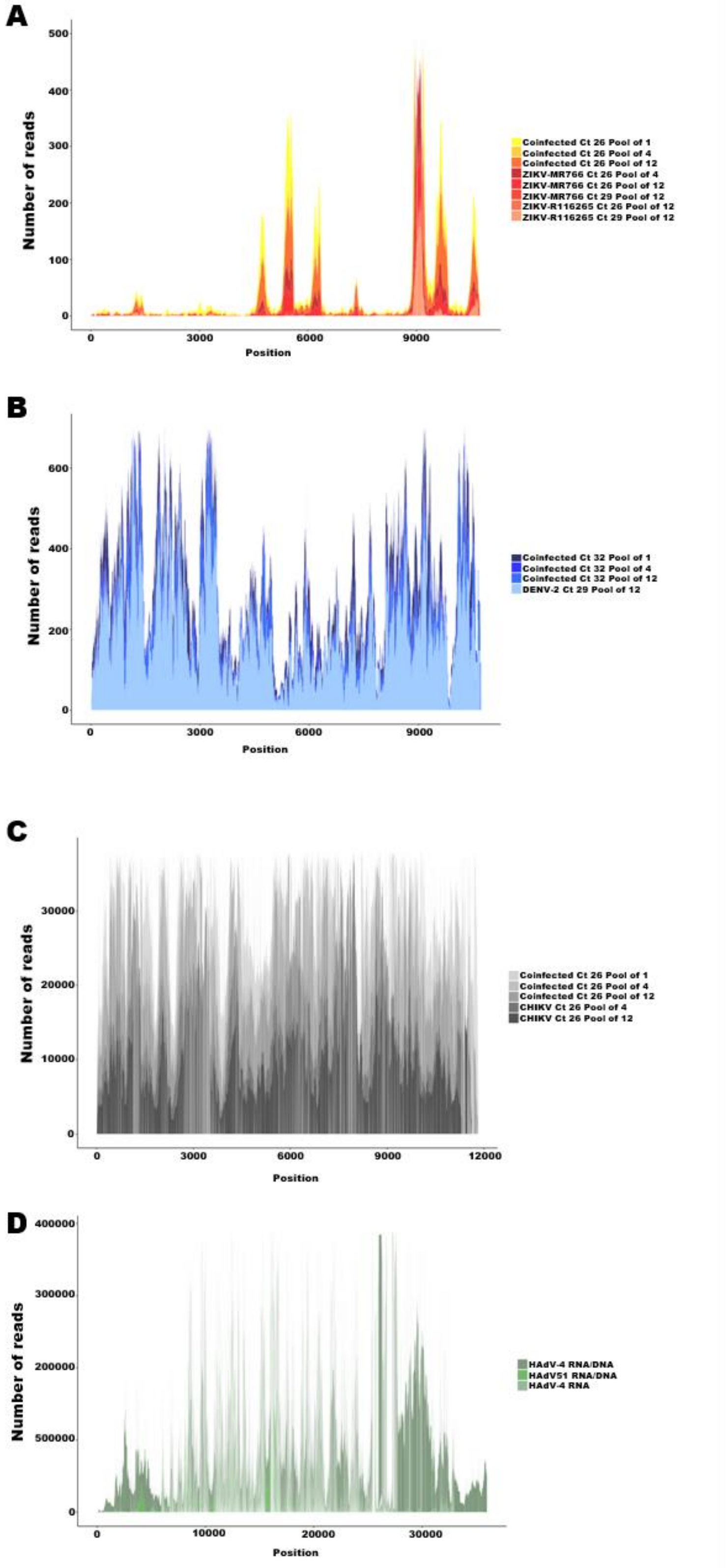
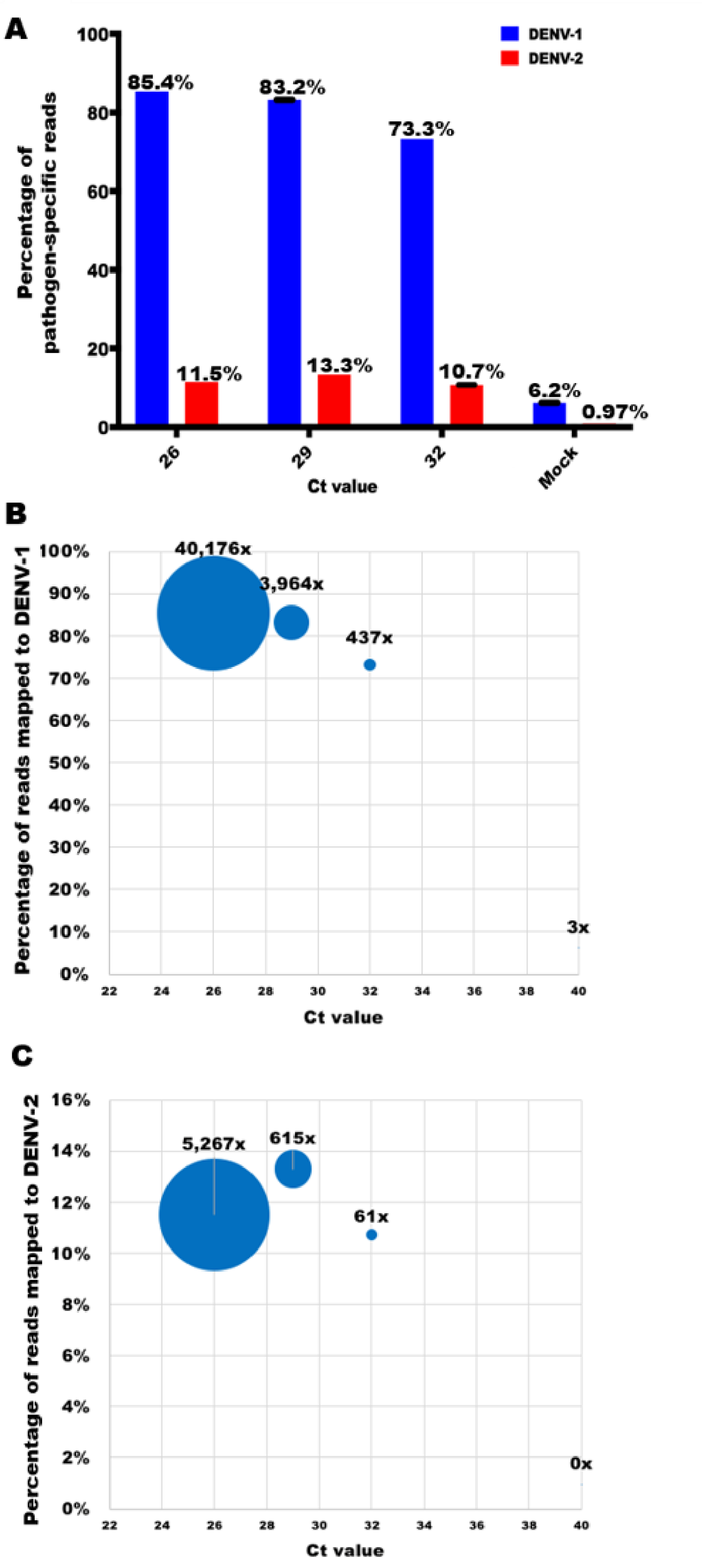


Figure 10. Recovery of full DENV genome at titers below limit of detection by conventional shotgun HTS or amplicon-based sequencing**.** DENV-1 and DENV-2 RNA were spiked into human serum RNA at a range of GE corresponding to Ct 26-32, in duplicate, and libraries were prepared using target enrichment in pools of four. Corresponding estimated genome equivalents per mL as extrapolated from a standard RT-qPCR curve are listed below the axis. Mock samples consisted of human serum RNA extract only. A The proportion of total reads that map to DENV-1 or DENV-2 at each Ct value. Error bars show standard error of two replicates. B-C The proportion of reads that map to DENV-1 or DENV-2, respectively, at each spike-in level. Bubble size corresponds to depth of coverage of the viral genome (average of two replicates).


A
DENV-1
DENV-2
Percentage of pathogen-specific reads
100
80
60
40
20
0
85.4%
11.5%
83.2%
13.3%
73.3%
10.7%
6.2%
0.97%
26
29
32
Mock
Ct value
B
Percentage of reads mapped to DENV-1
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
40,176x
3,964x
437x
3x
22
24
26
28
30
32
34
36
38
40
Ct value
C
Percentage of reads mapped to DENV-2
16%
14%
12%
10%
8%
6%
4%
2%
0%
5,267x
615x
61x
0x
22
24
26
28
30
32
34
36
38
40
Ct value

value crossed 29 (Figures 10B and 10C). As expected, the remaining reads that did not map to DENV-1 or DENV-2 were derived from human genes in spiked-in human serum RNA extract that were pulled down by the control probes, as well as the sequencing control library for PhiX.

DISCUSSION

A major challenge faced in virus detection as well as virus sequencing is the difficulty to detect divergent strains of viruses typically present at low titers amongst a robust host or environmental background. Small viral genomes present at low concentrations are effectively drowned out by signal from host nucleic acid and from commensal microorganisms. A variety of methods have been employed to increase the viral signal in high-throughput sequence data, including amplicon sequencing, but for viruses like DENV, with its genome of less than 11 kb in size, even amplicon sequencing is regarded as an inefficient approach for samples with Ct values of 29 or higher (30). Such limitations have been of particular concern for U.S. Department of Defense (DoD) laboratories tasked with biosurveillance and biodefense activities in regions with limited material resources and human expertise. Part of the motivation for this effort was to provide DoD laboratories operating in austere environments new tools aimed at enhancing on-site sequencing capacity when engaged in Force Health Protection (FHP) activities.

We demonstrate here that hybridization-based viral target enrichment yields robust coverage of small genomes from clinical samples, even yielding full-length, deeply covered genomes at concentrations whereby current amplicon sequencing protocols may be expected to fail. Moreover, we demonstrate that hybridization-based target enrichment can allow for not only detection, but also genetic characterization such as strain-level discrimination, even at very low concentrations of virus. The capability to detect and discriminate between multiple serotypes of a virus within a complex sample at clinically relevant concentrations by using this enrichment method increases the utility of high throughput sequencing for biosurveillance and for infectious disease diagnostics. For both biosurveillance and clinical sequencing, assay cost and time are important considerations. We have demonstrated that more extensive pooling and multiplexing can be performed to reduce cost and time without sacrificing the assay's ability to detect at least two strains of related virus and a variety of unrelated viruses in one sequencing reaction.

To date, the published viral target enrichment studies vary in focus and include characterization of EBOV during a recent outbreak in West Africa (16; 116) and detection of multiple viral families within clinical samples (32; 133). While the probes employed in the studies published to date vary in length from 50- to 120-mers, enrichment methods also can differ by the number of probes and target viruses included in a set. An additional potential protocol difference is the number of samples pooled, which ranges from a single sample to 12 (16; 116; 127). The current recommendation by Illumina for viral enrichment is to pool four samples (204).

The experiments described here systematically test enrichment of a single library as well as pools of four or 12 libraries and include a variety of titers of as many as three viruses within a single sample and as many as 12 samples within an enriched pool. For all conditions, even with more extensive pooling and multiplexing, we observed a dose-dependent response to varying Ct values even in co-infected clinical samples. A dose dependent response was also observed by O'Flaherty *et al.* in two co-infected samples containing Respiratory Syncytial virus and human Coronavirus OC43 spiked-in each at Ct 28 or 32 (133). Interestingly, although there was the expected dose-dependent effect on the proportion of sequencing reads derived from IFV as the concentration of spiked IFV gRNA increased, there was a slight decrease in the proportion of MERS-CoV-derived sequencing reads in samples at the upper end of the IFV gRNA concentration range. This was not expected given that MERS-CoV was present in each replicate at a constant, high level. We hypothesize that this may be due to saturation of the streptavidin beads used to capture probe-cDNA hybridized fragments. Further experimentation will be required to test that hypothesis.

Efficacy of probes varies by homology to the viral target, as evidenced by our results and those in the literature (127; 133). For example, the reference sequence used to design the probe set for DENV-1 exhibits 74% nucleotide identity over 35% of the length of the closest sequenced reference for the DENV-2 strain that was spiked. It is possible that this overlap, which is not shared by the DENV-2 probes and DENV-1 spike-in, contributed to an overrepresentation of DENV-1 reads in the experiments presented in Figures 3 and 4. Additionally, by comparison to the other richly covered strains of virus tested in multiplexed samples, there was an underwhelming coverage for both strains of ZIKV. It is possible that the quality of RNA from these viruses was less than the other RNA spike-ins, or that the probe panel for ZIKV is less efficacious when used in combination with the entirety of the probe set. The probes for ZIKV were synthesized and added later after all the other probes were combined (in response to the recent outbreak) and therefore it is possible that the comparatively lower performance of the ZIKV probes is due to a difference in quantitation of the ZIKV probe set.

We observed that HAdV-51 (species D) genetic material was efficiently enriched even though the only adenoviruses used to design the probe panel were species B, C, and E. This experiment indicates that the protocol works as well for this particular DNA virus as it does for RNA viruses. Limits of detection may vary by viral target, which may explain why previously published experiments showed differences between DNA and RNA viruses (133). Nucleotide identity between human adenovirus species D (the species to which HAdV-51 belongs) and the other human adenovirus species HAdV-A, B, C, and E was reported by Kaneko *et al.* to range from 58.73% to 69.35% (86). In our study, the probe panel containing probes for species HAdV-B, C, and E effectively enriched for the entirety of the HAdV-D genome. This suggests that when using long (80-mer) probes designed against several species of virus, related non-targeted species may also be enriched without being specifically included in the panel, if the nucleotide identity among them is at least 60-70%, and if multiple related species are targeted by the probes in the panel (in this case three species). This cross-reactivity for related human pathogens could prove to be a useful feature, by allowing for enrichment of more relevant viruses without added cost spent to increase the number of probes.

Our findings demonstrate that breadth of coverage does not suffer from extensive pooling but that deeper depth of coverage is gained by limiting the number of samples pooled. Extensive pooling makes hybridization-based enrichment sequencing more economical. Viral target enrichment could be applied as an economical approach to sequencing viruses known to mutate quickly and therefore evade other assays, fastidious organisms, or complex samples of limited volume. For example, this method could be prescribed to a scenario in which multiple serotypes of a virus such as DENV are expected to be present in a sample but detection is prohibited via conventional methods such as amplicon sequencing due to low titers. Viral target enrichment designed for a broad panel of targets could also be useful to the infectious disease field by enabling detection of low-titer viruses present in clinical samples taken from patients suffering from symptoms of unknown etiology. Another applicable use of a broad probe panel could be to perform environmental sampling. The aforementioned applications often involve complex samples of limited volume, for which this method is ideal. An important caveat to this approach is that while viral target enrichment is an economical method by which to reduce background noise in a metagenomic sample, probe design requires prior knowledge of the closest-sequenced genome for each viral target. Amplicon sequencing may be the best approach for previously known samples and unbiased whole shotgun sequencing may be more appropriate for a virus-rich sample. None of these approaches obviates the use of amplification by polymerase chain reaction or the potential introduction of sequence errors, and so standard quality analyses by computational methods should always be employed. As the development and testing of probe sets for viral target enrichment expands and continues, the application of this technique could make HTS more economical for routine use in Force Health Protection activities including biosurveillance, biodefense and outbreak investigations.

Acknowledgements and Funding

The authors would like to thank LCDR Gabriel Defang from NMRC for providing samples of MERS-CoV Vero cell culture matrix and LTC Richard Jarman for providing CHIKV RNA.

This work was funded by U.S. Navy, Office of Naval Research, In-House Laboratory Independent Research (ILIR) Program and WUN A1417. The funding bodies had no role in the design of this study nor in the collection, analysis, or interpretation of data, nor in the writing of this manuscript.

CHAPTER 3: The temporal RNA virome patterns of a lesser dawn bat (*Eonycteris spelaea*) colony revealed by deep sequencing

This work accepted for publication as: **Adrian C. Paskey, Justin H. J. Ng, Gregory K. Rice, Wan Ni Chia, Casandra W. Philipson, Randy J.H. Foo, Regina Z. Cer, Kyle A. Long, Matthew R. Lueder, Kenneth G. Frey, Theron Hamilton, Ian H. Mendenhall, Gavin J. Smith, Lin-Fa Wang, Kimberly A. Bishop-Lilly. 2020.** "The temporal RNA virome patterns of a Lesser Dawn Bat (*Eonycteris spelaea*) colony revealed by deep sequencing." *Virus Evolution.*

The work presented here is the sole work of A.C.P. with the following exceptions: J.H.J.N., W.N.C., J.H.F., I.H.M. and G.J.S. captured and maintained the bat colony and performed sample collection. G.K.R., C.W.P., R.Z.C., K.A.L. and M.R.L. performed computational analyses. K.A.B.-L. and L.-F.W. conceived the experimental design.

ABSTRACT

The virosphere is largely unexplored and the majority of viruses are yet to be represented in public sequence databases. Bats are rich reservoirs of viruses, including several zoonoses. In this study, high throughput sequencing of viral RNA extracted from swabs of four body sites per bat per timepoint is used to characterize the virome through a longitudinal study of a captive colony of fruit nectar bats, species *Eonycteris spelaea* in Singapore. Through unbiased shotgun and target-enrichment sequencing, we identify both known and previously unknown viruses of zoonotic relevance and define the population persistence and temporal patterns of viruses from families that have the capacity to jump the species barrier. To our knowledge, this is the first study that combines probe-based viral enrichment with high-throughput sequencing to create a viral

profile from multiple swab sites on individual bats and their cohort. This work demonstrates temporal patterns of the lesser dawn bat virome, including several novel viruses. Given the known risk for bat-human zoonoses, a more complete understanding of the viral dynamics in South-eastern Asian bats has significant implications for disease prevention and control. The findings of this study will be of interest to U.S. Department of Defense personnel stationed in the Asia-Pacific region and regional public health laboratories engaged in emerging infectious disease surveillance efforts.

INTRODUCTION

One Health, or the concept that humans, animals and environmental health are intrinsically linked, has provided a lens to study possible cross-species transmission of viruses from bats to humans or other amplifying hosts. The majority of viruses is previously unknown or not represented in public sequence databases, making virome characterization a particularly challenging task. Approximately 263 viruses from 25 families are known to infect humans (23; 61), but viruses of 40,000 species are estimated to infect mammals. Of those viruses, approximately 10,000 are estimated to have zoonotic potential (22). Innovative advancements in unbiased high throughput sequencing, coupled with increased computational power, have broadened the capacity for viral discovery in recent years (93). Describing and classifying previously unknown viruses and sharing them in public sequence databases not only helps the scientific community to better understand basic biology, but can ultimately improve detection and facilitate the prediction of viral emergence, and hence help prevention or mitigate future disease outbreaks (93).

More than half of all human infectious diseases result from zoonotic pathogens, and of those, 75% have emerged from wildlife reservoirs (197). Bats are the most widely distributed land animals (20), represent the second-most speciose mammalian order at 1,300 species, and harbor a significantly higher proportion of zoonotic viruses (136; 178). As the number of bats that roost in urban areas continues to increase due to anthropogenic land changes bringing bats closer into contact with livestock and humans, spillover from bats globally, especially in South-eastern Asian, has gained recognition as a potential source of pandemic infections (143). In particular, phylogenetic data suggests that bats host the progenitor strains of alpha- and beta-coronaviruses that infect humans or other incidental hosts, such as severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), porcine epidemic diarrheal virus (PEDV) and swine acute diarrhea syndrome coronavirus (SADS-CoV), with the latter two being agriculturally important (36; 110; 210). Bats are also the natural reservoir for other medically important viruses, including henipaviruses, lyssaviruses and filoviruses (10; 43; 202).

Bats are ecosystem service providers, acting as pollinators, seed dispersers, and insect consumers as well as producing guano that is used for fertilizer (95). Unfortunately, the displacement of bats through agricultural land conversion and urbanization increases the probability of human-bat interactions and increases the risk of zoonotic spillover via amplifying hosts (97; 145). The recognition of bats as a reservoir of infectious diseases, initially driven by the SARS-CoV outbreak, has led to an increase in bat-borne virus surveillance (70; 180). A comprehensive understanding of the viral dynamics in South-eastern Asian bats would have significant implications by informing pathogen surveillance, prevention and intervention.

While the majority of viral surveillance has focused on the detection of known emerging threats or their near-neighbors, there is a recent history of broadly characterizing the bat virome with high throughput sequencing (HTS). The bat virome can be defined as all of the viruses that exist within a single bat or a population of bats. Surveillance often focuses on viruses of zoonotic potential. However, a significant proportion of previously-reported viruses detected in virome studies of bats are not known to infect humans. For example, it has been reported that a high proportion of viruses detected in the bat virome reflect diet-associated viruses (13; 154). Phage-related sequences have also been explored in the literature (205) but are excluded from this work. Additionally, family-specific PCRs often focus on conserved internal genes (such as the RNA-directed-RNA polymerase) and are unable to provide information on surface proteins which are responsible for cellular entry and can be used to predict receptor usage (147).

Viral persistence and shedding in bats are, in part, driven by birth pulses, social contact, roost size, flight and migration (147). As the only truly flying mammals, bats possess a suite of characteristics that includes unique immunological factors to accommodate for the physiologically taxing nature of flight. These mechanisms may have evolved to minimize inflammation from the production of oxygen free radicals during flight, which in turn reduces damage to DNA (26; 169). Bats tolerate most viral infections without displaying symptoms due to a tighter control of immune responses that have a higher basal expression of certain defense genes (such as interferon and interferon-stimulated genes), and at the same time have multiple mechanisms to control for over-induction of inflammatory genes (165; 211). It is heretofore unknown whether bats serve as viral reservoirs, maintaining persistent infection at the population level by repeated viral infection of naïve individuals (juveniles with waning maternal antibodies) or through long-term persistence of virus within individual bats (165). In response to conflicting reports in the literature with regard to the persistence of viruses, longitudinal studies on captive colonies provide a controlled environment to fill the knowledge gap (145).

This longitudinal study of a captive bat colony presents a unique opportunity to study the viral genomes that persist or circulate within a closed community. Herein, we present a comprehensive RNA virome analysis and longitudinal evaluation of viral population persistence in a captive colony of lesser dawn bats where we characterized the virome by addressing population-level viral dynamics over time. We collected head, body, oral and rectal swabs from each bat (excluding pregnant and newborn bats) at three or four month intervals over the course of 18 months. RNA was extracted from each swab to perform shotgun and target enrichment sequencing from six time points from April 2016 to September 2017. These datasets were analyzed to ascertain the RNA virome diversity and how it changed over the study course in both individuals and cohorts. The aim of this study was to characterize the RNA virome by addressing population-level viral dynamics. Herein, we present a comprehensive RNA virome analysis and longitudinal evaluation of viral population persistence in a captive colony of lesser dawn bats.

MATERIALS AND METHODS

Bat colony structure and sampling strategy

To establish a breeding colony of lesser dawn bats, wild-caught bats were brought into captivity in April 2016. Sixteen bats resided in the colony throughout the study, two were euthanized and four were born or added to the colony after April 2016 **(Figure 1).** New bats were introduced to the colony in July 2017 and housed in separate cages situated approximately one meter apart. Head, body, oral and rectal swabs were collected from each individual bat every three or four months over the course of 18 months while the bats were kept in captivity.

Bats were housed in stainless steel mesh cages with ample room for roosting and flight. Each set of cages housed a maximum of 25 bats. The top of the cage consisted of wire meshing, allowing the bats to hang without obstruction. Burlap was strategically placed at the corner of the cage to provide seclusion. Bats were fed a liquid diet mixture of water, apple juice, glucose powder, low fat milk powder, powdered pollen, Wombaroo and slices of watermelon, mango or papaya daily at dusk. Swabs were collected quarterly for health screening purposes. Head, body, oral and rectal swabs were obtained using polyester tipped swabs and stored in 2 ml screw cap micro tubes (Sarstedt, Germany) containing 500 µl viral transport media (VTM, 10% Bovine Serum Albumin, 20% Antibiotics-Antimycotic in milli-Q water) at -80°C. Prior to head, body and rectal sample collection, swab tips were soaked in phosphate buffer saline (PBS).

Each swab site is defined as follows: oral and rectal swabs were collected by inserting a polyester tipped swab into the mouth or rectum of each bat, body swabs were collected from the fur spanning under the left wing of each bat and head swabs were collected from the fur between the ears on top of each bat's head. Internal (oral and rectal) swabs were used to characterize potential shedding of viruses and external (head and body) swabs were used in conjunction with internal swabs to characterize colony-level detection of viruses.

Nucleic acid extraction and sequencing

Total RNA was extracted from head, body, oral and rectal swabs of each bat using a QIAGEN RNeasy Kit with on-column DNase digestion (Qiagen; Valencia, CA). RNA was eluted twice with RNase-free water. RNA was extracted with the aim to sequence genomic RNA and transcripts from RNA viruses, as well as transcripts from nonencapsidated DNA viruses. A maximum of 18 µl of the extract was used as input to prepare RNA TruSeq libraries. Illumina's recommendations for the RNA TruSeq protocol were followed with a modification of fragmentation time to four minutes as described by Blackley *et al.* to account for potentially degraded RNA samples (Illumina; San Diego, CA) (16)**.** Conventional HTS (shotgun) libraries were multiplexed in pools of 24 for sequencing on the NextSeq 500 platform using v2 chemistry with 2x150 bp read lengths. Post-library enrichment probe targets and preparation methods were previously described by Paskey *et al.*; samples were probed in pools of 12 and multiplexed for sequencing on the MiSeq platform using v3 chemistry with 2x300 bp read lengths (140).

Bioinformatic analyses

Each sample was processed both by using VirusSeeker (virus discovery pipeline (209)) and MetaSPAdes assembler v3.11.1 (132). The *Eonycteris spelaea* genome was removed from each sample by read mapping to assembly GCA_003508835 (186) using bbmap v37.78 (19) prior to MetaSPAdes assembly or by VirusSeeker (209). VirusSeeker is a virus discovery pipeline that stitches paired reads together into a single read in addition to performing assembly using Newbler. Paired reads and contigs are then classified as potentially viral or discarded after being compared with viral databases using BlastX and BlastN algorithms (cutoff < e-5). False positives are then removed by comparing the candidate viral sequences to the complete nucleotide and non-redundant protein databases (cutoff < e-10). Viral reads as determined by VirusSeeker were normalized by number of reads per sample (formula below) and taxonomic assignments were filtered to exclude the possibility of sample carryover and only include assignments based on more than one read to semi-quantitatively evaluate abundance among samples.

$$Normalized\ reads = \frac{\frac{\#\ reads\ for\ one\ viral\ taxa\ in\ one\ sample}{total\ \#\ QC\ reads\ in\ that\ sample}}{virus\ genome\ length\ (nt)} \times 1000$$

Quality-controlled (QC) reads are defined as trimmed, deduplicated reads that did not map to bacterial and eukaryotic reference sequences with greater than 70% nucleotide identity using the tools described as follows. Quality control and removal of non-viral reads was performed using fastqc v0.11.5, bbmap and bbduk v37.78 (6; 19). *De novo* assembly was performed using MetaSPAdes (average contig length 198.65 bp); reads were mapped back to contigs for validation and sequenced relatives were determined by DIAMOND using the BlastX algorithm against RefSeq viral protein sequences from NCBI as of 4 December 2018 (3; 18; 132). Results were visualized using MEGAN 6 (81). Phage-related sequences are excluded from this work. Detailed analysis of contigs and reads was performed with CLC Genomics Workbench V11 (QIAGEN Bioinformatics; Redwood City, CA).

The International Committee for the Taxonomy of Viruses (ICTV) defines viral species as "a monophyletic group of viruses whose properties can be distinguished from those of other species by multiple criteria" and identity cutoffs vary by genus and are determined by natural and experimental host range, cell and tissue tropism, pathogenicity, vector specificity, antigenicity and the degree of relatedness of genomes (82). In this study, cutoffs of 90% or higher at the amino acid level and 95% or higher at the nucleotide level for the definition of a virus were used in viral classification. Sequences that were not sufficiently related to known species were classified using the following naming convention: "novel [name of viral family] virus."

Taxonomic assignments were grouped as zoonotic-related, dietary-associated or "other" based on the literature. Zoonotic-related viruses were defined as having a near-neighbor that was previously known to cause disease in vertebrates. Dietary-associated viruses were defined as plant-associated taxa that were associated with components of the bats' diet. "Other" viruses included species that could have been detected in the environment, were not previously associated with human infection or could have possibly existed as misclassified host material due to the possibility of integration (i.e., retroviruses) or similarity to mammalian genomes.

The evolutionary history of a novel filovirus-like nucleotide sequence detected in the sample was inferred by using the Maximum Likelihood method based on the General Time Reversible model (130) with 100 bootstrap replications. The tree with the highest log likelihood (-39135.14) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The analysis involved ten full-length filovirus genome sequences (NC_014373, NC_014372, NC_002549, NC_004161, NC_006432, NC_016144, KX371887, NC_024781, NC_001608, NC_03945) and one filovirus consensus sequence from Swab 340. There were a total of 3588 nucleotide positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (94).

Principal component analysis

Data for 29 zoonotic-related viruses that persisted for more than three collection dates were evaluated by principal component analysis. Evaluated elements included the number of collection dates for which the virus was detected, the total number of times the virus was detected, the length of the virus genome and binary code for the following parameters: human infectivity, cytoplasmic replication, segmented vs. non-segmented genome, vector-borne, single-stranded genome, RNA vs. DNA genome and enveloped vs. non-enveloped virion. The data was clustered using kmeans, k=4. The number of clusters was determined by using factoextra to evaluate the optimal k for the first three components by the Silhouette and Elbow methods (88). The first three principal components represented 65% of the variance and the first two principal components represented 48% of the variance. This analysis was performed using built-in statistical packages in R v3.4.1 (167).

Rarefaction curves

Rarefaction curves were produced using zoonotic-related reads as classified by VirusSeeker (209) to evaluate the extent at which mammalian viral diversity was recovered by using zoonotic-related species detected in each sample. The vegan package was utilized in R v3.4.1 for this analysis (40).

Statistical evaluation of potential confounders

To evaluate the potential for confounding influences on this analysis, we evaluated analysis of variance (ANOVA) for viral abundance. Features of concern include NextSeq 500 batch grouping, bat ID number, length of sample storage time and virus genome length. No statistically significant difference was found among groups ($p>0.1$). Taken together, these analyses suggest that taxonomic classifications of viruses were not biased by potentially confounding features of the study. We performed a two-way ANOVA to evaluate the impact of changing the collection month or number of samples sequenced on the number of taxa detected. The p-value for collection month was 0.06 and 0.018 for the number of samples sequenced per time point. There was no significant synergistic effect for collection month and number of samples sequenced ($p>0.1$). We interpret these results to indicate that a change in the number of sequencable samples would impact the number of detected viral taxa.

This analysis was performed using built-in statistical packages in R v3.4.1 (167).

RESULTS

Colony structure and sampling strategy.

To establish a breeding colony of lesser dawn bats, wild-caught bats were brought into captivity in April 2016. Head, body, oral and rectal swabs were collected from each individual bat every three or four months over the course of 18 months while the bats were kept in captivity. Total RNA was extracted from each swab and 210 swabs were successfully sequenced by shotgun sequencing. Among them, 134 swabs were also sequenced using target enrichment sequencing, resulting in 0.027% of QC target-enriched reads to be classified as viral (**Figure 11, Table 6**).

Figure 12 provides an overview of whole-colony level data, including all swab sites (head, body, oral, rectal). The most frequently detected zoonotic agent-related viruses belong to *Orthomyxoviridae, Coronaviridae, Astroviridae, Reoviridae, Picornaviridae,* and *Paramyxoviridae.* The most abundant and consistently-detected zoonotic-related viruses were unclassified betacoronaviruses, Rousettus bat coronavirus GCCDC1 and influenza A virus. Ubiquitous viruses that are native to bats but not of concern to humans such as unclassified herpesvirus, bat retroviruses and anelloviruses were classified as other or host-specific, and were frequently detected (52; 58; 139; 151). Dietary-associated viruses such as watermelon silver mottle tospovirus, an unclassified totivirus, unclassified crinivirus and potyvirus were frequently detected.

There was no significant difference in total number of viruses detected at each time point between zoonotic-related, other or dietary viruses (**Figure 13, Table 7**). Normalized abundance varied among time points. Among individual bats, there was variation in both abundance and distribution. Few complete or nearly-complete viral genomes were recovered through this study due to the complexity and microbial "noise" of metagenomic samples. The complete genome for Rousettus bat coronavirus GCCDC1 was recovered, as well as a nearly-complete genome for a novel bat mumps virus and partially complete novel bat paramyxovirus. All other zoonotic-related taxonomic assignments were made using partial gene coverage. Overall, data obtained from these 210 swabs represent massive virus diversity and broad variability in abundance among swab types and taxa. Analysis was limited to viruses determined by abundance and

Figure 11. Colony structure and sample overview. Total RNA from head, body, oral and rectal swabs was sequenced using two different, complementary methods: unbiased shotgun sequencing for broad detection of all organisms including potentially very divergent viruses, and target enrichment sequencing for more sensitive detection of 84 viruses of concern to human health and/or biosurveillance as well as their near neighbors (140). The green bar represents the original resident bat colony, the blue bar represents births and additions to the bat colony and the red bar represents 14 new bats that were introduced to the colony in July 2017. Samples from the 14 new bats are not included in this study. Solid circles represent samples that were sequenced by shotgun and/or target enrichment. Open circles represent samples that were not sequenced due to RNA input constraints.

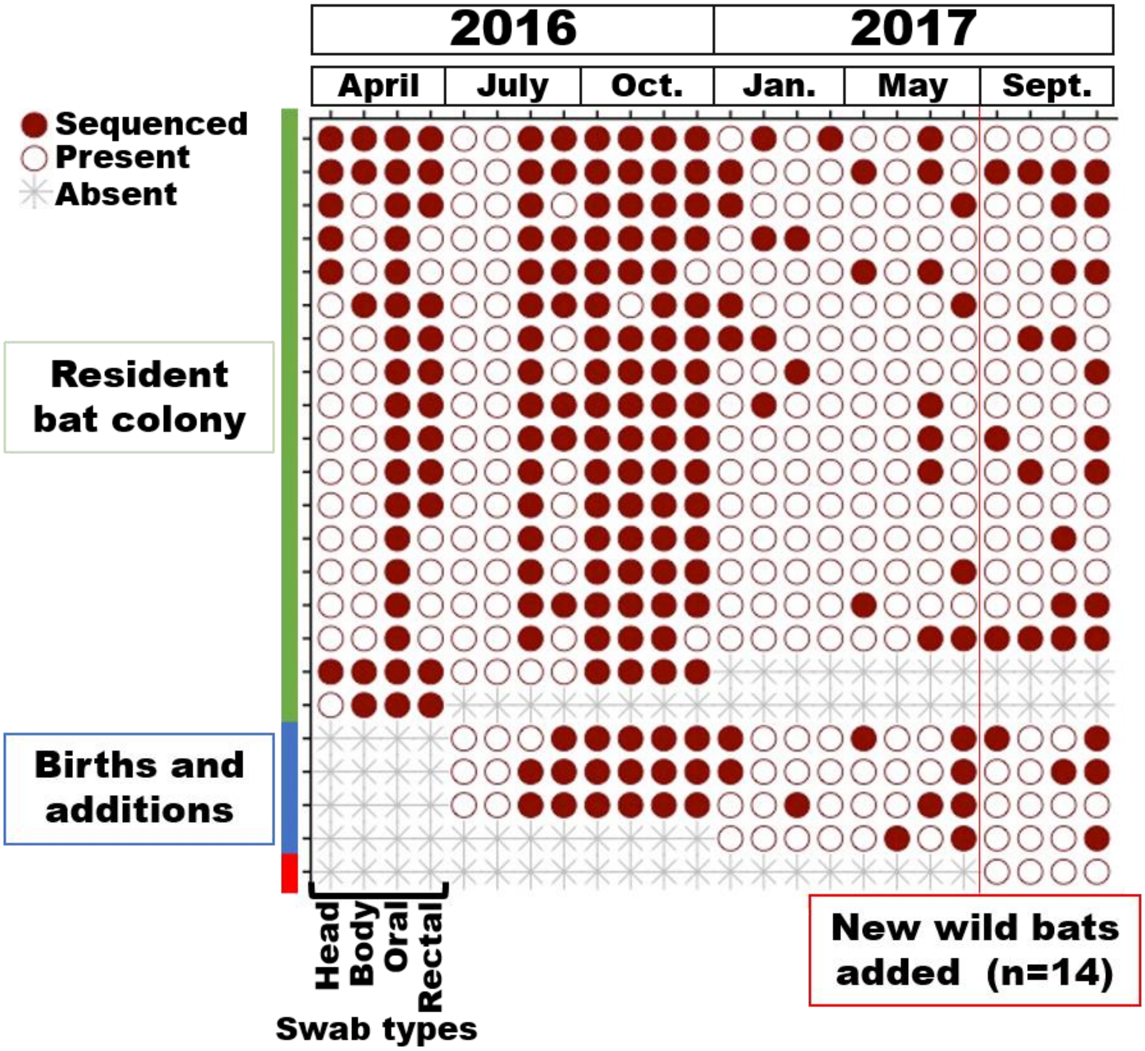
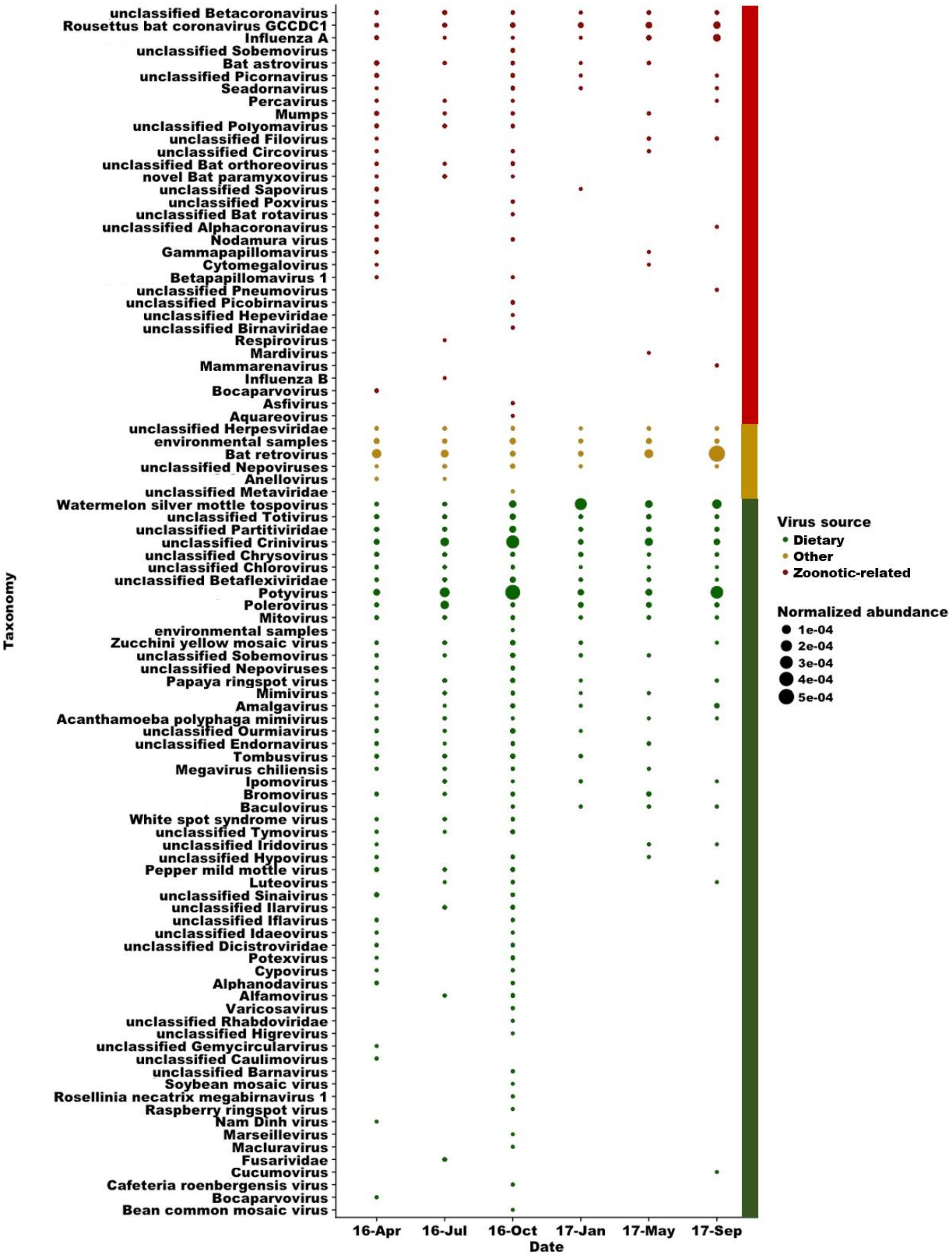

2016
2017
April
July
Oct.
Jan.
May
Sept.
● Sequenced
○ Present
✳ Absent
Resident bat colony
Births and additions
Head
Body
Oral
Rectal
Swab types
New wild bats added (n=14)

Table 6. Sequencing statistics

Date	# Swabs sequenced		# QC reads		% QC viral reads	
	Enriched	Shotgun	Enriched	Shotgun	Enriched	Shotgun
Apr-16	21	41	2,349,630	229,356,123	0.041%	0.000208%
Jul-16	27	28	2,067,041	315,310,157	0.004%	0.000261%
Oct-16	73	77	7,147,206	961,825,033	0.017%	0.000325%
Jan-17	1	14	202,706	50,295,480	0.226%	0.000560%
May-17	2	24	431,291	22,940,899	0.011%	0.000466%
Sep-17	10	26	3,102,742	129,229,830	0.046%	0.000337%
Total	134	210	15,300,616	1,708,957,522	0.027%	0.002157%

Figure 12. Classifiable members of the lesser dawn bat virome from unbiased shotgun sequencing data were arranged by virus source (red, Zoonotic-related; gold, Other; green, Dietary) and collection date. A dot representing each taxon detected is scaled to normalized abundance.


unclassified Betacoronavirus
Rousettus bat coronavirus GCCDC1
Influenza A
unclassified Sobemovirus
Bat astrovirus
unclassified Picornavirus
Seadornavirus
Percavirus
Mumps
unclassified Polyomavirus
unclassified Filovirus
unclassified Circovirus
unclassified Bat orthoreovirus
novel Bat paramyxovirus
unclassified Sapovirus
unclassified Poxvirus
unclassified Bat rotavirus
unclassified Alphacoronavirus
Nodamura virus
Gammapapillomavirus
Cytomegalovirus
Betapapillomavirus 1
unclassified Pneumovirus
unclassified Picobirnavirus
unclassified Hepeviridae
unclassified Birnaviridae
Respirovirus
Mardivirus
Mammarenavirus
Influenza B
Bocaparvovirus
Asfivirus
Aquareovirus
unclassified Herpesviridae
environmental samples
Bat retrovirus
unclassified Nepoviruses
Anellovirus
unclassified Metaviridae
Watermelon silver mottle tospovirus
unclassified Totivirus
unclassified Partitiviridae
unclassified Crinivirus
unclassified Chrysovirus
unclassified Chlorovirus
unclassified Betaflexiviridae
Potyvirus
Polerovirus
Mitovirus
environmental samples
Zucchini yellow mosaic virus
unclassified Sobemovirus
unclassified Nepoviruses
Papaya ringspot virus
Mimivirus
Amalgavirus
Acanthamoeba polyphaga mimivirus
unclassified Ourmiavirus
unclassified Endornavirus
Tombusvirus
Megavirus chiliensis
Ipomovirus
Bromovirus
Baculovirus
White spot syndrome virus
unclassified Tymovirus
unclassified Iridovirus
unclassified Hypovirus
Pepper mild mottle virus
Luteovirus
unclassified Sinaivirus
unclassified Ilarvirus
unclassified Iflavirus
unclassified Idaeovirus
unclassified Dicistroviridae
Potexvirus
Cypovirus
Alphanodavirus
Alfamovirus
Varicosavirus
unclassified Rhabdoviridae
unclassified Higrevirus
unclassified Gemycircularvirus
unclassified Caulimovirus
unclassified Barnavirus
Soybean mosaic virus
Rosellinia necatrix megabirnavirus 1
Raspberry ringspot virus
Nam Dinh virus
Marseillevirus
Macluravirus
Fusarividae
Cucumovirus
Cafeteria roenbergensis virus
Bocaparvovirus
Bean common mosaic virus
Taxonomy
16-Apr
16-Jul
16-Oct
17-Jan
17-May
17-Sep
Date
Virus source
● Dietary
● Other
● Zoonotic-related
Normalized abundance
● 1e-04
● 2e-04
● 3e-04
● 4e-04
● 5e-04

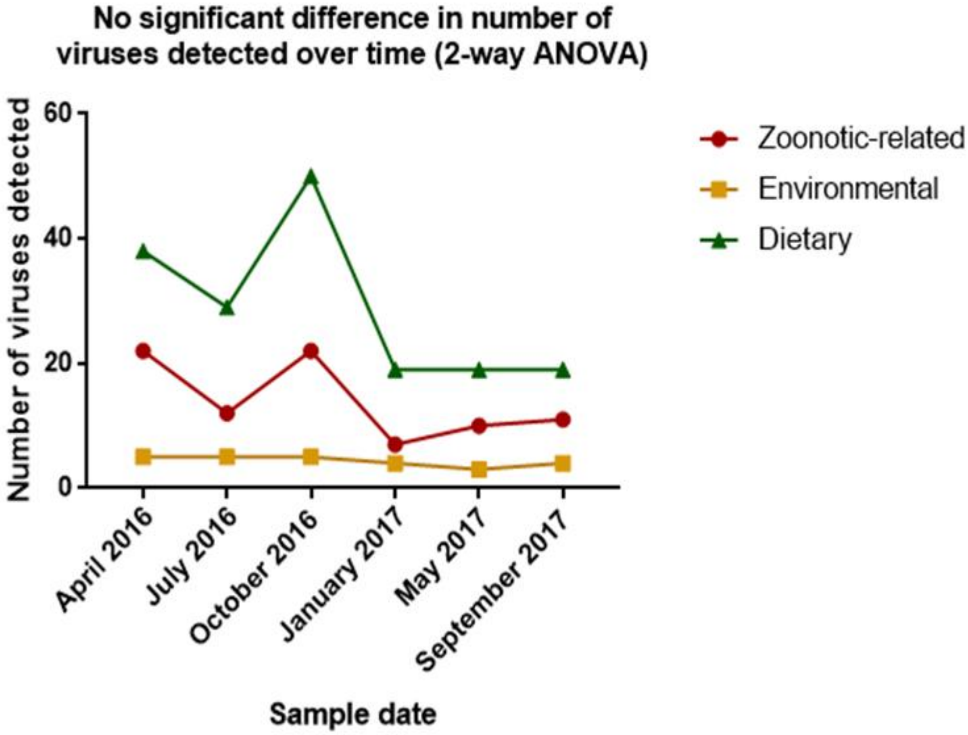


Figure 13. There was no significant difference in number of viruses detected over time (2-way ANOVA).

Table 7. The number of viruses detected using VirusSeeker at each time point in unbiased shotgun sequencing data are listed, parsed by date and virus source.

Virus source	**N = # viruses detected at each time point**					
	Apr-16	**July-16**	**Oct-16**	**Jan-17**	**May-17**	**Sept-17**
Zoonotic-related	22	12	22	7	10	11
Other	5	5	5	4	3	4
Dietary	38	29	50	19	19	19
Total	**63**	**46**	**74**	**30**	**32**	**34**

significance with a known sequenced relative, therefore excluding unmatched reads.

Comparison of two different sequencing strategies.

Target enrichment for known viruses of biosurveillance concern resulted in a simplified dataset consisting of primarily zoonotic-related taxonomic classifications. Data from target enrichment sequencing was used to guide bioinformatic analysis of shotgun sequencing data.

The number of zoonotic-related taxonomic assignments was compared between matched swabs sequenced both by shotgun and by target enrichment sequencing. Of 317 taxa assigned by VirusSeeker to shotgun data, 33.1% of assignments were zoonotic-related. Conversely, 72.1% of 44 target-enriched assignments were zoonotic-related, a result which is expected as the probe panel used here for target enrichment intentionally biases our dataset toward known viruses of biosurveillance concern (**Figure 14**). It is important to note that the number of taxonomic assignments was limited to previously known, classifiable references and cannot account for unknown sequences. For reference, approximately 5.3 million base pairs (Mbp) of sequence were unequivocally classified as viral. 293 giga base pairs (Gbp) of sequence were identified as potential viral sequence, including unclassifiable viral reads or “dark matter” that did not match any sequence in GenBank (92).

Previously we demonstrated binding tolerance as low as 79% nucleotide identity in a contrived, complex sample containing human adenovirus strains 4 and 51, as well as 89% nucleotide identity for Zika virus strains MR766 and R11265 in contrived samples (140). Therefore, the capacity for target enrichment probes to bind slightly off-target but related (near-neighbor) sequences in the current samples was evaluated. In a subset of

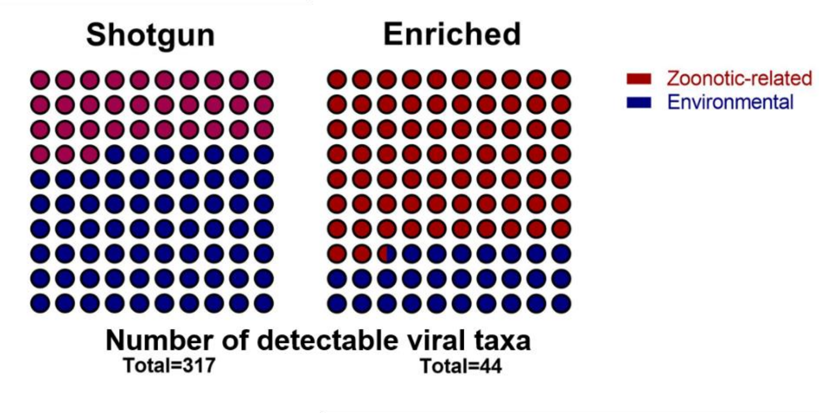


Figure 14. The proportion of zoonotic-related taxonomic assignments made for 134 matched sample sets (unbiased shotgun sequencing data matched with target enriched sequencing data derived from same swab) is illustrated in maroon. VirusSeeker made 317 taxonomic assignments using conventional HTS (shotgun) data and 44 taxonomic assignments using target-enriched data.

enriched reads classified by VirusSeeker as belonging to *Filoviridae, Orthomyxoviridae, Paramyxoviridae, Picornaviridae* and *Pneumoviridae*, the lowest percent identity at the nucleotide level between probe and captured sequence was 88%. Taken together, these data demonstrate that target enrichment probes exhibit a binding tolerance that could permit the enrichment from environmental samples of previously unknown near-neighbors that have yet to be represented in public sequence databases.

Evaluation of a dynamic virome in a captive colony of bats.

We considered whether trends in detected viruses could be observed despite the condition of captivity, where drivers of infection dynamics such as contact rates, food abundance and quality, environmental perturbations, reproduction and population density were partially controlled or minimized (147). While numerous zoonotic-related viruses such as bat orthoreovirus, mumps-like bat paramyxovirus, unclassified polyomavirus, unclassified sapovirus, and unclassified bat rotavirus did not persist over time, several viruses were detectible at five or six out of six collection dates (**Figure 15**). Persistence within the colony could only be measured to the fifth collection date because new, wild-caught bats were introduced to the colony in July 2017, prior to the final swab collection date. Thus, new viruses could have been introduced or reintroduced to the original colony. Notable zoonotic-related viruses that were detectable at multiple collection dates belonged to the families *Astroviridae, Orthomyxoviridae, Picornaviridae, Paramyxoviridae, Reoviridae* and *Coronaviridae*. Other and dietary-associated viruses were detected more consistently, as expected, due to frequent reintroduction from food and environmental sources, with the exception of notable diet changes in the first year of the study.

Figure 15. Selected zoonotic-related viruses that were detected at two or more time points in unbiased shotgun sequencing data are graphed with swab collection date on the x-axis and the normalized abundance on the y-axis. The gray line displays the average normalized abundance and the positive and negative standard deviation about the mean is represented by a dotted line. Normalized abundance for individual swabs, with type coded by shape, is represented by points. The size of the point corresponds to the normalized abundance, which is scaled to be consistent across all line graphs despite the free-y-axis for each individual line graph. There are few significant differences in normalized viral abundance over time for each virus.

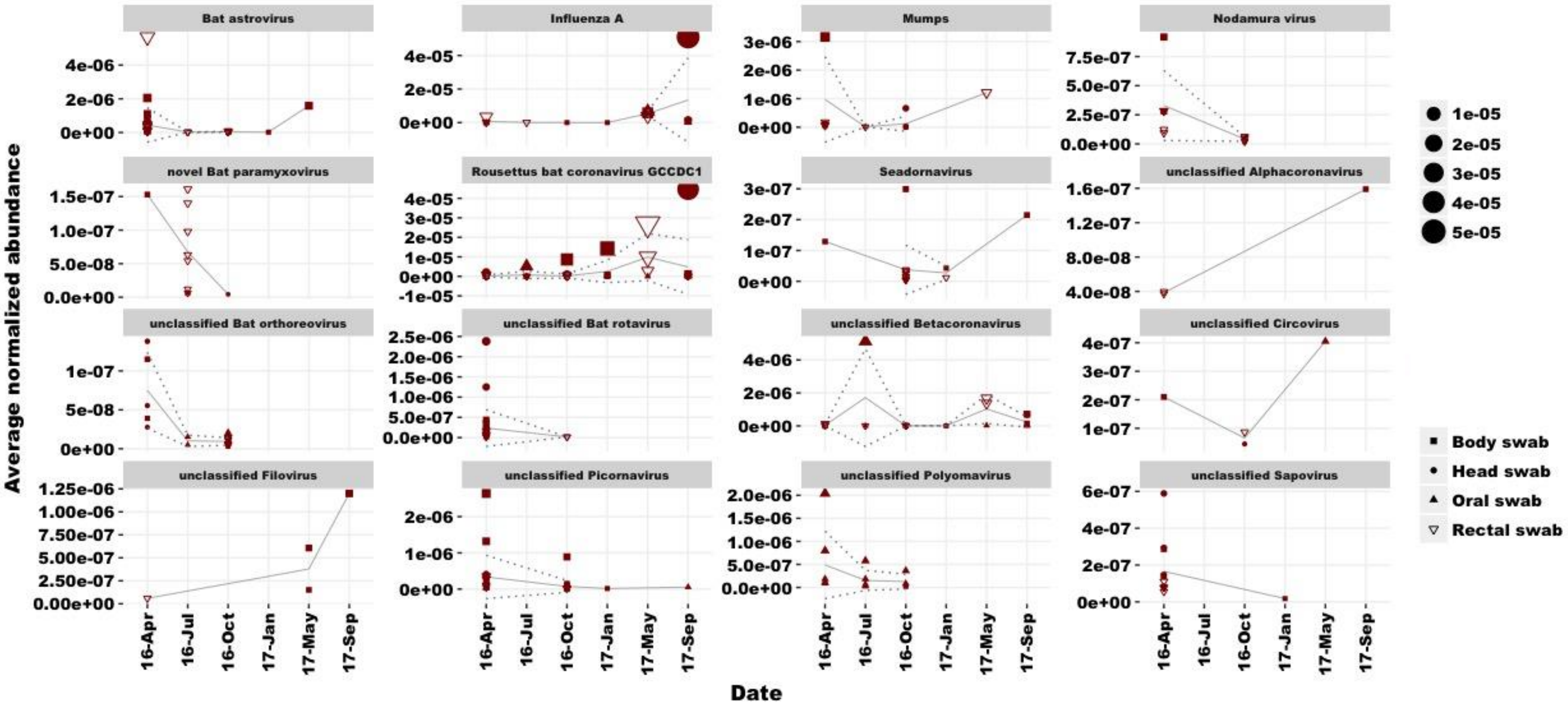

Bat astrovirus
4e-06
2e-06
0e+00
Influenza A
4e-05
2e-05
0e+00
Mumps
3e-06
2e-06
1e-06
0e+00
Nodamura virus
7.5e-07
5.0e-07
2.5e-07
0.0e+00
novel Bat paramyxovirus
1.5e-07
1.0e-07
5.0e-08
0.0e+00
Rousettus bat coronavirus GCCDC1
4e-05
3e-05
2e-05
1e-05
0e+00
-1e-05
Seadornavirus
3e-07
2e-07
1e-07
0e+00
unclassified Alphacoronavirus
1.6e-07
1.2e-07
8.0e-08
4.0e-08
unclassified Bat orthoreovirus
1e-07
5e-08
0e+00
unclassified Bat rotavirus
2.5e-06
2.0e-06
1.5e-06
1.0e-06
5.0e-07
0.0e+00
unclassified Betacoronavirus
4e-06
2e-06
0e+00
unclassified Circovirus
4e-07
3e-07
2e-07
1e-07
unclassified Filovirus
1.25e-06
1.00e-06
7.50e-07
5.00e-07
2.50e-07
0.00e+00
unclassified Picornavirus
2e-06
1e-06
0e+00
unclassified Polyomavirus
2.0e-06
1.5e-06
1.0e-06
5.0e-07
0.0e+00
unclassified Sapovirus
6e-07
4e-07
2e-07
0e+00
16-Apr
16-Jul
16-Oct
17-Jan
17-May
17-Sep
Average normalized abundance
Date
● 1e-05
● 2e-05
● 3e-05
● 4e-05
● 5e-05
■ Body swab
● Head swab
▲ Oral swab
▽ Rectal swab

Table 8 summarizes biosurveillance-related viruses detected in this study and reported in the literature, listing viral family, sampling country, references, virus, and total reads if the result is from this study (25).

Variation of viral abundance at both the colony and individual level.

Data from oral and rectal swabs were used as an indicator of potential replication in and shedding from particular bats, while data from head and body swabs (presumably derived from contact with other, shedding, bats but not necessarily indicating shedding from the particular bat on whose head or body it is identified) were used to characterize the overall diversity of viruses present within the colony. Bat coronavirus was previously found in the small intestine of experimentally infected *Leschenault's rousettes (Rousettus leschenaultii* (183)); therefore, we hypothesized that coronaviruses would be frequently detected from rectal swabs. In fact, rectal swabs contained the greatest proportion of all coronaviruses detected (30.3%), followed by body swabs (28.3%), head swabs (23.7%) and oral swabs (17.8%). We hypothesize that the generally high abundance of viruses on the skin of bat bodies is due to the gregariousness of lesser dawn bats.

Fractional abundance was used to evaluate detection of viruses on the colony level and is graphed in **Figure 16**. The zoonotic agent-related viruses detected within the colony are consistent with the literature (**Figure 16A**) and many dietary-associated viruses have been associated with elements of the colony's diet (**Figure 16B**) (122-124; 205). Results are biased toward RNA viruses due to our technical approach and it is likely that DNA virus diversity is not fully represented in these analyses. This dataset does, however, represent RNA from multiple related viruses that were detected at various time points throughout the study. **Figure 17** displays the sporadic detection of even the most

Table 8. Overview of biosurveillance-relevant virus detection as compared to the literature

Viral family	Sampling country	References	Virus	Total Reads (this study)
Adenoviridae	Singapore	Viruses 2019, 11(3):pii: E250	*Adenoviridae* sp. DNUS-Bat-Polymerase-Esp	NA
Arenaviridae	Singapore	This study	Mammarenavirus	2
Asfarviridae	Singapore	This study	Asfivirus	6
Astroviridae	Singapore	This study	Bat astrovirus	294
Astroviridae	Singapore	One Health 2017, 4:27-33	Mamastrovirus sp.	NA
Astroviridae	Laos	Infect Genet Evol 2016, 47:41-50	Mamastrovirus sp. PREDICT MAstV-13 strains	NA
Caliciviridae	Singapore	This study	unclassified sapovirus	42
Circoviridae	Singapore	This study	unclassified circovirus	8
Coronaviridae	Singapore	This study	Rousettus bat coronavirus GCCDC1	2765
Coronaviridae	Singapore	This study	unclassified Betacoronavirus	84
Coronaviridae	Singapore	This study	unclassified Alphacoronavirus	6
Coronaviridae	Singapore	Transbound Emerg Dis 2017, 64(6):1790-1800	Bat betacoronavirus Es/Singapore/2014	NA
Coronaviridae	Cambodia	Infect Genet Evol 2016, 48:10-18	Bat coronavirus RK strains	NA
Coronaviridae	China	Virol Sin 2018, 33(1):87-95	Bat coronavirus GCCDC1 BatCoV *Eonycteris spelaea*/Mengla/2016	NA
Coronaviridae	Laos	Infect Genet Evol 2016, 48:10-18	Coronavirus PREDICT CoV-22 PREDICT_CoV-22/LAP11-D0063	NA
Coronaviridae	Cambodia	Infect Genet Evol 2016, 48:10-18	Rousettus bat coronavirus HKU9 strains	NA
Coronaviridae	Laos	Infect Genet Evol 2016, 48:10-18	Rousettus bat coronavirus HKU9 PREDICT-LAP strains	NA
Filoviridae	Singapore	This study	unclassified filovirus	22

Filoviridae	China	Emerg Infect Dis 2017, 23(3):482-486	Bat filovirus *Eonycteris spelaea*/China/2009, 2015 strains	NA
Flaviviridae	Singapore	Viruses 2019, 11(3):pii: E250	*Flaviviridae* sp. DNUS-Bat-E-Esp	NA
Hepeviridae	Singapore	This study	unclassified *Hepeviridae*	2
Herpesviridae	Singapore	This study	unclassified *Herpesviridae*	609
Herpesviridae	Singapore	This study	Percavirus	30
Herpesviridae	Singapore	This study	Cytomegalovirus	4
Herpesviridae	Singapore	This study	Mardivirus	2
Herpesviridae	Malaysia	Unpublished	unidentified herpesvirus acc_AB125970	NA
Nodaviridae	Singapore	This study	Nodamura virus	43
Orthomyxoviridae	Singapore	This study	Influenza A virus	36
Orthomyxoviridae	Singapore	This study	Influenza B virus	6
Papillomaviridae	Singapore	This study	Betapapillomavirus 1	4
Papillomaviridae	Singapore	This study	Gammapapillomavirus	6
Papillomaviridae	Singapore	Viruses 2019, 11(3):pii: E250	*Papillomaviridae* sp. DNUS-Bat-E1-Esp	NA
Paramyxoviridae	Singapore	This study	Mumps	248
Paramyxoviridae	Singapore	This study	novel bat paramyxovirus	112
Paramyxoviridae	Singapore	This study	Respirovirus	2
Paramyxoviridae	China	Viruses 2014, 6(5):2138-54	Henipavirus YN12069/CHN/2012	NA
Paramyxoviridae	Singapore	Viruses 2019, 11(3):pii: E250	*Paramyxoviridae* s. DNUS-Bat-L-9, N	NA
Parvoviridae	Singapore	This study	Bocaparvovirus	16
Parvoviridae	Singapore	Viruses 2019, 11(3):pii: E250	*Parvoviridae* sp. DNUS-Bat-VP1, VP2	NA
Picornaviridae	Singapore	This study	unclassified Picornavirus	222
Picornaviridae	Singapore	Viruses 2019, 11(3):pii: E250	*Picornaviridae* sp. DNUS-Bat-3D-Esp, Polyprotein-Esp	NA
Pneumoviridae	Singapore	This study	unclassified pneumovirus	2
Polyomaviridae	Singapore	This study	unclassified polyomavirus	166
Polyomaviridae	Singapore	Viruses 2019, 11(3):pii: E250	*Polyomaviridae* sp. DNUS-Bat-V7-Esp	NA
Reoviridae	Singapore	This study	Seadornavirus	40
Reoviridae	Singapore	This study	unclassified bat rotavirus	333
Reoviridae	Singapore	This study	unclassified bat orthoreovirus	112

Reoviridae	Singapore	Viruses 2019, 11(3):pii: E250	Orthoreovirus sp. DNUS-Bat-L2, M2-Esp	NA
Reoviridae	Philippines	Arch Virol 2017, 162(6):1529-1539	Pteropine orthoreovirus Samal-24	NA
Reoviridae	Philippines	Arch Virol 2017, 162(6):1529-1539	Pteropine orthoreovirus Talikud strains 73, 74, 81, 83	NA
Reoviridae	Singapore	Viruses 2019, 11(3):pii: E250	Rotavirus sp. DNUS-Bat-Vp1, Vp7-Esp	NA
Retroviridae	Singapore	This study	Bat retrovirus	1807
Retroviridae	China	Viruses 2014, 6(5):2138-54	Bat gammaretrovirus comp48905_c0	NA

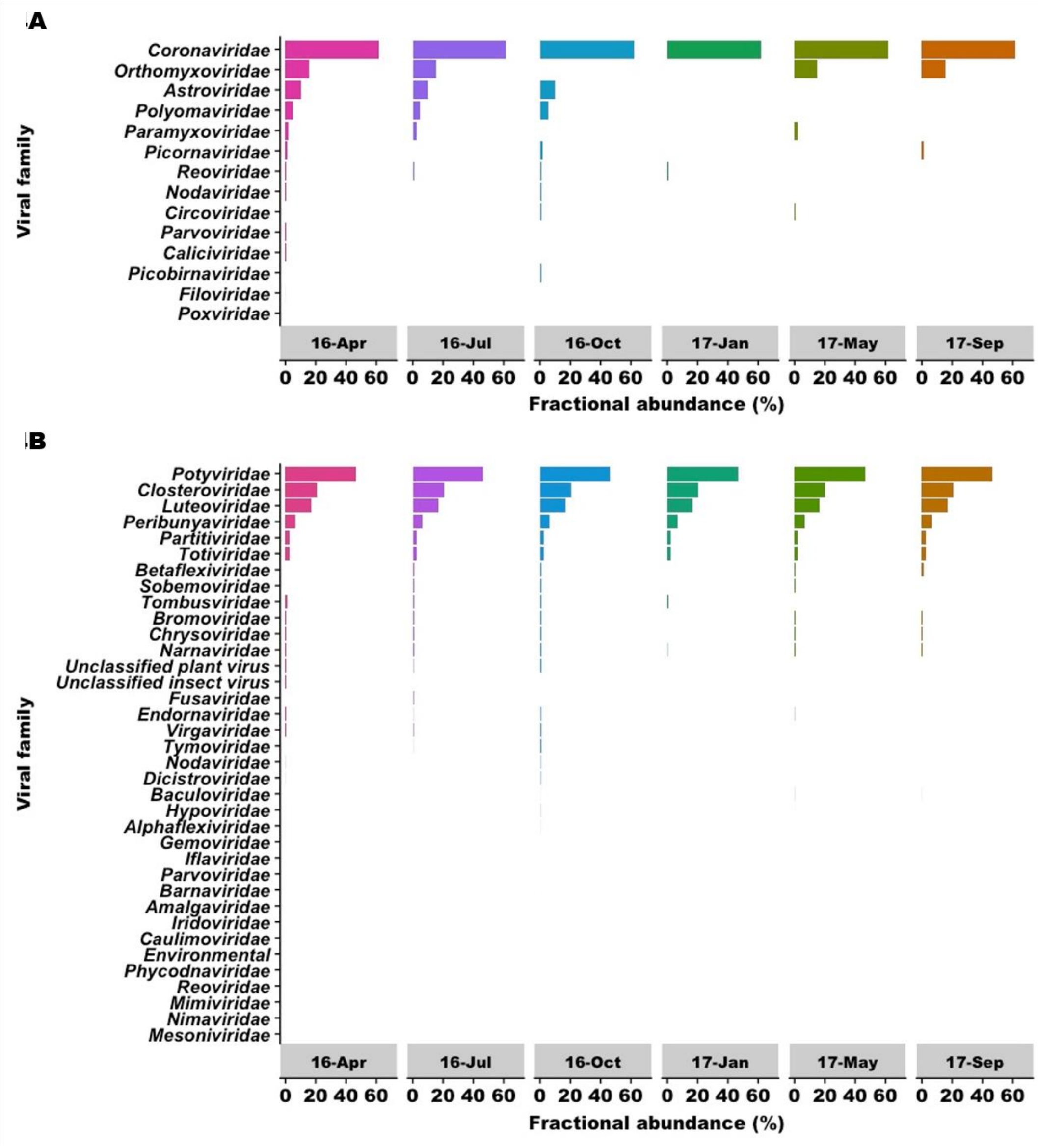


Figure 16. The proportion of reads from unbiased shotgun RNA sequencing data representing viral families in each virus source are graphed. 4A includes zoonotic agent-related viruses and 4B includes dietary-associated viruses.

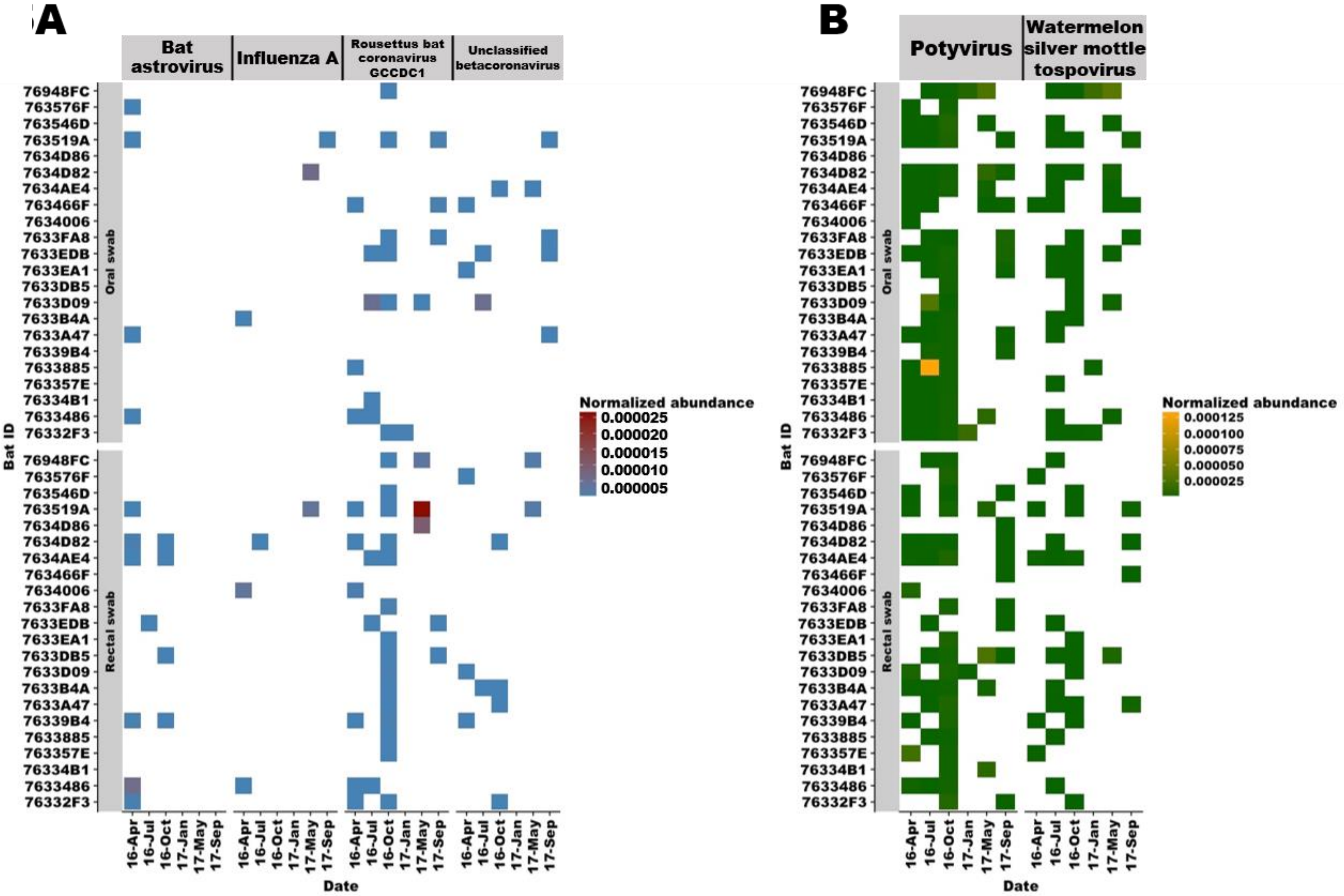
frequently detected zoonotic-related viruses. Notably, this inconsistency is in contrast to the more consistent detection of other and dietary viruses over time. One should note that bat 7634D86 was born into the colony and first sampled in 2017. Additionally, bat 763576F was euthanized between the October 2016 and January 2017 time points.

Persistence of certain viral populations at the colony level.

Figure 17 highlights four zoonotic-related and two dietary viruses, which were "more persistent" than other viruses detected. Individual bats rarely shed specific viruses during consecutive time points. Overall, viruses were not restricted to individual bats, but rather were detected in multiple bats at numerous time points. Dietary virus shedding was more consistent and detected at the overall highest normalized abundance. We observed that bat astrovirus was not detected beyond the October 2016 time point. Influenza A virus was detected sporadically, and never in the same shedder swab type from the same bat at subsequent time points. This pattern could reflect intermittent shedding or cycles of infection, transmission and reinfection or persistent virus replication at levels near or slightly below the limit of detection for our assay. In this study, we observed read support in 24 swabs for an influenza A-like virus by both shotgun and enrichment sequencing. No full sequences for all eight segments were detected, preventing unequivocal strain typing.

Rousettus bat coronavirus GCCDC1 and unclassified Betacoronavirus were detected most frequently of all zoonotic-related viruses and display remarkable population persistence (**Figures 15 and 17**). Although coronaviruses are commonly detected in South-eastern Asian bats, to our knowledge, this is the first time the cross-family recombinant Rousettus bat coronavirus GCCDC1 has been detected outside of China (80; 135). Overall, we observed that each of these viruses was

Figure 17. Abundance of select taxa detected from unbiased shotgun sequencing data derived from shedding sites is displayed in a heat map. Individual bat identification numbers are listed along the y-axis, parsed by oral and rectal swab types. Sample collection date is listed along the x-axis, divided by specific viral taxa and virus source. Figure 5A includes zoonotic-related viruses bat astrovirus, influenza A, Rousettus bat coronavirus GCCDC1 and unclassified betacoronavirus. Figure 5B includes dietary viruses potyvirus and watermelon silver mottle tospovirus. Reads were normalized to virus size and total number of reads in each sample.


A
Bat astrovirus
Influenza A
Rousettus bat coronavirus GCCDC1
Unclassified betacoronavirus
Oral swab
Rectal swab
Bat ID
76948FC
763576F
763546D
763519A
7634D86
7634D82
7634AE4
763466F
7634006
7633FA8
7633EDB
7633EA1
7633DB5
7633D09
7633B4A
7633A47
76339B4
7633885
763357E
76334B1
7633486
76332F3
16-Apr
16-Jul
16-Oct
17-Jan
17-May
17-Sep
Date
Normalized abundance
0.000025
0.000020
0.000015
0.000010
0.000005
B
Potyvirus
Watermelon silver mottle tospovirus
Normalized abundance
0.000125
0.000100
0.000075
0.000050
0.000025

detected in multiple bats and subsequently detected in different bats at later time points.

To investigate the inherent qualities that contribute to population persistence, 29 zoonotic-related viruses that persisted at the population-level for more than three collection dates were evaluated by principal component analysis of multivariate characteristics (human infectivity, cytoplasmic replication, segmented vs. non-segmented genome, vector-borne, single-stranded genome, RNA vs. DNA genome and enveloped vs. non-enveloped virion).

We hypothesized that there is an association between the frequency of detection of a virus and known capacity to infect humans (coded 0 if not known to infect humans or 1 if known to infect humans as defined by EcoHealth Alliance), therefore indicating a propensity for spillover (136). In fact, cluster one had a higher known average capacity to infect humans (0.92) and also included the most frequently detected viruses as compared to clusters two through four (0.5, 0.44, 0.2, respectively). The hypothesis was supported by unsupervised clustering and we further analyzed the clustering results to make several observations (**Figure 18**).

Cluster one (13 members; gold points) was composed of single-stranded RNA viruses. Significantly more of these viruses are known to infect humans, and none of these viruses is vector-borne. Cluster two (2 members; red points) was composed of the enveloped, double-stranded DNA viruses: asfivirus and unclassified bat poxvirus. This group had the largest genomes. Cluster three (9 members; grey points) is represented by segmented RNA viruses. Cluster four (5 members; green points) was composed of nonenveloped DNA viruses. Overall, the viral members of the largest cluster, Cluster one, also included the largest number of human-infecting viruses that are associated with

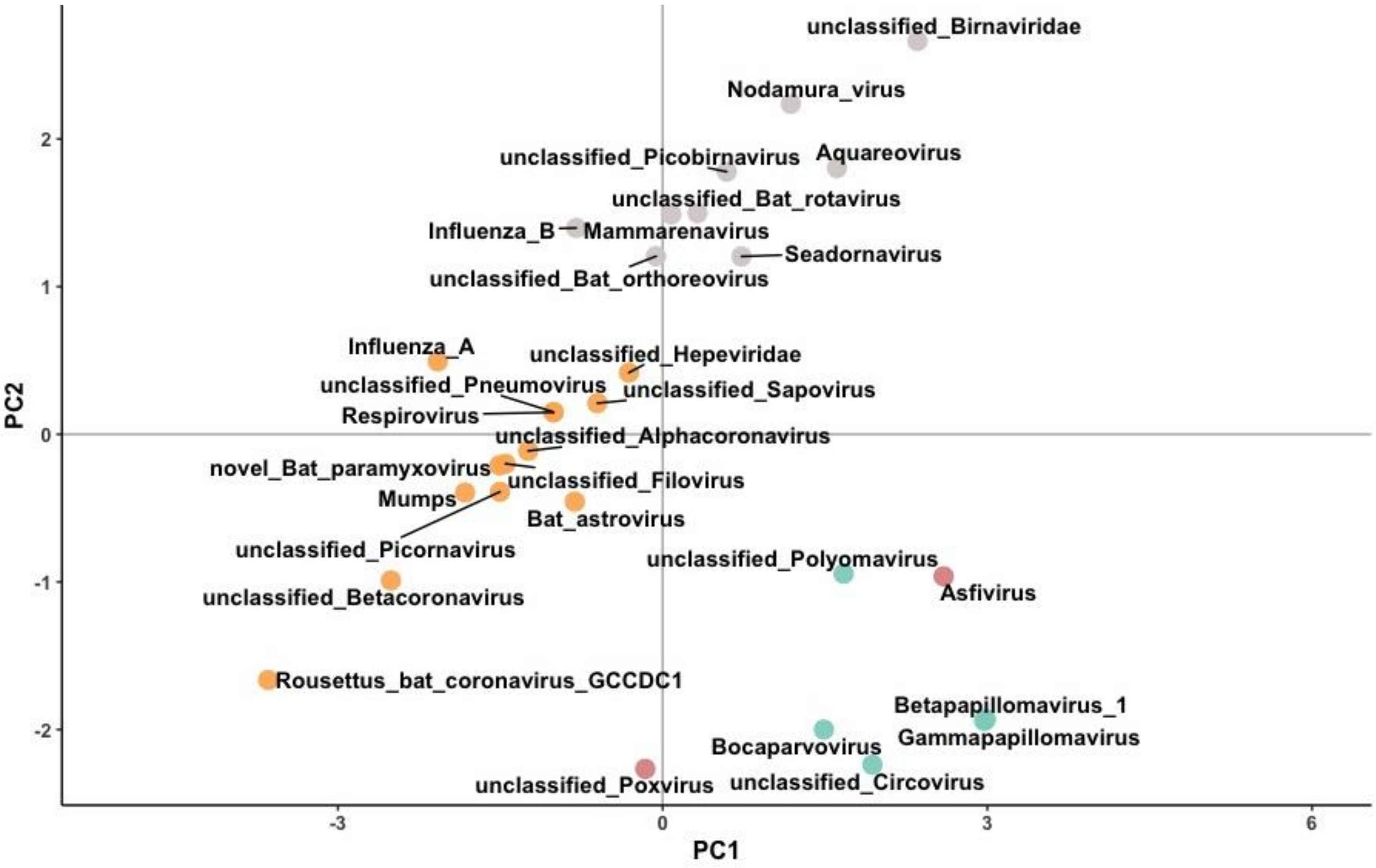


Figure 18. Principal component analysis of frequently-detected, zoonotic-related viruses (k=4) from unbiased shotgun sequencing data. Data are graphed in PC1 and PC2 space, representing 48% of the total variance. Cluster number (color): 1 (gold), 2 (red), 3 (grey), 4 (green).

multivariate characteristics that are prone to zoonotic spillover such as those possessed by orthomyxoviruses, paramyxoviruses, and coronaviruses (181).

In addition to the vast complexity of each swab type, we observed a surprisingly even distribution of detection among swab types. As shown in **Figure 19**, many viruses are detected in all four swab types. As previously discussed, there is a consistent pattern of dietary and other viruses that contrasts with the sporadic detection of zoonotic-related viruses. Some individual bats appear to shed or carry a larger number of zoonotic-related taxa as compared to other individuals. We hypothesize that this is influenced by both the immunological predispositions and behavior of each bat.

DISCUSSION

To our knowledge, this is the first comprehensive, longitudinal study to evaluate virome dynamics in a colony of old world bats. Beyond the challenge of limited known viral sequences, it can be difficult to detect viral genomes amidst the "noise" of bacterial commensals in a complex environmental sample (54). Computational methods for virome characterization rely on database-independent analyses such as *de novo* assembly to obtain scaffolded or complete viral genomes that share little homology but can be compared to distantly-related, published references. In particular, Meta-SPAdes is an effective tool for *de novo* assembly of virome data (132; 166). With limited tools to classify viruses from complex samples, in addition to the diversity of previously unknown viral genomes, the full characterization of a virome requires extensive manual analysis.

Two sequencing methods, unbiased shotgun sequencing and viral target enrichment, were utilized to evaluate the extent of viruses of biosurveillance concern.

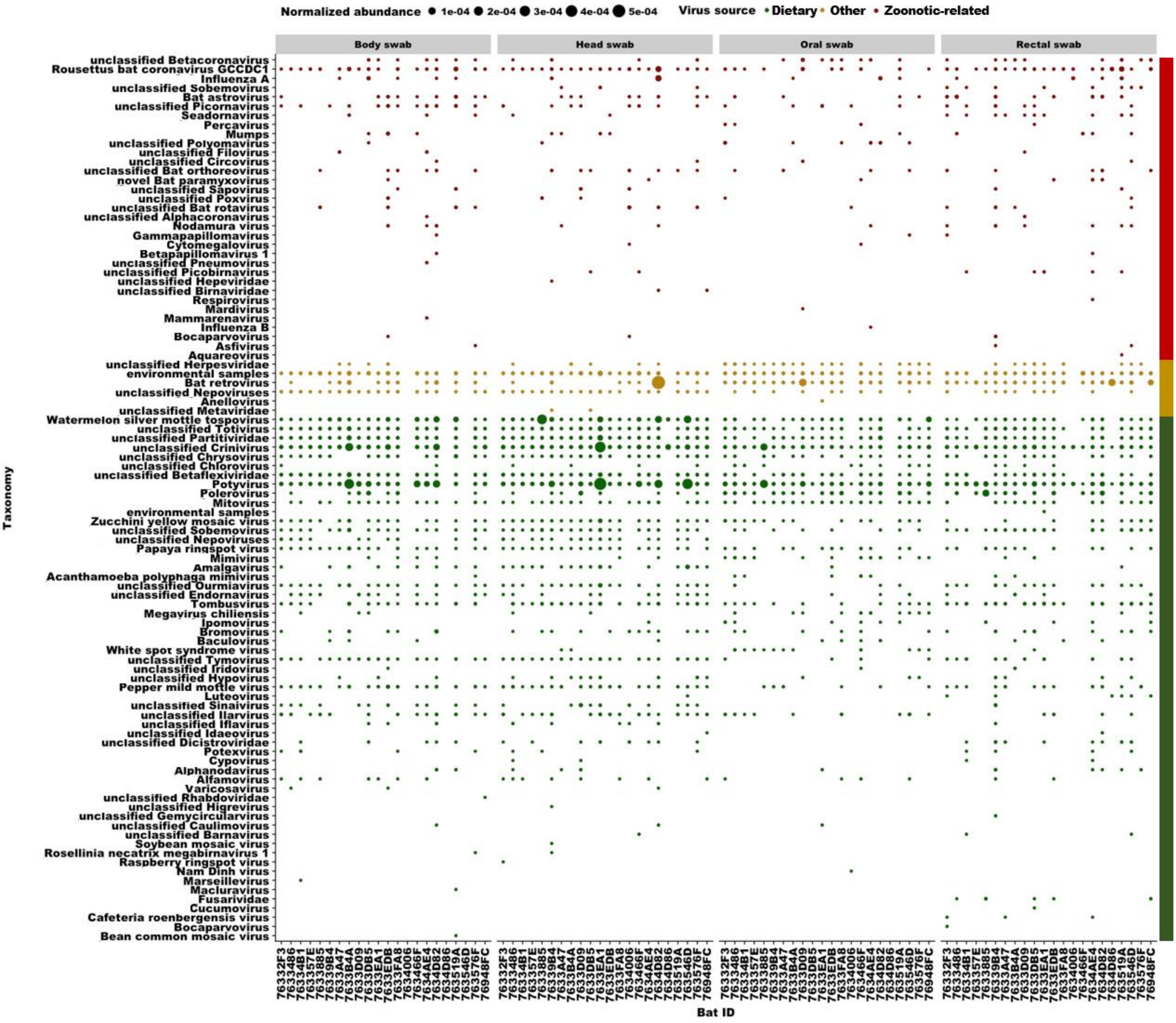


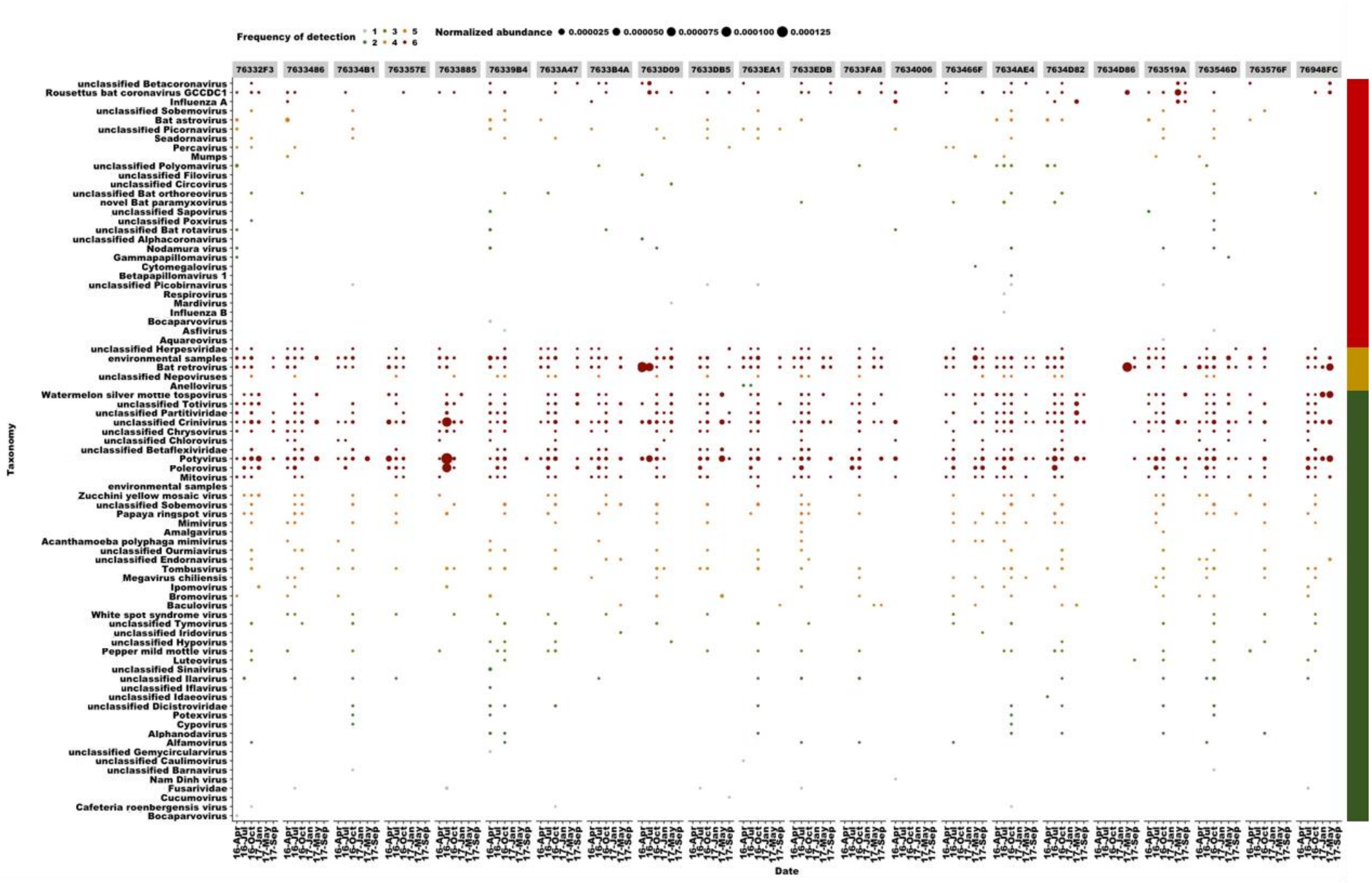
Figure 19. Each virus detected in unbiased shotgun sequencing data is listed along the y-axis, organized by zoonotic-related (red), other (yellow) and dietary (green). The identification code for each bat present in the colony at any time point is along the x-axis, and data is parsed by swab type (head, body, oral, rectal swabs).

In so doing, we were able to evaluate the strengths and limitations of target enrichment, while also obtaining informative data with regard to the colony-level virome in a captive colony of bats. Our results showed that there is sufficient binding tolerance to enrich for near-neighbors of probe targets. This flexibility enabled for the detection of viruses of biosurveillance concern in our samples, however, unbiased shotgun sequencing was necessary for characterization of all known viruses. Despite the improvement in detection of classifiable reads gained by the use of target enrichment, less than 1% of target-enriched or shotgun reads that were possibly viral could be unequivocally classified. This is consistent with reports that viral dark matter can represent as much as 90% of sequences (92). Nevertheless, approximately 5.3 Mbp of 293 Gbp of viral sequence were unequivocally classified. Coverage of detectable taxonomic groups varied in breadth and depth, as is typical in sequencing of metagenomic samples (54). To address the challenge of low coverage depth in data including viruses with varied genome lengths, a normalization approach that takes into account the nucleotide length of the virus genome was used and with a requirement of both forward and reverse read coverage.

Virus detection rates vary with several factors, including specimen type, cold chain logistics, date of sampling, host age and taxonomy. Common surveillance methodologies include culture/isolation, serological assays, conventional and quantitative polymerase chain reaction (PCR) assays, high throughput sequencing (HTS) and target enrichment sequencing (206). Each of the aforementioned methods has advantages and disadvantages with regard to portability and necessary prior knowledge about the viruses present within a given sample. Rarefaction curves were produced using VirusSeeker data to evaluate the extent at which mammalian viral diversity was recovered by using zoonotic-related species detected in each sample **(Figure 20)**. The rarefaction curves illustrate that some samples were more completely sampled by comparison and that there is variation in the detection of zoonotic-related species per sample. One caveat to this study is that there is a correlation between the number of detected taxa and the number of samples successfully prepared for sequencing. Multivariate factors may contribute to the failure of samples that were unable to be sequenced, including low concentration or quality RNA, decreased shedding of certain viral taxa, behavior changes in the bat or technical error in preparation.

Consistently detected dietary-associated viruses such as watermelon silver mottle tospovirus could be attributed to the bat's diet being supplemented with watermelon as part of their enrichment regime. For this reason, dietary-associated viruses that were consistently reintroduced by the diet or environment were used as a comparison alongside zoonotic-related viruses that are not found in dietary sources. Interestingly, several relevant zoonotic agent-related viruses that were detected at four or more time points have been associated with prior viral spillover (*Coronaviridae, Reoviridae, Paramyxoviridae*) from bats to humans. Our findings are consistent with the existing literature (122-124), in that Pteropodidae family bats have been previously demonstrated to be associated with viruses of the following: *Paramyxoviridae, Adenoviridae, Herpesviridae, Astroviridae, Coronaviridae, Rhabdoviridae, Polyomaviridae, Flaviviridae, Iflaviridae, Hepadnaviridae, Bunyavirales, Togaviridae, Caliciviridae, Orthomyxoviridae, Papillomaviridae, Reoviridae, Retroviridae, Filoviridae, Parvoviridae* and *Circoviridae* (206).

Figure 20. Classified viruses detected in unbiased shotgun sequencing data are arranged by zoonotic, other and dietary source along the y-axis. Within each group, taxa are arranged by descending frequency of detection. The x-axis marks sample collection dates and is parsed by individual bat identification number (top x-axis label). Points are color-coded to frequency of detection, with the most frequently detected taxa plotted in red.



🟥 **Zoonotic-related**
🟨 **Other**
🟩 **Dietary**

We observed population persistence of the following viral families: *Astroviridae, Orthomyxoviridae, Picornaviridae, Paramyxoviridae, Reoviridae* and *Coronaviridae*. Of these families, *Paramyxoviridae, Reoviridae* and *Coronaviridae* are each known to be associated with spillover from bats to humans (7; 43; 181). This knowledge should inform future directions for biosurveillance of viruses within each of the aforementioned families that have the potential to cross the species barrier from bats to humans or amplifying hosts.

The circulation of orthomyxoviruses and filoviruses in wild reservoirs is relevant to public health and these viruses were sporadically detected in the colony. Although influenza A virus has previously been detected in South American bats (171; 172), the capacity for old world bats to serve as a reservoir for influenza A-like viruses was only recently discovered (85). Questions have been raised with regard to the geographic range of bat-borne influenza viruses and the capacity for reassortment of bat-borne orthomyxoviruses with known human influenza A viruses (198). Serologic evidence of filoviruses has been reported in lesser dawn bats (98) but, to our knowledge, sequence data has not been previously reported. It is apparent from both target enrichment and shotgun sequence data that a novel filovirus was present in one rectal swab and two body swabs, but there was not significant breadth of coverage as compared to any previously known filovirus reference genomes. The closest sequenced relative is Mengla virus, which was recently isolated from Rousettus bats in China (202) (**Figure 21**).

Zoonotic-related viruses were intermittently detected in samples from individual bats. One caveat of this study lies within the variability of nucleic acids that could be sequenced. Results with regard to individual bats should be interpreted with caution because it is unknown whether this phenomenon was observed due to intermittent shedding below the limit of detection or cycles of recurrent infection (susceptible-infectious-recovered-susceptible [SIRS] or susceptible-infectious-latent-infectious [SILI] models of infection (147)). However, one compelling finding of this study was that multiple viruses exhibited population persistence despite the isolation from contact with new, wild bats for 20 months. We interpret these results to indicate that persistent viruses possess a propensity for spillover due to their comparatively higher recurrence in shedder sites at high enough abundance for detection. These viruses share multivariate qualities with viruses known to infect humans (**Figure 22**) and could represent viruses that are most likely to cross the species barrier to humans or intermediate hosts in the manner that henipaviruses and SARS-CoV caused outbreaks in the past. The persistence of *Paramyxoviridae, Reoviridae* and *Coronaviridae* family viruses without external reintroduction to this captive community indicates that lesser dawn bats may serve as a maintenance host. As changing human factors such as urbanization impact the potential for disease interface by increasing urban-adapted (synanthropic) animal populations (71), targeted assays such as PCR, for these persisting viruses should continue to be conducted by regional public health laboratories engaged in emerging infectious disease surveillance efforts.

Furthermore, viruses that were detected less frequently may be of a lesser concern for biosurveillance as they did not exhibit robust, colony-level persistence. In particular, this study utilized swabs of the exterior of the bat (head and body) to evaluate the virus population. This could be a useful sample site for future surveillance to extrapolate population-level infection and recapitulates the observation that host-microbiome dynamics of gregarious species such as bats should be observed on the colony rather than an individual level (90). Taken together, we conclude that noninvasive surveillance methods that target the body of bats not only detect viruses shed within the colony, but can also represent viral populations dispersed throughout the entire colony. As shown in **Figure 7**, the results across internal (oral and rectal) and external (head and body) swabs are homogeneous and reflect the viral populations for the entire colony. External swabs could be informative targets for sample collection. In conclusion, we have provided novel insight to the virome of South-eastern Asian bats that should be used to inform future surveillance methods in the region.

Acknowledgements and Funding

The authors would like to thank Dr. Danielle E. Anderson for assistance with the bat colony and Dr. Jens Kuhn for careful reading of the manuscript.

This work was funded by U.S. Navy, Office of Naval Research, In-House Laboratory Independent Research (ILIR) Program and WUN A1417. The work in the Singapore team was funded by the Singapore National Research Foundation grants (NRF2012NRF-CRP001-056 and NRF2016NRF-NSFC002-013). The funding bodies had no role in the design of this study, nor in the collection, analysis, or interpretation of data, nor in the writing of this manuscript.

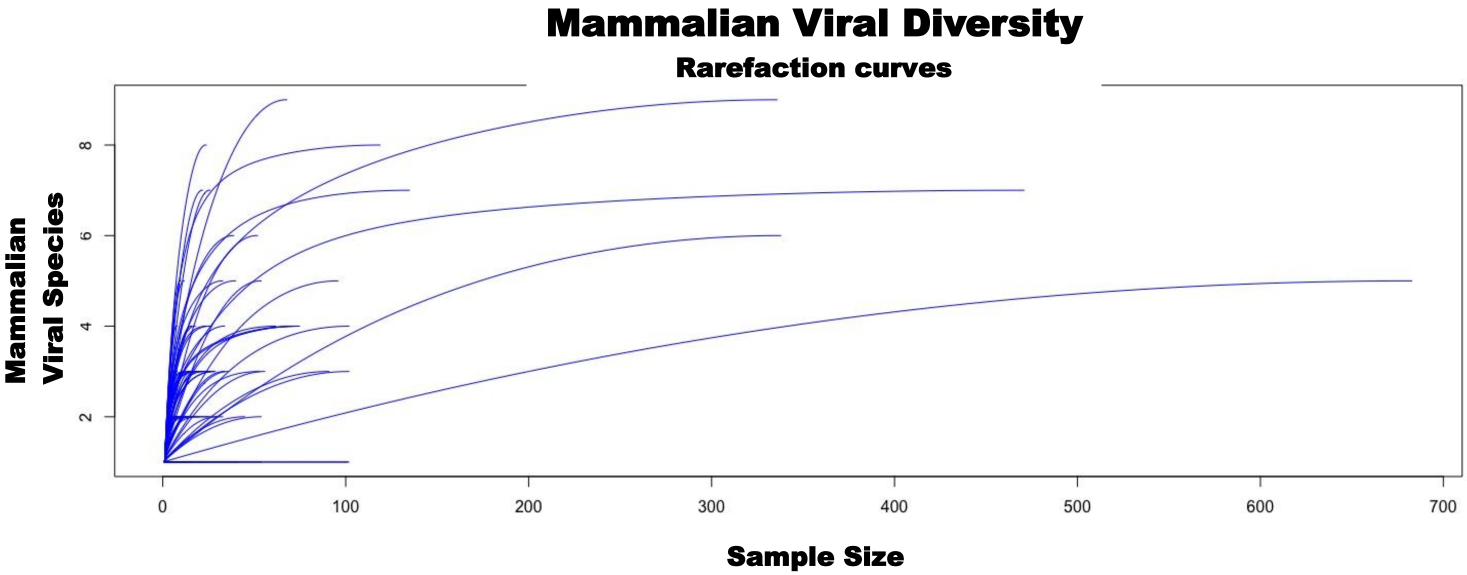


Figure 21. Rarefaction curves were produced using VirusSeeker data to evaluate the extent at which mammalian viral diversity was recovered by using zoonotic-related species detected in each sample. The rarefaction curves illustrate that some swabs were more completely sampled by comparison and that there is variation in the detection of zoonotic-related species per sample. It is probable that with deeper sequencing or greater starting concentrations of RNA, more mammalian viruses could be detected.

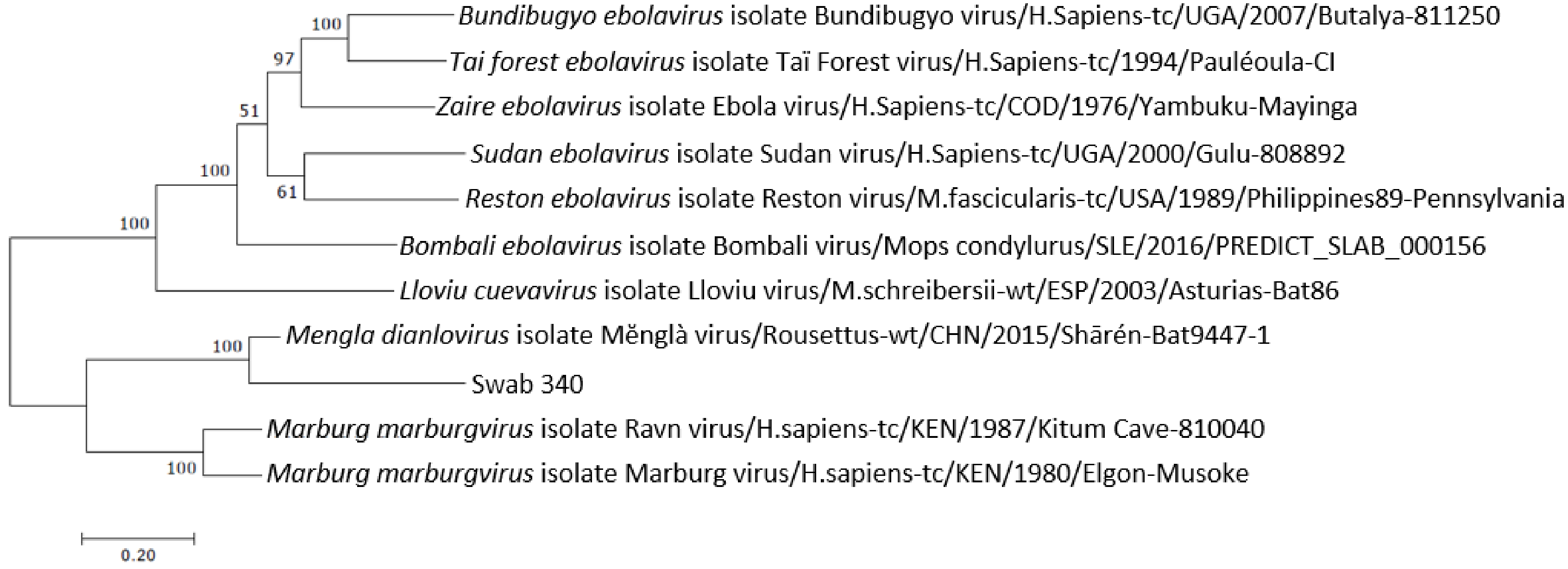

100
97
51
100
61
100
100
100
Bundibugyo ebolavirus isolate Bundibugyo virus/H.Sapiens-tc/UGA/2007/Butalya-811250
Tai forest ebolavirus isolate Taï Forest virus/H.Sapiens-tc/1994/Pauléoula-CI
Zaire ebolavirus isolate Ebola virus/H.Sapiens-tc/COD/1976/Yambuku-Mayinga
Sudan ebolavirus isolate Sudan virus/H.Sapiens-tc/UGA/2000/Gulu-808892
Reston ebolavirus isolate Reston virus/M.fascicularis-tc/USA/1989/Philippines89-Pennsylvania
Bombali ebolavirus isolate Bombali virus/Mops condylurus/SLE/2016/PREDICT_SLAB_000156
Lloviu cuevavirus isolate Lloviu virus/M.schreibersii-wt/ESP/2003/Asturias-Bat86
Mengla dianlovirus isolate Měnglà virus/Rousettus-wt/CHN/2015/Shārén-Bat9447-1
Swab 340
Marburg marburgvirus isolate Ravn virus/H.sapiens-tc/KEN/1987/Kitum Cave-810040
Marburg marburgvirus isolate Marburg virus/H.sapiens-tc/KEN/1980/Elgon-Musoke
0.20

Figure 22. Phylogenetic analysis of a novel bat filovirus demonstrates that the unclassified bat filovirus is most closely related to Mengla dianlovirus. A maximum likelihood phylogeny using a consensus sequence of reads from a body swab was generated using MEGA7 based on the General Time Reversible model (94; 130).

CHAPTER 4: Detection of recombinant Rousettus bat coronavirus GCCDC1 in lesser dawn bats (*Eonycteris spelaea*) in Singapore

This work is in preparation for publication as: **Adrian C. Paskey, Justin H. J. Ng, Gregory K. Rice, Wan Ni Chia, Casandra W. Philipson**, **Randy J.H. Foo, Regina Z. Cer, Kyle A. Long, Matthew R. Lueder, Xiao Fang Lim, Kenneth G. Frey, Theron Hamilton, Danielle E. Anderson, Eric D. Laing, Ian H. Mendenhall, Gavin J. Smith, Lin-Fa Wang, Kimberly A. Bishop-Lilly. 2020.** "Detection of recombinant Rousettus bat coronavirus GCCDC1 in lesser dawn bats (*Eonycteris spelaea*) in Singapore."

The work presented here is the sole work of A.C.P. with the following exceptions: J.H.J.N., W.N.C., R.J.H.F., X.F.L., D.E.A., I.H.M., and G.J.S. were responsible for the maintenance and origination of the captive bat colony. G.K.R., C.W.P., R.Z.C. and M.R.L. contributed bioinformatic scripts that were modified to perform analyses or generate figures. K.A.B-L., L-F.W., T.H., K.G.F. and E.D.L. provided valuable guidance for the analyses of these data and assisted with manuscript preparation.

ABSTRACT

Rousettus bat coronavirus GCCDC1 (RoBat-CoV GCCDC1) is a cross-family recombinant coronavirus that has previously only been reported in wild-caught bats in Yúnnan, China. We report the persistence of a related strain in a captive colony of lesser dawn bats captured in Singapore. Genomic evidence of the virus was detected using targeted enrichment sequencing and further investigated using deeper, unbiased high throughput sequencing. RoBat-CoV GCCDC1 Singapore shared 96.52% similarity with RoBat-CoV GCCDC1 356 (NC_030886) at the nucleotide level and had a high

prevalence in the captive bat colony. It was detected in five of six sampling time points across the course of eighteen months. A partial segment 1 from an ancestral Pteropine orthoreovirus makes up the recombinant portion of the virus, which shares high similarity with previously reported RoBat-CoV GCCDC1 strains that were detected in Yúnnan, China. RoBat-CoV GCCDC1 is an intriguing, cross-family recombinant virus with a geographical range that expands farther than was previously known. The discovery of RoBat-CoV GCCDC1 in Singapore indicates that this recombinant coronavirus exists in a broad geographical range and can persist in bat colonies long-term.

INTRODUCTION

Human coronaviruses have been studied for just over half a century and are a priority for research as a consequence of recent high-profile disease outbreaks (119; 173). The first coronavirus detection from bats was reported in 2005 following increased surveillance of wildlife in response to the SARS-CoV outbreak (148). Since then, 3,796 coronavirus variants have been detected in feces, swabs, tissue, blood, and urine of wild collected bats from surveillance research spanning 58 countries (Database of Bat-Associated Viruses as of November 2019) (25). Of the seven known human coronaviruses (HCoVs), four are regarded to cause mild to limited disease and circulate endemically: HCoV-NL63, HCoV-229E, HCoV-OC43 and HKU1 (31). Of the remaining HCoVs, severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle Eastern respiratory syndrome coronavirus (MERS-CoV) are WHO R&D Blueprint priority pathogens that represent potential epidemic threats lacking effective on-hand countermeasures (119; 174). The most recently discovered human coronavirus, SARS-CoV-2, was first detected in December 2019, is known to cause mild to severe respiratory disease in humans, and is most similar to a bat SARS-like CoV (111). SARS-CoV and MERS-CoV are transmitted to humans through palm civets and camels, respectively, where the terrestrial mammals act as amplifying intermediate hosts (97). Phylogenetic reconstructions suggest that the putative ancestor of both viruses are both bat coronaviruses (9(1). pii: E41.; 28; 110). Although members of the coronavirus genus have been found to infect a wide diversity of animal hosts, alpha- and betacoronaviruses are thought to have bat origins, whereas gamma- and deltacoronaviruses are thought to be of avian origin (191; 193).

The viral family *Coronavridae* is so named due to the morphology of the virion, which resembles a crown; thus, the name is derived from the Greek word for crown (191). Coronavirus genomes undergo a high frequency of recombination, particularly in the spike gene, which is responsible for viral entry into host cells (164). This potentially leads to an increase in the capacity to infect a wider range of hosts and/or altered virulence (164). Coronaviruses are enveloped positive-sense, single-stranded RNA viruses with nonsegmented genomes approximately 30 kb in length (64). Subgenomic RNA generated during viral replication increases the likelihood for homologous recombination with coinfecting viruses by way of template switching, resulting in novel variants (164). Recombination between caliciviruses and retroviruses, which belong to different virus families but infect the same avian host, has been implicated in the generation of novel avian and porcine circoviruses (35). Viral recombinants that emerge from animal reservoirs may have the potential to infect a broader spectrum of intermediate hosts or spill over into human populations. For this reason, recombination of potentially zoonotic viruses has been studied extensively with the intent to predict the emergence of virulent, recombinant coronaviruses (164).

SARS-CoV, SARS-CoV-2, MERS-CoV and Rousettus bat coronavirus HKU9 (RoBat-CoV HKU9) are ancestrally related as bat-borne betacoronaviruses (111). RoBat-CoV HKU9 was first reported in 2007 and is one of numerous betacoronaviruses discovered during biosurveillance sampling in the wake of the SARS-CoV outbreak to identify the animal source (195). In addition to hosting coronaviruses, bats are regarded as the primary animal hosts of an exceptionally diverse level of viruses, including paramyxoviruses (e.g. Nipah virus), filoviruses (e.g. Marburg virus) and orthoreoviruses (4; 43; 69; 136). Orthoreoviruses are nonenveloped, viruses with ten genome segments comprised of double-stranded RNA. Multiple fusogenic orthoreoviruses, defined by their ability to cause cell syncytia, circulate in Southeast Asian bats and are known to cause disease in humans (190). As animal reservoirs for diverse families of viruses, bats represent a mammalian host in which co-infecting viruses potentially recombine. A cross-family recombinant virus was discovered in a population of wild Leschenault's rousette (*Rousettus leschenaultii*) bats in 2016 in China (80). This virus, Rousettus bat coronavirus GCCDC1 (RoBat-CoV GCCDC1), possesses a partial segment 1 from Pteropine orthoreovirus incorporated into the backbone of Rousettus bat coronavirus HKU9 between the N and NS7a genes (195).

Following the discovery of RoBat-CoV GCCDC1 in Leschenault's rousette in 2016, in Yúnnan, China, this virus was further detected in an lesser dawn bat (*Eonycteris spelaea*) population known to co-roost with Leschenault's rousette (114). It is thought that the dense population of bat roosts and gregarious behavior contribute to the persistence of RoBat-CoV GCCDC1 (135). In an analysis of viral families, coronaviruses were found to have a level of topological distance from known hosts that suggests frequent cross-species transmission among mammals and host switching (59). In this study, we conducted longitudinal virome analysis of a wild-caught, captive colony of lesser dawn bats in Singapore. We have previously reported the virome analysis of this colony (Paskey *et al*. 2020 Virus Evolution), which was captured from a wild colony known to host a lineage D betacoronavirus (122). Using a hybridization-based targeted enrichment sequencing approach (140), we detected genomic evidence of RoBat-CoV GCCDC1 in this colony of lesser dawn bats and then performed unbiased shotgun sequencing. This is now the third report of this recombined coronavirus/orthoreovirus in a bat host, and the first report of the persistence of this virus in bats beyond the geographical region of Yúnnan, China. The findings of this study are of significant relevance to Asia-Pacific regional public health laboratories due to the implications for disease prevention and control.

RESULTS

Prevalence of GCCDC1 in the captive colony

RoBat-CoV GCCDC1 was observed in 72 of 206 samples that were collected from a captive colony of lesser dawn bats in 2016 and 2017 and characterized via shotgun sequencing. The virus was detected in five of six sampling dates in head, body, oral and rectal swabs (**Table 9**). Previous longitudinal studies that evaluated the persistence of RoBat-CoV GCCDC1 in wild bats in Yúnnan province, China reported a prevalence of 39.3% in 2014, 35.6% in 2015 (135), as well as another report of 5.26% in

Table 9. Summary of prevalence of RoBat-CoV GCCDC1 in swabs collected over the course of 18 months.

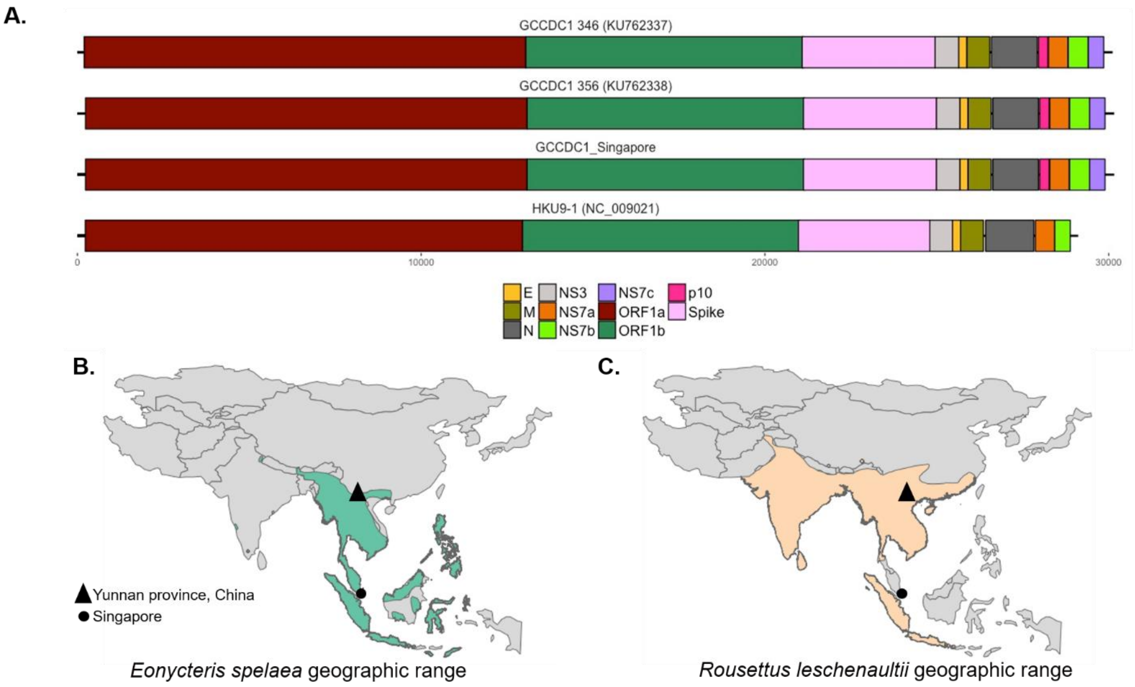
Sampling date	# Swabs sequenced	# Swabs positive for GCCDC1 (%)	# Bats sampled	# Bats positive for GCCDC1 (%)
April 2016	41	11 (26.8%)	18	8 (44.4%)
July 2016	28	4 (14.3%)	19	4 (21.1%)
October 2016	75	55 (73.3%)	20	20 (100%)
January 2017	14	1 (7.1%)	11	1 (9.1%)
May 2017	21	0 (0%)	15	0 (0%)
September 2017	27	1 (3.7%)	13	1 (7.7%)

2015 and 18.87% in 2016 (114). The average prevalence among sequenced swabs in this study of captive bats was 38.1% in 2016 and 3.6% in 2017.

Comparison to previously discovered cross-family recombinant coronaviruses

The genome of RoBat-CoV GCCDC1 includes a p10 gene from segment S1 of an ancestral orthoreovirus inserted between the RoBat-CoV HKU9 nucleocapsid (N) and NS7a genes (**Figure 23A**). The same transcription regulatory sequence (TRS) motif, 5'-ACGAAC-3', is shared between RoBat-CoV GCCDC1 and RoBat-CoV HKU9, with the identical alteration of one nucleotide to 'TCGAAC' in the intergenic TRS before the envelope gene in both RoBat-CoV GCCDC1 and HKU9 (80; 195). Previously reported RoBat-CoV GCCDC1 sequences detected in China share a high identity (114). The nucleotide similarity between the full genome detected in Singapore and RoBat-CoV GCCDC1 356 is 96.52%. We report notable similarity between strains despite host and geographic differences. RoBat-CoV GCCDC1 was first detected in China where the geographic range of *E. spelaea* and *R. leschenaultii* overlap (triangles, **Figure 23B-C**).

Next, we investigated viral genes that would reflect host adaptation or interspecies codon usage bias (11; 165). Dinucleotide analysis of each gene and rho value calculation was utilized to evaluate any host replication biases reflected among *E. spelaea* and *R. leschenaultii* species. A dinucleotide bias among strains of RoBat-CoV GCCDC1 has not been previously investigated and we hypothesized that such a bias, if detected, may result from circulation in distinct bat host species. Upon investigation, we detected no significant difference in rho value or notable variation at the nucleotide level in the reovirus-derived segment p10 (**Figure 24**). Furthermore, we detected a high level of conservation for p10 at the amino acid level between *E. spelaea* (Singapore), *E. spelaea*



Eonycteris spelaea geographic range

Rousettus leschenaultii geographic range

Figure 23. RoBat-CoV GCCDC1 Singapore is genetically similar to the strain detected in China and its host is found in a wide geographic range. (A) The distribution of orthoreovirus p10 gene across three strains of RoBat-CoV GCCDC1: 346, 356 and Singapore, as well as a putative parental relative, RoBat-CoV HKU9-1. Note the presence of the p10 gene insertion (dark pink) and NS7c (purple) only in the RoBat-CoV GCCDC1 strains. The geographic distribution of (B) *E. spelaea* and (C) *R. leschenaultii* across Southeast Asia, respectively (51).

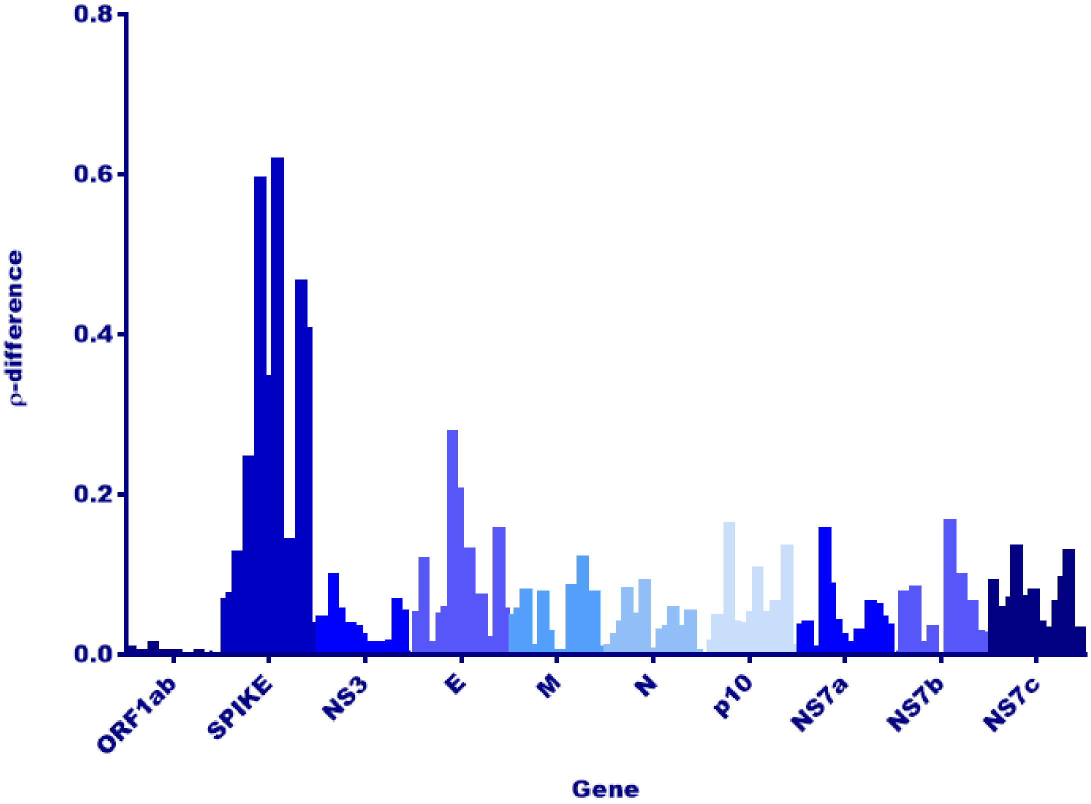


Figure 24. Dinucleotide usage analysis among RoBat-CoV GCCDC1 strains. By two-way ANOVA, the rho-difference for the spike gene is significantly different from all other genes (alpha .05, p < .0001). It is not surprising that the spike gene is the region with the greatest variation. This may indicate that while RoBat-CoV GCCDC1 356 and Singapore may be highly related, they function differently.

(China) and *R. leschenaultii* RoBat-CoV GCCDC1 (**Figure 25**). The p10 amino acid sequence of the Singapore strain is 98.6% similar at the nucleotide level to previously published strains 346 and 356, falling within the 'Group A' categorization of p10 sequences described by *Obameso et al.* (135).

Detection of the recombinant genome despite geographic and host differences

We did not detect the putative parental strain (RoBat-CoV HKU9) of this recombinant in the sampled cohort of captive lesser dawn bats. We extrapolate from our observations that the recombinant is circulating in wild lesser dawn bats and did not arise as a recombinant in the captive colony because the backbone of RoBat-CoV GCCDC1 is distinct at the nucleotide level from RoBat-CoV HKU9. Moreover, the p10 segment is genetically distinct from known orthoreoviruses. The phylogenetic analysis of the full nucleotide sequence as compared to previously published references for strains of GCCDC1 and RoBat-CoV HKU9 (**Figure 26**) illustrates the relatedness of RoBat-CoV GCCDC1 sequences. The paired reads and contigs that span both junctional regions of p10 provides evidence that the sequence data represent a true recombinant virus. Furthermore, we detected the shedding of RoBat-CoV GCCDC1 using both NextSeq500 and MiSeq platforms by shotgun and targeted enrichment sequencing, respectively.

Discussion

The evolution of viruses of pandemic potential, such as bat-borne coronaviruses, is significant to public health due to the risk of spillover. Viral recombination in host reservoirs is a concern for public health, as these events can increase the potential for spillover. Here, we report genomic evidence of the recombinant RoBat-CoV GCCDC1 in a population of lesser dawn bats in Singapore. Interestingly, this recombinant virus strain

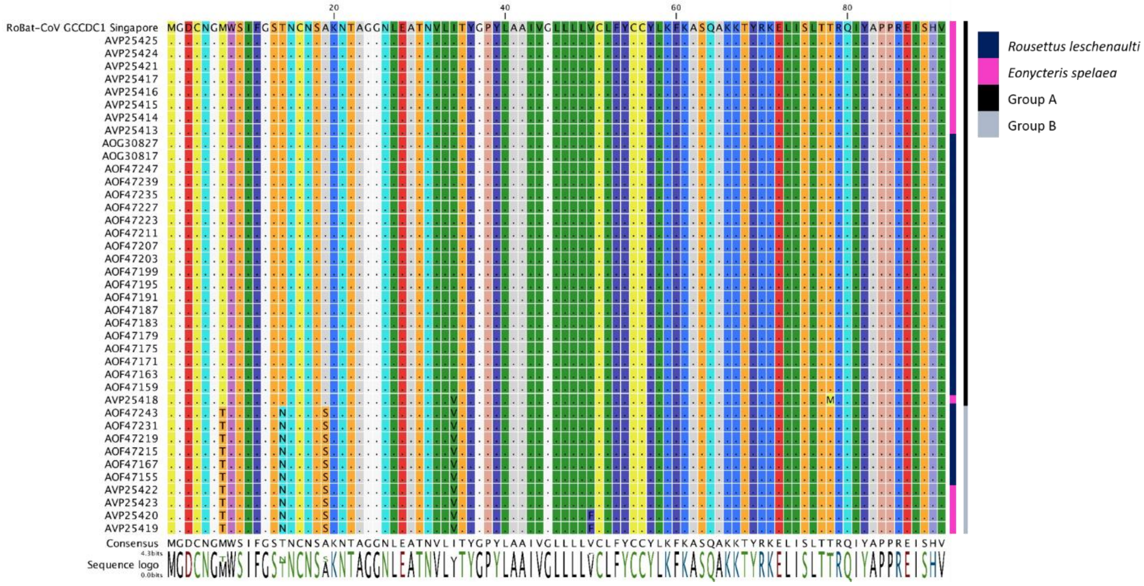


Figure 25. Multiple sequence alignment (MSA) of the amino acid sequence for p10 from RoBat-CoV GCCDC1 of the strain detected in this study and p10 sequences reported from Yúnnan province, China (80; 114; 135). The colored bars to the right of the MSA indicate that the sequence was detected in *R. leschenaultii* (dark blue) or *E. spelaea* (pink). Groups A (black) and B (grey) were previously defined in the literature (135).

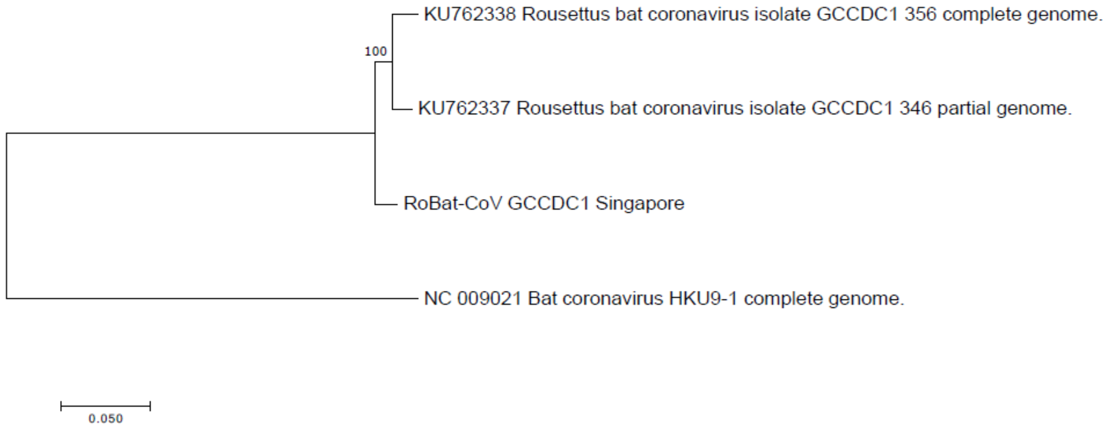

KU762338 Rousettus bat coronavirus isolate GCCDC1 356 complete genome.
100
KU762337 Rousettus bat coronavirus isolate GCCDC1 346 partial genome.
RoBat-CoV GCCDC1 Singapore
NC 009021 Bat coronavirus HKU9-1 complete genome.
0.050

Figure 26. Phylogenetic analysis of RoBat-CoV GCCDC1 Singapore strain as compared to previously published references and related RoBat-CoV HKU9. Nucleotide-level analysis was performed using full genome alignments and the maximum likelihood method based on the General Time Reversible model in MEGA7 with 100 bootstrap replicates (94).

exhibits genetic conservation as compared to strains initially detected in Yúnnan, China. WTherefore, we have demonstrated that a cross-family recombinant coronavirus persists in a captive colony of bats and is similar at the nucleotide level to previously discovered strains, despite geographic and host differences. There is high similarity between RoBat-CoV GCCDC1 Singapore and previously reported strains by gene arrangement (**Figure 23A**), high conservation at the amino acid level within the p10 insertion (**Figure 24**), and across the whole genome at the nucleotide level (**Figure 25**). One unusual element of the backbone of RoBat-CoV GCCDC1 is the presence of nonstructural protein-encoding gene NS7c at the 3’ end of the genome.

This genome arrangement is most similar to that of deltacoronaviruses, which can infect humans but are typically found in birds or pigs (192; 194). The betacoronavirus related to HKU9-1 is found in bats and does not possess NS7c (102). Taken together, it is possible that RoBat-CoV GCCDC1 is the product of two or more historical recombination events.

The virus was detected less frequently in the second year of the longitudinal study. It is unclear if prevalence of RoBat-CoV GCCDC1 was impacted during the first year of the study by unknown confounding factors. We hypothesize that RoBat-CoV GCCDC1 was able to persist in this lesser dawn bat colony but, with no immigration or emigration, it was not continually shed at levels that were detectable by our methods. This work indicates the ability for the virus to persist long-term within a captive colony of lesser dawn bats. It is important to note that this study evaluated the population persistence of RoBat-CoV GCCDC1 in a captive colony, which is unique from previous longitudinal reports in wild-caught bats that are exposed to variables such as weather changes or the ability to exchange genetic viral variants via dispersal and migration. The geographic distance and genetic similarity between strains provide insight to the possibility that this strain likely exists in the geographical region between Yúnnan province and Singapore. Additionally, unbiased biosurveillance assays could detect other, yet-to-be-discovered cross-family recombinants.

Evidence of coronavirus infection has been detected in 198 bat species by a variety of methods such as conventional PCR, quantitative PCR, serology, and HTS (25). Coronaviruses are of high priority for biosurveillance due to their propensity to evolve quickly, prior history of zoonotic spillover, and subsequent high mortality rates in humans (164). Surveillance for coronaviruses of pandemic potential often targets the polymerase gene by screening noninvasive samples via PCR assays. Immunoassays are also invaluable in detecting previous exposure of viruses of zoonotic potential in reservoirs but they do not provide information with regard to current virus infection (145). One shortcoming of these approaches is the inherent bias toward known viruses, specifically highly conserved genomic regions, which can be circumvented with HTS. Progenitor coronavirus strains related to SARS-CoV or SARS-CoV-2 may exist in a cohort of diverse viral variants in reservoirs (120) and this depth of information can be discovered via an unbiased approach like HTS. Genetic diversity of viruses within a single host reservoir may permit frequent transmission to incidental hosts that becomes problematic when a genetic variant capable of infecting the new host population spills over (79). Given the propensity of coronaviruses to recombine, unbiased HTS, as used in this study, provides insight to the diversity of coronavirus genomes circulating in bat reservoirs.

A continuation of sampling from Leschenault's rousette and lesser dawn bat colonies in other geographic regions could reveal deeper insight into the circulation of RoBat-CoV GCCDC1 variants in wild bats. These bats inhabit regions between Singapore and Yúnnan province, China and may also carry RoBat-CoV GCCDC1 based upon similarities in viruses detected among groups of the same species. For example, recent serological biosurveillance of lesser dawn bat populations in Singapore and Northeast India demonstrated that both populations had similar exposure to Asiatic filoviruses (42; 98). The geographic range of the lesser dawn bat extends across Southeast Asia and it is unknown whether these populations are panmictic. As RoBat-CoV GCCDC1 continues to circulate in co-roosting populations of multiple bat species, the spike gene will be under pressure as demonstrated in **Figure 24**. While the cellular receptor for RoBat-CoV GCCDC1 and HKU9 is unknown, surveillance for adaptation and mutation of the spike gene should be performed to estimate risk of tropism for receptors found in intermediate amplifying hosts or humans.

Coronaviruses may emerge following random mutations permissive to infection of intermediate amplifying hosts and/or recombination events that result in a large pool of variants that could infect humans, potentially without an intermediate host (120). Due to the knowledge gap with regard to the circulation of recombinant viruses, an understanding of the prevalence of unique, recombinant viruses will provide an advantage to predict the innate features of a virus with greater propensity for spillover. RoBat-CoV GCCDC1 is an intriguing, cross-family recombinant virus with a geographical range that expands farther than was previously known.

METHODS

Bat colony structure and sampling strategy

Sampling was previously described in Chapter Three. In summary, bats were housed in stainless steel mesh cages with ample room for roosting and swabs were collected quarterly for health screening purposes. Head, body, oral and rectal swabs were obtained using polyester tipped swabs and stored in 2 ml screw cap micro tubes (Sarstedt, Germany) containing 500 µl viral transport media (VTM, 10% Bovine Serum Albumin, 20% Antibiotics-Antimycotic in milli-Q water) at -80°C.

Nucleic acid extraction and sequencing

Extraction and sequencing methods were previously described in Chapter Three. In summary, RNA was extracted from head, body, oral and rectal swabs of each bat using a QIAGEN RNeasy Kit with on-column DNase digestion (Qiagen; Valencia, CA). RNA was eluted twice with RNase-free water. A maximum of 18 µl of the extract was used as input to prepare RNA TruSeq libraries. Conventional HTS (shotgun) libraries were multiplexed for sequencing on the NextSeq500 platform using v2 chemistry with 2x150 bp read lengths. Post-library enrichment probe targets and preparation methods were previously described by *Paskey et al*; samples were probed in pools of 12 and multiplexed for sequencing on the MiSeq platform using v3 chemistry with 2x300 bp read lengths (140).

Bioinformatic analyses

Rousettus bat coronavirus GCCDC1 was first detected in target-enrichment samples by read-mapping to Rousettus bat coronavirus isolate GCCDC1 356 (NC_030886) using CLC Genomics Workbench V11 (QIAGEN Bioinformatics;

Redwood City, CA). Further analysis for prevalence of the related strain was performed using shotgun data mapped using bbsplit with parameter adjustment of minid = .75 (19). Results were filtered to require both forward and reverse reads covering more than 100 bases. Multiple sequence alignments, as well as variant analysis of contigs and reads were performed with CLC Genomics Workbench V11 (QIAGEN Bioinformatics; Redwood City, CA). Dinucleotide variation was evaluated by calculating Rho using seqinr package for R (24; 167).

Phylogenetic analysis

Molecular phylogenetic analysis was performed with 100 bootstrap replicates using the Maximum Likelihood method based on the General Time Reversible model in MEGA7 (94). The tree with the highest log likelihood (-74447.85) is shown. The analysis involved 4 nucleotide sequences: the genome sequence generated from this study (RoBat-CoV GCCDC1 Singapore), GCCDC1 356 (KU762338), GCCDC1 346 (KU762337) and HKU9-1 (NC_009021). All positions containing gaps and missing data were eliminated. There were a total of 28,712 positions in the final dataset.

Geographic range of bats

Rousettus leschenaultii and *Eonycteris spelaea* range data was obtained in shapefiles format from the International Union for Conservation of Nature and Natural Resources (ICUN) (51). Ranges were mapped using ‘tmap’ in R (167; 168).

VERTEBRATE ANIMAL CARE AND SAFETY

All bats were housed and handled by Duke-National University of Singapore Medical School animal facilities. Trained laboratory personnel provided daily care for the animals according to the guidelines agreed upon by Duke-NUS Institutional Animal Care and Use Committee (2015/SHS/1088) and the Agri-Food and Veterinary Authority of Singapore. All samples were noninvasive.

FUNDING

This work was funded by U.S. Navy, Office of Naval Research, In-House Laboratory Independent Research (ILIR) Program and WUN A1417. The work by the Singapore team was funded by the Singapore National Research Foundation grants (NRF2012NRF-CRP001-056 and NRF2016NRF-NSFC002-013). The funding bodies had no role in the design of this study, nor in the collection, analysis, or interpretation of data, nor in the writing of this manuscript.

CHAPTER 5: Relatives of rubella virus in diverse mammals portend challenges for global rubella eradication

This work submitted for publication as: **Andrew J. Bennett*, Adrian C. Paskey*, Arnt Ebinger*, Florian Pfaff, Grit Premer, Dirk Hoper, Jens H. Kuhn, Kimberly A. Bishop-Lilly, Martin Beer, Tony L. Goldberg. 2019.** "Relatives of rubella virus in diverse mammals portend challenges for global rubella eradication."

The work presented here is the sole work of A.C.P. with the following exceptions: A.J.B., T.L.G., F.P., G.P and D.H. collected samples. A.E., F.P., G.P., D.H. and M.B. discovered and characterized rustrela virus in Germany. K.A.B.-L, A.J.B., T.L.G and J.H.K. contributed to data analysis and interpretation. *Indicates shared first-author.

ABSTRACT

We describe the first known close relatives of rubella virus (*Matonaviridae*: *Rubivirus*) in apparently healthy cyclops leaf-nosed bats (*Hipposideros cyclops*) in Uganda (ruhugu virus), in an acutely encephalitic donkey (*Equus asinus*), and in a Bennett's tree-kangaroo (*Dendrolagus bennettianus*) in a German zoo (rustrela virus). Ruhugu and rustrela viruses share identical genomic architecture with rubella virus and are near phylogenetic outgroups to all known human rubella genotypes. Surprisingly, an important neutralizing epitope of rubella virus's primary antigenic protein E1 is almost completely conserved in the homologous ruhugu virus E1 protein, suggesting serological cross-reactivity between the two viruses. Furthermore, modeling of E1 homotrimers in the post-fusion state predicts similar membrane fusion capacity for ruhugu virus and rubella virus. These functionally relevant similarities, and the ability of rustrela virus to infect both placental and marsupial mammals, raise concerns about ongoing global efforts

to control and eliminate rubella. Such efforts may have to reassess the specificity of diagnostic tests, consider monitoring bats and other wildlife for rubella and rubella-like viruses, evaluate cross protection of rubella virus vaccines for novel members of the *Matonaviridae*, and potentially vaccinate populations in "interface" environments where zoonoses most often occur.

INTRODUCTION

Rubella, first described in 1814 (117) is an acute, contagious human infectious disease typically characterized by rash, low-grade fever, adenopathy, and conjunctivitis (99). Research during the 1940s to 1960s revealed that rubella (at the time also referred to as "German measles") contracted during the first-trimester of pregnancy is directly associated with severe congenital birth defects, miscarriage, and stillbirth. Rubella virus (RuV), currently the sole member of the riboviriad family *Matonaviridae* (genus *Rubivirus*), is the etiologic agent of rubella and causes fetal pathology after transplacental transmission. Extensive rubella epidemics have occurred worldwide due to the high transmissibility of RuV ($R0$ = 3.5–7.8)(48). Safe, efficacious, live-attenuated RuV vaccines, including the measles/mumps/rubella (MMR) vaccine, are now deployed worldwide and have successfully decreased global rubella incidence. However, ≈100,000 cases of congenital rubella syndrome still occur annually (99) and RuV can persist in the human body for years (41). Furthermore, RuV infection of adults is generally underreported, as 30–50% of adult cases are subclinical (66). High priority areas for rubella vaccination include countries in Sub-Saharan Africa, where rubella virus circulates widely and primarily infects young children. RuV eradication is considered rapidly achievable because of the effectiveness of available vaccines and the lack of known animal reservoirs.

Results and Discussion

Here we report the discovery of ruhugu virus (RuhV) and rustrela virus (RusV), the first close relatives of RuV. RuhV was found in oral swabs of 50% of apparently healthy cyclops leaf-nosed bats (Hipposideridae: *Hipposideros cyclops* Temminck, 1853) inhabiting tree roosts (hollow cavities in trees) in Kibale National Park, western Uganda. RusV was found in tissues from an encephalitic donkey (*Equus asinus* Linnaeus, 1758) and a tree-kangaroo (*Dendrolagus bennettianus* de Vis, 1886) in a zoo in Germany (Figure 27). Infected bats in Uganda all appeared healthy at the time of sampling, but the zoo animals in Germany had died or were humanely euthanized after rapid onset of unresponsiveness, convulsions, and other severe neurologic signs.

In Uganda, bat roosts ranged in size from 1–8 bats, and all bats in 5 roosts were captured and sampled. Using molecular and metagenomic methods, RuhV RNA was detected in 5/9 (55.6%) males and 5/11 (45.5%) females in 4 of 5 (80%) of roosts (prevalence 50%; 95% CI 29.9–70.1%). This high prevalence and frequency of positive roosts suggest that apparently healthy cyclops leaf-nosed bats are reservoir hosts, rather than incidental hosts, of RuhV. In Germany, both the donkey and the Bennett's tree-kangaroo were submitted for pathology and diagnostic testing after succumbing to severe neurologic disease, and histopathological analysis revealed a non-purulent meningoencephalitis in both animals.

Using molecular and metagenomic methods, the presence of RusV was confirmed in brain samples of both animals and in the liver of the donkey (Data Table 11). The acute and severe nature of RusV infection in these two very different animals at the same

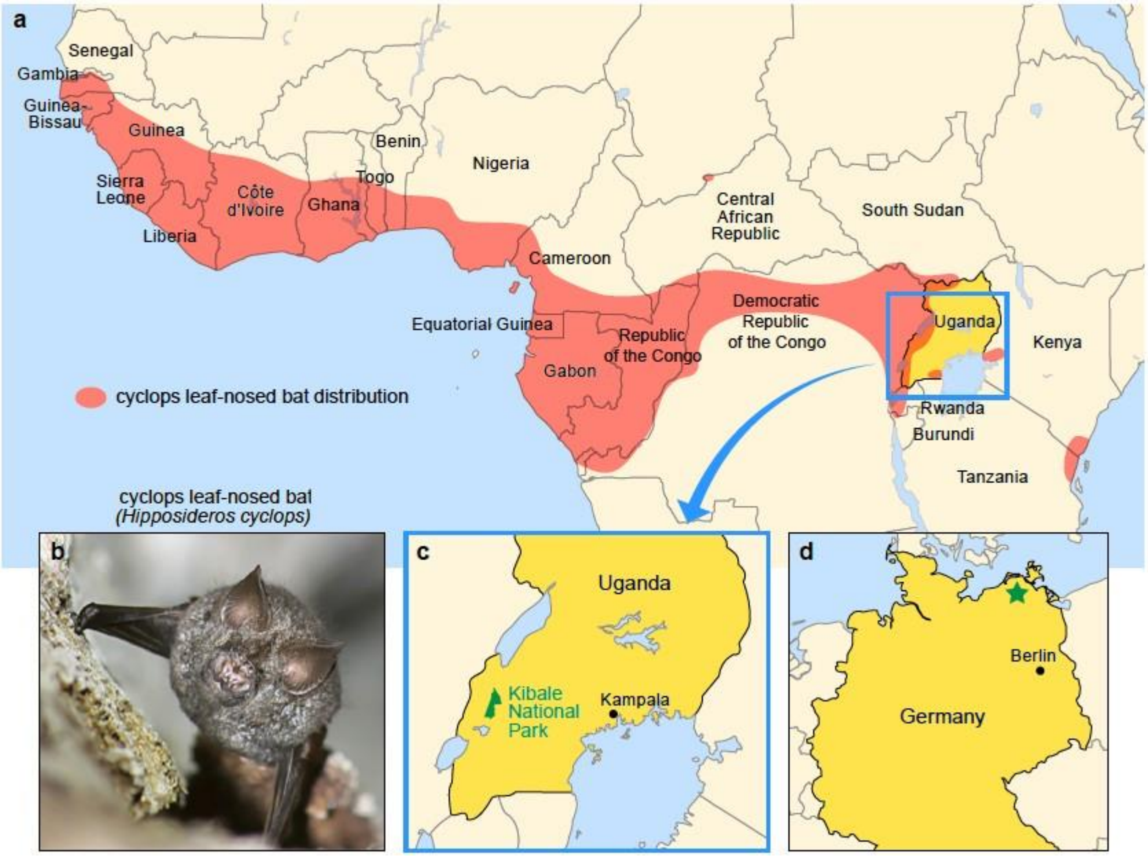


Figure 27. Geographic locations of rubiviruses and their hosts. a) Summary map of estimated cyclops leaf-nosed bat distribution in Africa (red) and Uganda (blue box). b) Male cyclops leaf-nosed bat in Kibale National Park, Uganda (photo credit: Caley Johnson). c) Location of bat sample collection and discovery of ruhugu virus (Kibale National Park, Uganda, green). d) Location of zoo animals and discovery of rustrela virus in Germany (southern Baltic Sea region, green star).

location suggests that they were not natural hosts, but rather spill-over hosts, of RusV.

The genome organizations of RuV, RuhV and RusV are identical (Figure 28a), consisting of two large open frames (ORFs). In RuV, the longer (5′) of the two ORFs encodes nsPP/p200, a nonstructural polyprotein that is post-translationally cleaved into two proteins, p150 (methyltransferase and protease) and p90 (helicase and RNA-directed RNA polymerase). RuhV nsPP/p200 is 6,146 nt long and shares 56.6% nucleotide and 58.6% amino acid identity with RuV nsPP/p200. RusV nsPP/p200 is 6,158 nt long and shares 52.5% nucleotide and 45.2% amino acid identity with RuV nsPP/p200. In RuV, the smaller (3′) of the two ORFs is transcribed as a subgenomic RNA that encodes sPP, the structural polyprotein. RuV sPP is processed into 3 structural proteins: CP (capsid protein) and E1 and E2, the two envelope proteins (2). RuhV sPP is 3,296 nt long and shares 53.8% nucleotide and 51.5% amino acid identity with RuV sPP, whereas RusV sPP is 3,129 nt long and shares 52.5% nucleotide and 40.1% amino acid identity with RuV sPP. Genetic similarities among the viruses vary across the sequences of endpoint protein products and are generally lowest in a hyper-conserved region within the Y domain of P150 (Table 10, Figure 28).

RuhV (named for Ruteete Subcounty, Uganda, and the Tooro word for insectivorous bat, *obuhuguhugu*) is an outgroup to all known RuV genotypes (Figure 29b). RusV (named for its rubivirus-like genome and the Strela Sound of the Baltic Sea in Germany) is a close outgroup to the RuV/RuhV clade of viruses (Figure 29b). This topology is consistent with the higher similarity of RuhV to RuV in each of the five mature peptides of the protein-coding viral genome (Extended Data Table 1, Figure 29). A phylogeny including outgroups from other viral families (Figure 30) shows that the

Table 10. Genetic similarities among ruhugu, rustrela and rubella viruses

	Nucleotide position (5'-3')		*Amino acid residues*		*Amino acid identity (%)*			*GC content (%)*		
Genome feature	*RuhV*	*RusV*	*RuhV*	*RusV*	*RuhV*[a]	*RusV*[a]	*RuhV–RusV*	*RuhV*	*RusV*	*RuV*[b]
Complete genome	1–9622	1–9755	N/A	N/A	56.2	43.0	36.9	62.1	70.4	69.6
Non-structural polyprotein	44–6190	59–6217	2049	2053	58.6	45.2	46.6	61.1	70.1	70.0
p150 protease	44–3751	59–3775	1236	1239	48.2	33.3	34.4	61.4	72.0	71.4
p90 replication complex	3752–6190	3776–6217	813	813	75.7	65.5	66.4	60.8	67.7	67.8
Structural polyprotein[c]	6266–9562	≈6542–9672	1099	1042	51.5	40.1	20.0	64.3	71.5	69.4
Capsid protein[c]	6266–7216	≈6542–7269	317	242	52.5	39.4	20.6	60.3	75.1	73.1
E2 envelope protein	7217–8101	7270–8211	295	314	43.2	26.4	19.1	67.1	72.2	71.0
E1 envelope protein	8102–9562	8212–9672	487	487	55.8	50.9	20.2	65.5	69.2	66.3

[a] Ruhugu and rustrela virus amino acid identities compared to rubella virus strain F-Therien (RefSeq accession numberNC_001545).

[b] GC content for rubella virus strain F-Therien (RefSeq accession number NC_001545).

[c] The rustrela virus genome contains two gaps: p150 ≈413 nt and capsid≈102 nt.

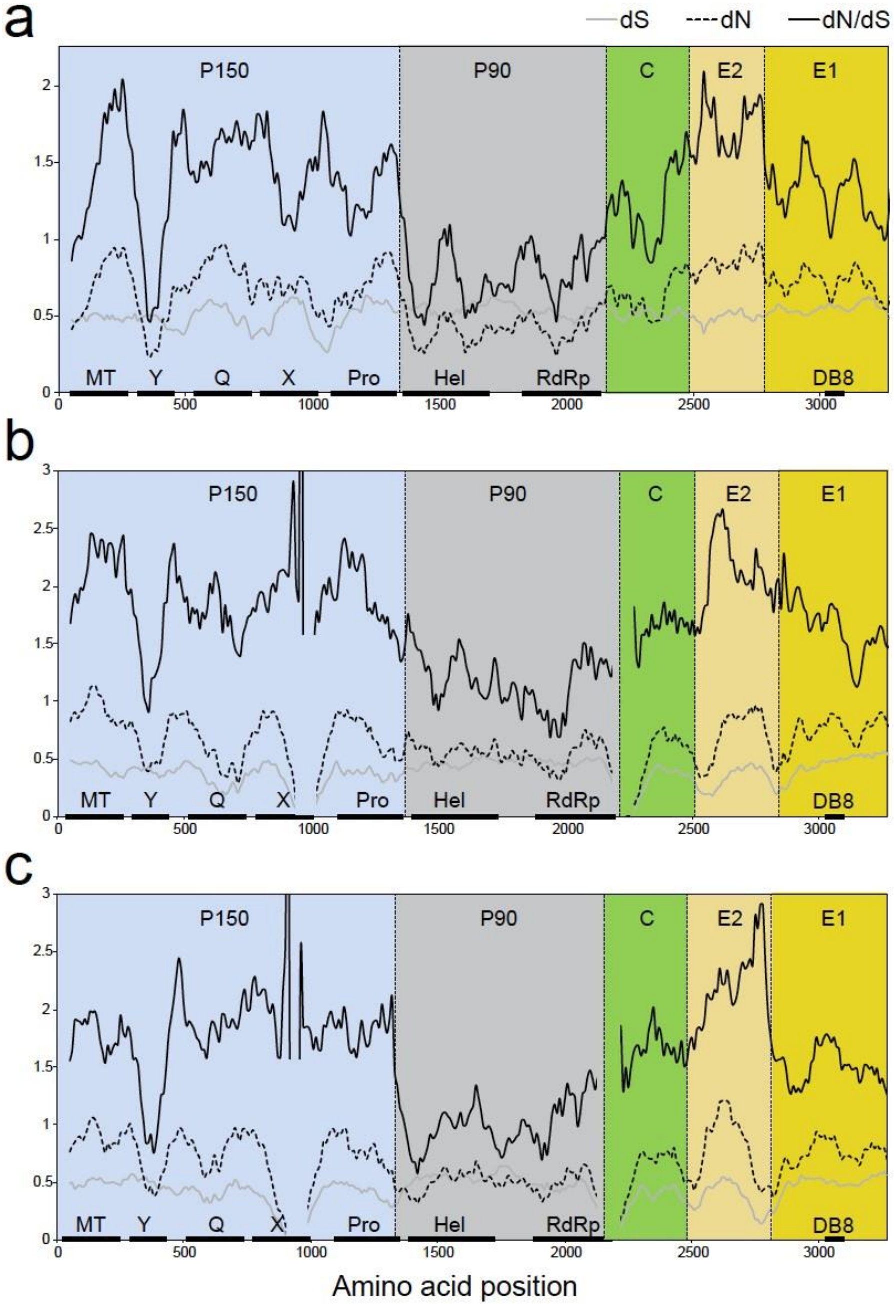


Figure 28. Average substitution rates at non-synonymous (dN; dashed lines) and synonymous (dS; grey lines) sites, and the ratio of dN/dS (solid lines), for aligned, concatenated amino acid sequences comparing RuV and RuhV (a), RuV and RusV (b) and RuhV and RusV (c) using sliding window analysis (100 residue window size, 10 residue steps). Protein domains are labeled on the X axes: MT=methyltransferase; Y, Q, and X=domains of unknown function; Pro=protease; Hel=helicase; RdRp=RNA-directed RNA polymerase; DB8=disulfide bond 8 epitope.

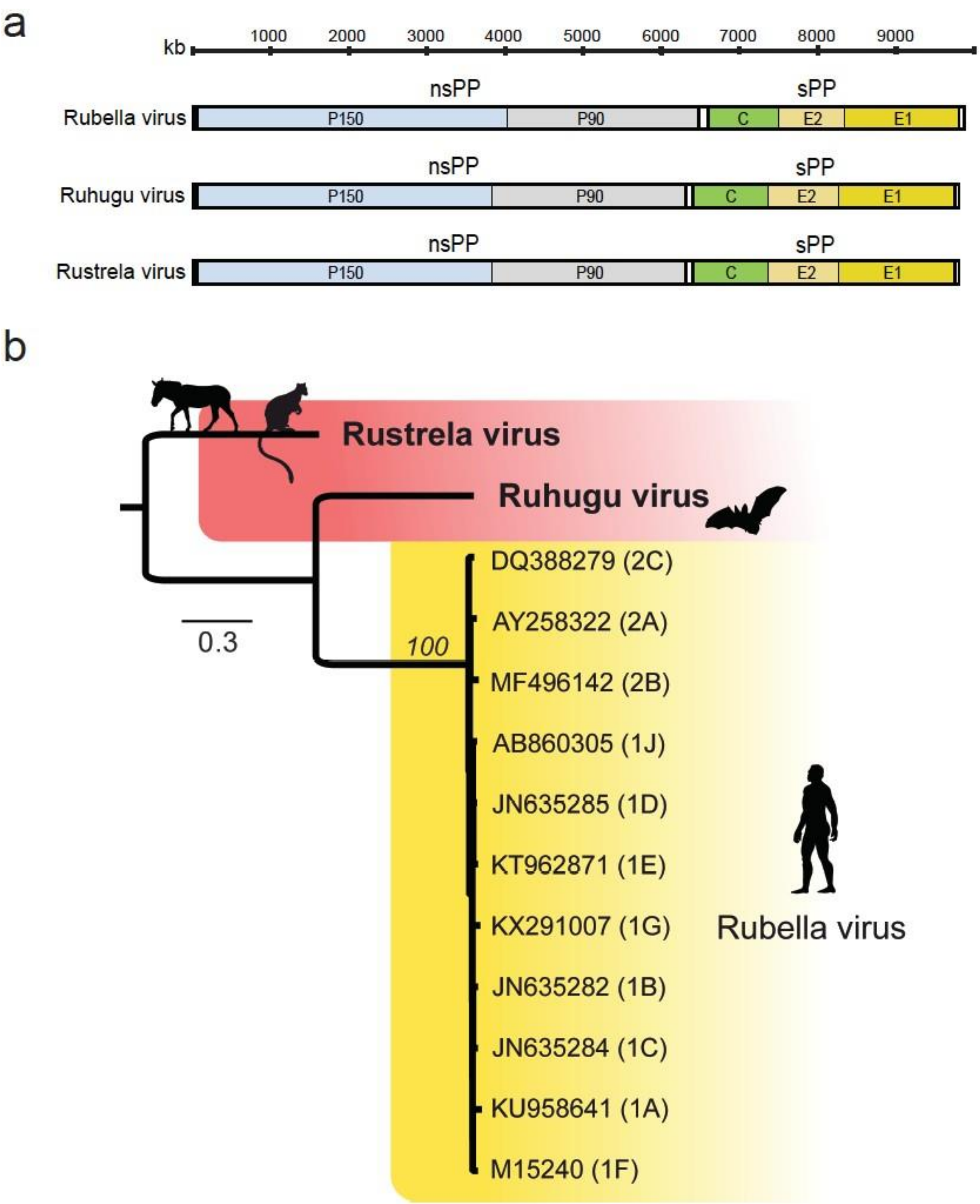


Figure 29. Evolutionary relationships among rubiviruses. a) Comparative genome architecture of RuV, RuhV, and RusV. b) Maximum likelihood phylogenetic tree of rubella virus genotypes 1A–1J and 2A–2C, ruhugu virus, and rustrela virus. Numbers beside nodes indicate bootstrap values (percent; only values for major branches are shown); the scale bar indicates amino acid substitutions per site.

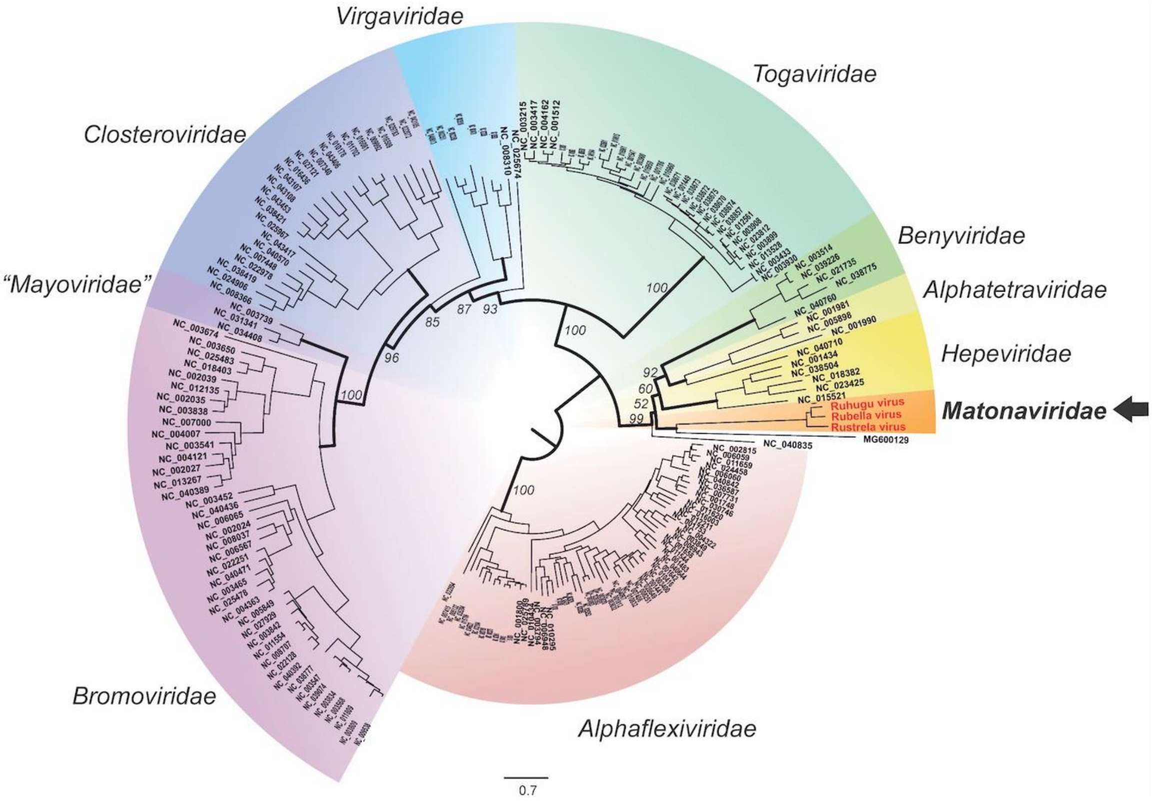


Figure 30. Phylogenetic tree of members of the genus *Rubivirus* (rubella, ruhugu, and rustrela viruses, provisionally; highlighted in red) included in the family *Matonaviridae* and selected viruses from related families in the alphavirus supergroup ("*Alsuviricetes*") using a conserved region of the viral RNA-directed RNA polymerase (IQ-TREE v1.6.12; VT+F+R8; 250,000 UF bootstraps). Viruses indicated by lack of shading are currently unclassified.

matonavirids are only distantly related to viruses of any other family, and that no congenerics of RuV have previously been identified.

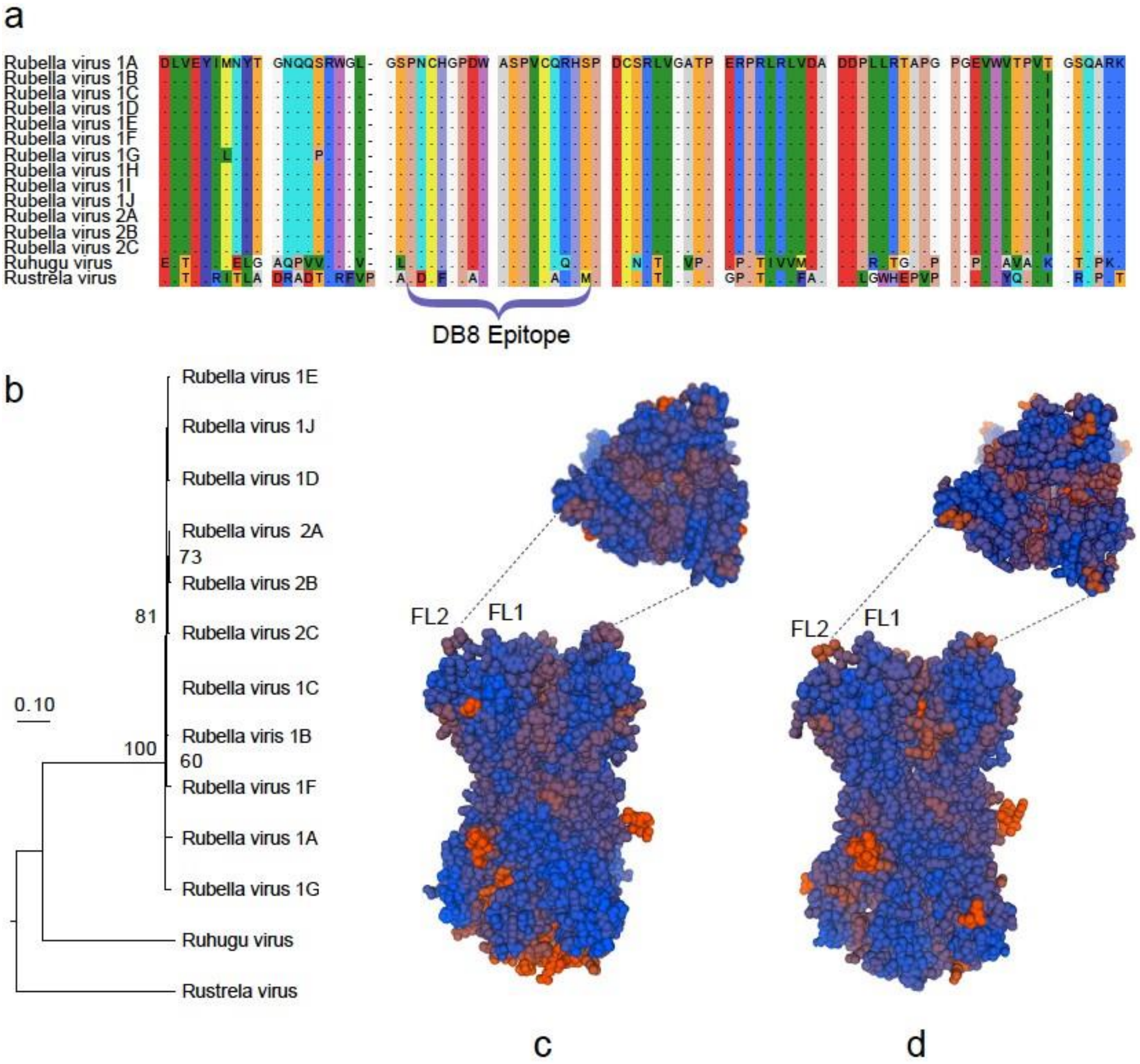
E1, a receptor-binding, class II fusion protein (44), is the primary RuV antigen E1contains a well-characterized, immune-reactive region (amino acid residue positions 202–283) that includes dominant linear B- and T-cell epitopes (Figure 31a) An important neutralizing epitope maps to amino acid positions 223–239 of this region at E1 disulfide bond 8 (DB8). In humans, antibodies to the DB8 epitope appear to protect against congenital rubella syndrome following intrauterine infection (44; 149), whereas a lack of antibodies to this epitope in newborns is associated with congenital rubella syndrome, which can lead to persistent RuV infection for up to 13 months (27). The mechanism of neutralization appears to involve blocking of E1 trimerization, which is necessary for virion fusion with the host cell plasma membrane. Surprisingly, the RuhV DB8 epitope differs from that of RuV by only 1 amino acid residue (R237Q) near the C-terminus of the 17-residue linear epitope (Figure 31a), despite greater amino acid-level divergence across E1 (Figure 31b). By contrast, RusV differs from RuV at 5 amino acid residues within the same region (Figure 31a).

The modeled secondary structures of trimeric RuhV and RusV E1 are homologous to RuV E1 with 100% confidence, and homology-based modeling of RuhVE1 quaternary structure predicts with high confidence that RuhV and RusV E1 proteins form homotrimers in the post-fusion state (44) (Figure 31c and 31d). RuhV E1 fusion loops (FL1:residues 87–92; FL2: residues 130–136) are predicted to support the unusual metal ioncomplex necessary for E1-mediated RuV membrane fusion due to presence of Asn 87and Asp 135 (homologous to RuV Asn 88 and Asp 136, respectively, Figure 31c . By contrast, FL2 of RusV is predicted to be less similar to RuV, due to two amino acid replacements, P134A and T135A, the latter being a transition from a polar to a non-polar residue (Figure 31d). Across the RuV, RuhV, and RusV genomes, regions of markedly stabilizing selection are evident immediately upstream of the putative methyltransferase domain of P150, in the RdRp domain of P90, and proximal to the aforementioned DB8 epitope of E1for RuhV and RuV (Figure 29).

Whether RuhV infects animals other than cyclops leaf-nosed bats remains unknown. However, the discovery of RusV in an acutely ill donkey and Bennett's tree-kangaroo indicates that rubella-like viruses can infect animals as divergent as placental and marsupial mammals. This observation also suggests that RusV in these hosts was likely transmitted from an animal of another species, which, given the likely bat reservoir host of RuhV, may be a bat or another small mammal. An important step in determining the potential risk that RuhV, RusV and related viruses might pose to global rubella eradication efforts will be to assess their ability to infect humans, bats, and non-chiropteran animals. If rubiviruses capable of zoonotic transmission exist, such viruses could conceivably fill the niche left empty after RuV eradication in humans. Surveys of cyclops leaf-nosed bats and other mammals for rubiviruses are therefore exigent.

Cyclops leaf-nosed bats are insectivorous microbats primarily found in lowland rainforests from Senegal to Tanzania, but they also inhabit coastal, montane, and swamp forests and disturbed and agricultural landscapes where they could contact people (Figure 27). Surveys of these forest-dwelling bats would not be trivial because of their wide geographic range, spatially distributed populations, and small roost sizes. Moreover, the near identity of the RuV and RuhV DB8 epitope suggests that serologic methods to

Figure 31. Comparisons of the rubella, ruhugu, and rustrela viruses E1 envelope glycoproteins. a) Amino acid alignment of an 85 amino acid residue-long immunoreactive region of E1 for RuhV, RusV, and 13 RuV genotypes. b) Maximum likelihood phylogeny of complete RuhV and RusV E1 amino acid sequences with representatives of all RuV genotypes, rooted based on the position of outgroup taxa. Numbers beside nodes represent bootstrap values (%); only values ≥ 50% are shown. The scale bar indicates amino acid substitutions per site. c) Homology-based model of RuhV E1 homotrimer quaternary structure in post-fusion state (with inset, top-down view of fusion surface). Global model quality estimates (QMEAN) indicates a good model fit for the structure as a whole relative to the crystal structure of the RuV E1 in post-fusion form (Protein Data Bank biological assembly 4adg.1). Local structural similarity between RuV and RuhV (estimated as QMEAN) is depicted by color, ranging from orange (low) to blue (high). d) Homology-based model of RusV E1 homotrimer quaternary structure in post-fusion state, as described above for RuhV. Key differences are seen in the modeled fusion loops (FL1, FL2): RuhV FL1 and FL2 are highly similar to those of RuV, whereas RusV FL2 residues differ considerably from those of RuV FL2.


a
Rubella virus 1A
Rubella virus 1B
Rubella virus 1C
Rubella virus 1D
Rubella virus 1E
Rubella virus 1F
Rubella virus 1G
Rubella virus 1H
Rubella virus 1I
Rubella virus 1J
Rubella virus 2A
Rubella virus 2B
Rubella virus 2C
Ruhugu virus
Rustrela virus
DLVEYIMNYT GNQQSRWGL- GSPNCHGPDW ASPVCQRHSP DCSRLVGATP ERPRLRLVDA DDPLLRTAPG PGEVWVTPVT GSQARK
L P
E.T...ELG AQPVV...V- .L........ ......Q... ..N.T..VP. .P.TIVVM.. ...R.TG..P ..P..AVA.K .T.PK.
..T..RITLA DRADT.RFVP .A.D.F..A. .....A..M. ....T..... GP.T...FA. ..LGWHEPVP ......YQ.I .R.P.T
DB8 Epitope
b
Rubella virus 1E
Rubella virus 1J
Rubella virus 1D
Rubella virus 2A
73
Rubella virus 2B
81
Rubella virus 2C
Rubella virus 1C
0.10
Rubella viris 1B
100
60
Rubella virus 1F
Rubella virus 1A
Rubella virus 1G
Ruhugu virus
Rustrela virus
FL2
FL1
FL2
FL1
c
d

diagnose rubivirus infection may prove non-specific in bats and other animals, including humans. Implicating RuhV as a zoonotic agent is currently speculative, but since bats are now widely recognized as hosts to many zoonotic viruses, we advise that this scenario ought not to be dismissed. The ability of RusV to infect both placental and marsupial mammals and to cause clinical disease resembling severe forms of rubella in humans (14; 56) reinforces such a precautionary stance. The World Health Organization's (WHO's) Global Measles and Rubella Strategic Plan aims to control or eliminate rubella and congenital rubella syndrome in 5/6 WHO regions by 2020 (65).

For this effort to succeed, accurate data on the host range and epidemiology of RuV, RuhV, RusV, and other as-yet undiscovered rubiviruses is paramount. Molecular and serologic tests that accurately distinguish RuV from its relatives may therefore be necessary. The discoveries of RuhV and RusV strongly suggest that other relatives of RuV exist in bats or other reservoir hosts and are able to infect mammals across wide taxonomic distances. Thus, our findings portend challenges for rubella eradication but also open doors to comparative studies of RuV that were heretofore not possible.

METHODS

Animal sampling and pathology

In Uganda, cyclops leaf-nosed bats were captured and released in Kibale National Park in June and July of 2017. Kibale is a 795-km^2 mid-altitude semideciduous forest park (0°13'–0°41"N, 30°19'–30°31"E) within the Albertine Rift, which is a region of exceptional biodiversity (Figure 28c). Bats were caught in mist nets (Avinet Portland, ME, USA) set in their flight path as they exited tree roosts at dusk and were kept in cloth bags until processing. Oral swabs were collected from each bat using sterile rayonpolyester-tipped swabs and preserved in 500 µl of TRI Reagent (Zymo Research, Irvine, CA, USA). Swabs were frozen at -20 °C within 3 h of sample collection and transported on ice for storage at -80 °C prior to further analyses. Animal collection and handling protocols were approved by the Uganda Wildlife Authority, the Uganda National Council for Science and Technology, and the University of Wisconsin-Madison Animal Care and Use Committee. Samples were shipped in accordance with international law and imported under PHS permit number 2017-07-103 issued by the US Centers for Disease Control and Prevention, Atlanta, GA, USA.

In Germany, a Bennett's tree-kangaroo and a donkey were submitted for general diagnostics and pathology in July 2018 and March 2019, respectively, after presenting with acute and severe neurologic signs. Both animals lived in the same small zoo close to the Baltic Sea coast in the northeast of Germany (Figure 27d). Histopathological analysis revealed a non-purulent meningoencephalitis in both animals and diffuse fatty liver cell degeneration and hemosiderosis in the donkey. Standard diagnostics, including exclusion diagnostics (e.g. for rabies virus, bornaviruses, West Nile virus, herpesviruses, *Listeria*, *Salmonella* or *Toxoplasma*) did not detect any specific viral, bacterial, or protozoan pathogens.

Metagenomic, molecular and bioinformatic analyses

RNA was purified from bat oral swabs using the Direct-zol RNA MicroPrep kit (Zymo Research, Irvine, CA, USA). Illumina RNA TruSeq libraries were then prepared, evaluated for quality, multiplexed, and sequenced with Illumina NextSeq 500 v2 chemistry and using 2x150-bp read lengths (Illumina, San Diego, CA, USA). After RuhV was first identified using the VirusSeeker virus discovery pipeline (209), deeper sequencing of two individual bat swab libraries was performed on an Illumina MiSeq sequencer using v3 chemistry and 2x300-bp read lengths. The cyclops leaf-nosed bat genome was removed *in silico* by mapping reads to assembly PVLB01000001 using bbmap (19) and discarding mapped reads. *De novo* assembly was then performed using MetaSPAdes following quality control and removal of non-viral reads using fastqc, bbmap, and bbduk (6; 19), reads were mapped back to contigs for validation, and sequenced relatives were determined by DIAMOND using the BlastX algorithm (3; 18). Results were visualized using MEGAN 6 (81). Detailed analyses of contigs and reads were performed with CLC Genomics Workbench V12 (QIAGEN Bioinformatics; Redwood City, CA).

Bennett's tree-kangaroo and donkey tissues were processed using published methods for metagenomic pathogen detection (199). Briefly, tissues were first disrupted using the Covaris cryoPREP system (Covaris, Woburn, MA, USA) and subsequently lysed in buffer AL (Qiagen, Hilden, Germany), followed by addition of TRIzol Reagent (Life Technologies, Darmstadt, Germany). After centrifugation, the aqueous phase was then transferred to RNeasy Mini Kit columns (Qiagen, Hilden, Germany) and processed according to the manufacturer's instructions. Total RNA from the cerebrum of the donkey and the Bennett's tree-kangaroo was used for library preparation (199) and sequencing on an Ion Torrent S5XL instrument with a 530 chip (Thermo Fischer Scientific, Waltham, MA, USA). Subsequently, the RIEMS software pipeline (157) was used for initial taxonomic assignment of reads.

After RusV was identified in the donkey using the methods described above, deeper sequencing was performed on the Ion Torrent S5XL and an Illumina MiSeq. The donkey genome was removed *in silico* by mapping reads to assembly ASM130575v1 using BWA (107), and unmapped reads were filtered and retained. Read data quality trimming, adapter removal, and quality control were performed using the 454 software suite (v3.0; Roche) and FastQC (6). *De novo* assembly was performed using SPAdes (132).

RusV-specific contigs were then identified by DIAMOND using the BlastX algorithm (3; 18) followed by an iterative mapping and assembly approach using the 454 software suite (v3.0; Roche), SPAdes, and Bowtie2 (101) for contig extension and verification. Results were visualized using Geneious (v11.1.5, Biomatters Ltd, Auckland, New Zealand). ORFs were identified by ORF Finder (implemented in Geneious). Conserved elements were identified by translated amino acid sequence alignment to RuV genomes using MUSCLE and subsequent annotation of p150, p90, and E1. The 5′end of E2 was identified by similar hydrophobicity and sequence pattern of the E2 signal peptide of RuV (77) located at the C terminus of CP using ProtScale (179) (window size: 3; relative weight for window edges:100 %; weight variation model: linear).

The presence of RusV was confirmed in formalin-fixed paraffin-embedded (FFPE) brain tissue samples from both the donkey and the Bennett's tree-kangaroo, and in the liver of the donkey, using an original RT-qPCR (Table 11). Total RNA from FFPE tissues was extracted using a combination of the Covaris truXTRAC FFPE total NA Kit and the Agentcourt RNAdvance Tissue Kit (Beckman Coulter, Indianapolis, IN, USA). RT-qPCR was then performed using the SensiFAST Probe No-ROX One-Step Kit (Bioline, Luckenwalde, Germany) with forward-primer (1063–1082, CGAGCGTGTCTACAAGTTCA), reverse-primer (1210–1228, GACCATGATGTTGGCGAGG), and 5′ nuclease probe (1152–1171, FAM-CCGAGGAGGACGCCCTGTGC-BHQ-1) on a Bio-Rad CFX96 qPCR Instrument (Bio-Rad, Hercules, CA, USA). Primer and probe specificity were verified by BLASTN36 *in silico* analyses and Sanger sequencing (Eurofins Genomics Germany GmbH, Ebersberg, Germany). β-actin was used as an internal inhibition control.

Phylogenetic analyses and protein functional domain predictions

Phylogenetic trees of aligned amino acid sequences were inferred using IQ-TREE software (v. 1.6.12) (131), with automated model selection (JTTDCMut+F+R3) and 250,000-500,000 ultrafast bootstrap replicates (76).

RuhV and RusV protein functional domain prediction and annotation were performed using the InterPro webserver , and the confidence of E1 structural homology was estimated using Phyre2 (89). Homology modeling of the quaternary structure of the post-fusion E1 homotrimer (Figures 31c and 31d) was performed using the SWISS-MODEL (184) workspace with model view by PV (15) and residue color corresponding to QMEAN score (12), with 53 c-terminal residues of E1 (representing the stem and transmembrane segment of the E1 linear peptide) removed prior to homotrimer modeling (91). Patterns of selection across the RuV, RuhV, and RusV genomes were examined using SNAP 2.1.1 (44).

Table 11. Rustrela virus distribution in animal tissues assessed by RT-qPCR

	Ct value	
Source	*Rustrela virus*	*ß-Actin*
Donkey		
Cerebrum (I)[a]	22.9	20.5
Cerebrum (II)[b]	29.2	23.7
Cerebrum (III)[b]	29.5	23.8
Brain stem[b]	30.5	25.6
Cerebellum[b]	30.6	24.9
Medulla oblongata[b]	33.9	29.1
Liver (I)[a]	35.9	21.7
Liver (II)[b]	-	27.5
Kidney[b]	-	23.2
Small intestine[b]	-	22.6
Bennett's tree-kangoroo		
Cerebrum[a]	30.2	31.1
Organ pool (I)[a]	35.5	28.7
Organ pool (II)[a]	-	30.8

[a] unfixed tissues

[b] formalin-fixed paraffin-embedded tissues

ACKNOWLEDGEMENTS AND FUNDING

We thank David Hyeroba (Makerere University Kampala, Uganda), Katusabe Swaibu (Makerere University Biological Field Station), and Jessica Carag (University of Wisconsin-Madison, Madison, WI, USA) for assistance in the field. We thank Laura Bollinger and Jiro Wada (NIH/NIAID Integrated Research Facility at Fort Detrick, Frederick, MD, USA) for their help improving the manuscript and figures. We thank Gregory K. Rice (Leidos and Biological Defense Research Directorate, Naval Medical Research Center – Frederick, MD, USA) for advice and assistance with bioinformatics scripts and Patrick Zitzow and Jenny Lorke (Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Greifswald—Insel Riems, Germany) for invaluable technical assistance.

This work was supported in part through the US National Institute of Allergy and Infectious Diseases (NIAID) Virology Training Grant No. T32 AI078985 (to University of Wisconsin-Madison), Battelle Memorial Institute's prime contract with NIAID under Contract No. HHSN272200700016I (to J.H.K.) and GEIS P0032_18_NM_05 (to K.A.B-L.), and by the Federal Ministry of Education and Research within the research consortium "ZooBoCo" (Grant No. 01KI1722A). Additional support was provided through the University of Wisconsin-Madison John D. MacArthur Professorship Chair, Institute for Regional and International Studies, and Global Health Institute.

Data availability. Sequence data that support the findings of this study have been deposited in GenBank with the accession numbers MN547623 and MN552442.

CHAPTER 6: Discussion

DISSERTATION SUMMARY

The goal of the work described in this thesis was to examine viral persistence in bat populations and, ultimately, inform ongoing biosurveillance efforts by utilizing a whole genome sequencing approach for pathogens of public health concern. Chapter Two discussed the validation of a hybridization-based target enrichment technique for viruses of biosurveillance concern using contrived samples. In Chapter Three, this target enrichment approach was applied side-by-side with unbiased sequencing in a longitudinal study that characterized the virome of a captive colony of lesser dawn bats that roost in Singapore. In Chapter Four, we identified a cross-family recombinant virus named Rousettus bat coronavirus GCCDC1, which was detected in the captive megabat colony. In Chapter Five, we reported the discovery of two novel rubella-like viruses. These contributions are evidence of the power of genomics and computational biology for biosurveillance and discovery of new viruses.

WET LAB AND COMPUTATIONAL METHODS DEVELOPMENT

Our research questions required us to perform methods development in the wet lab and for bioinformatics analysis. In an effort to minimize the large amount of bacterial and host “noise” that drowns out viral reads in metagenomic samples, we needed to optimize a hybridization-based target enrichment technique to suit the needs for this project and typical biosurveillance samples. We also needed to develop a new pipeline to efficiently and reliably identify divergent viral sequences amidst millions of reads per individual sample. To address our hypothesis of long-term persistence versus a decline in viral abundance over time, we found it necessary to develop a normalization technique that could be broadly applicable to metagenomic studies.

Hybridization-based target enrichment enhances viral detection

Through a collaboration with Illumina, a panel of biosurveillance-related probes was developed to complement existing panels designed to detect respiratory viruses and filoviruses (Chapter Two, Table 3). The probe sets were tested using bat guano or serum samples spiked with several known strains of viruses. Once the performance of the probe set was validated in a controlled set of samples, environmental samples (bat swabs) were tested. Results from shotgun and target-enrichment preparation were compared to evaluate the performance of the enrichment method in this sample type. We chose to evaluate enrichment by comparing the depth and breadth of coverage of viruses only detected in both sets for fair comparison, although we did not have a "ground truth" because the samples were environmental unknowns. We concluded that probe-based targeted enrichment is highly effective for targeted viral nucleic acid with a tolerance as low as 70% nucleotide identity (140). However, this approach may not be appropriate when the viruses in a sample are largely unknown. For example, in swabs collected from the understudied species *E. spelaea*, many of the detected viruses were previously unknown and significantly divergent from targeted genomes. Therefore, unbiased shotgun sequencing data was crucial for the comprehensive characterization of the bat virome. A future direction for this approach could be to expand the breadth of targeted viruses in the probe set to include genomes of newly discovered filoviruses and bat-associated viruses.

Development of a pipeline for data analysis

To execute bioinformatic data analysis following a massive sampling and sequencing effort, our team developed an efficient, modular pipeline that could use cluster computing to handle massive amounts of data generated by Illumina's NextSeq500 platform. Our goal was to automate quality control, including trimming and host sequence removal, for hundreds of samples while enabling manual analysis of results for individual samples. To achieve this goal, previously published and publicly available tools were combined in a modular format with in-house scripts. The pipeline utilized FastQC for quality evaluation, bbmap for host removal, clumpify and seqtk for duplicate sequence removal, bbduk for quality trimming, metaSPAdes for *de novo* assembly of trimmed reads, Meta-QUAST Paired End for assembly validation, bbmap to map trimmed reads to the contig assembly, and Diamond for blastX to determine similarity to known viruses (1; 6; 19; 132). MEGAN6 was used to visualize results, and CLC Workbench v11 was used to manually analyze data (81). An outline of the resultant pipeline is illustrated in Figure 32A; this pipeline is currently referred to as MetaDetector. This approach is modular and parameters for quality control or assembly can be modified to suit the input data. MetaDetector has since been adopted for use in projects beyond bat virome characterization in the Genomics & Bioinformatics department. This approach has also been reconstructed by our collaborators for the discovery of Rustrela virus. As an orthogonal validation for taxonomic classifications that were determined using our in-house pipeline, a newly published and publicly available discovery pipeline, VirusSeeker, was used to query all samples (209).

The VirusSeeker-Discovery pipeline was published in 2017 and was used to make taxonomic classifications of viruses in Chapters 3-5. As shown in Figure 32B, the pipeline trims raw sequences for quality and determines unique, unmapped reads by Cluster Database at High Identity with Tolerance (CD-HIT (57)). Reads are then assembled into contigs, if possible. Resultant contigs and outliers are again simplified by CD-HIT, then filtered for quality before analysis with BLAST from VS-Virome. The output of VirusSeeker was used for a variety of analyses, including determination of the family or genus of viral sequences, length of genomic segments, and homologs of encoded proteins. The novel feature of

VirusSeeker is the use of paired-end reads in a stitching script to join reads together to enhance sensitivity for the detection of highly divergent viruses (53). The advantage of joining two reads together to create longer reads may seem obvious; however, this is a method rarely used in popular bioinformatic pipelines. The original use of VirusSeeker in our department was for this project in an attempt to analyze the massive dataset produced by a NextSeq500 sequencing run on our cluster. This pipeline has been so informative for viral discovery and virome characterization that it has been adopted for routine use in our department and we have trained visiting scientists how to use it. We generated scripts that transfer output from VirusSeeker to make it more quantitative in nature for the purpose of downstream analyses. These scripts were developed before MetaDetector and we provide the scripts to subordinate labs that we train.

The use of orthogonal bioinformatic methods provided more confidence to our taxonomic assignments as we sought the most comprehensive solution to dealing with the large amount of data. These methods were used to analyze data from the longitudinal bat study (Chapter 2) and in the discovery of Ruhugu virus (Chapter 5). Limitations of VirusSeeker's assembler and binning approach result in fewer classified reads; however, the results, while underestimated, seldom contained false classifications. While neither method is perfect, the ability to compare VirusSeeker results to classified contigs assembled by MetaSPAdes is a solution that best utilizes the technologies available. In other words, identification of viruses by both methods provided confidence in our results.

In the case of novel viruses, taxonomic assignments followed a consistent pattern that indicated a need for manual analysis, which is not unusual in our experience and is not specific to viruses in bats; rather it is a function of virus divergence and database insufficiency. Results indicating multiple reads or contigs assigned to two or more closely-related viral species generally indicated the presence of a novel species and were subsequently combined after manual analysis. This thesis included understudied bats, lesser dawn bats and cyclops roundleaf bats, for which much of the virome was previously unknown. As such, there were numerous cases in which the results from VirusSeeker and our in-house pipeline were combined for further analysis in CLC Workbench.

Development of a normalization technique

In Chapter 3, normalized viral abundance was utilized to make colony-level observations of the captive colony of lesser dawn bats. A normalization technique was required for comparison among individual bat swabs. While there is no standard, community-accepted procedure for normalization across metagenomic samples, there has been considerable discussion in the literature about best practices for identifying biases that could be introduced during metagenomic analyses (62; 118). Normalization of individual taxa was performed using the following formula:

$$Normalized\ reads = \frac{\frac{\#\ reads\ for\ one\ viral\ taxa\ in\ one\ sample}{total\ \#\ QC\ reads\ in\ that\ sample}}{virus\ genome\ length\ (nt)} \times 1000$$

Results were filtered to exclude the possibility of sample carryover with a minimum cutoff of one stitched read. We determined that this normalization and cutoff approach was appropriate following evaluation of possible confounders and carryover among samples via analysis of variance and unsupervised clustering; these methods revealed no trends among potentially confounding elements. Unsupervised clustering was initially evaluated as a tool for analysis but was abandoned for normalization in favor of a less esoteric approach - analysis of variance. Virus genome length in nucleotides was selected for normalization to account for the

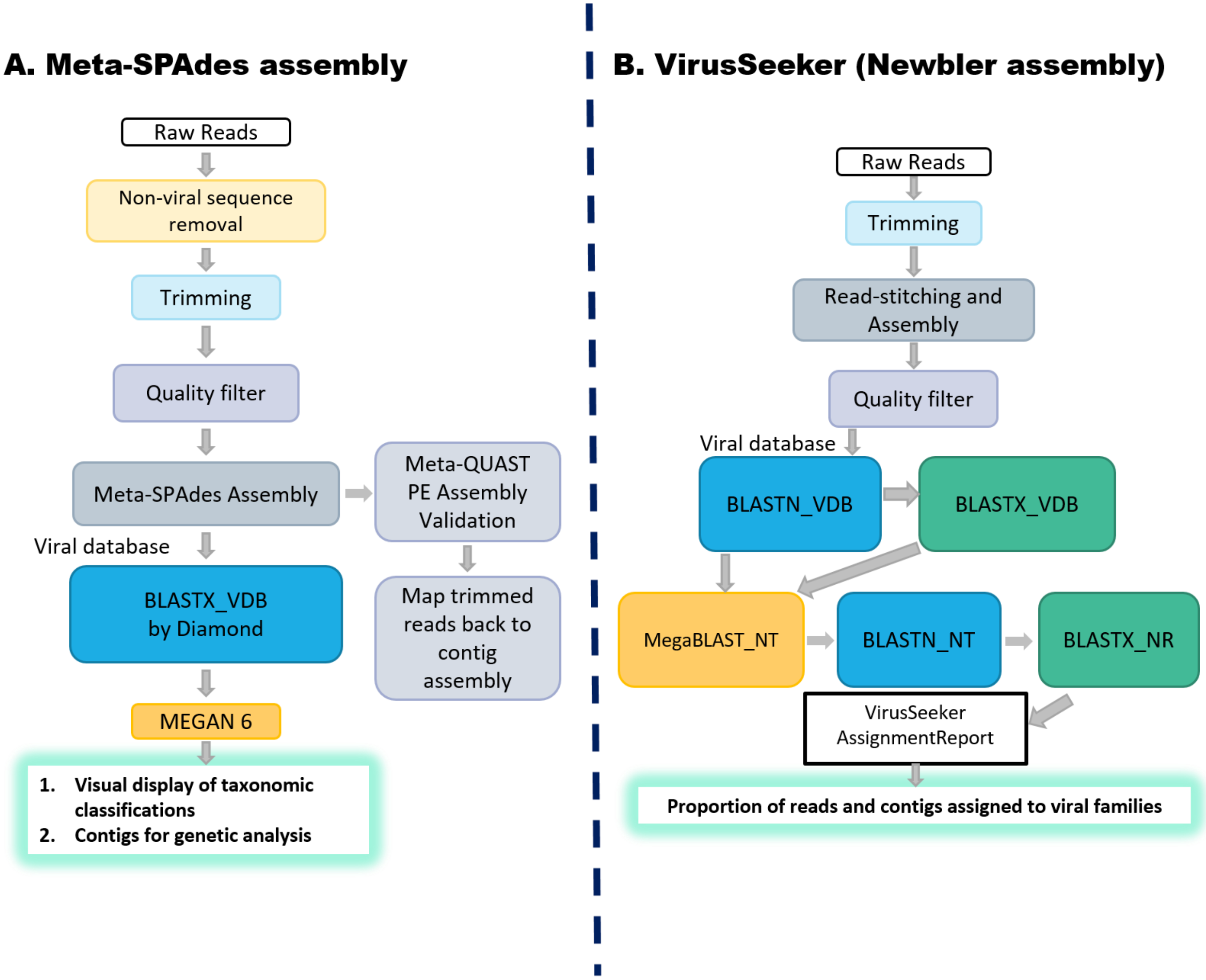

A. Meta-SPAdes assembly
Raw Reads
Non-viral sequence removal
Trimming
Quality filter
Meta-SPAdes Assembly
Meta-QUAST PE Assembly Validation
Map trimmed reads back to contig assembly
Viral database
BLASTX_VDB by Diamond
MEGAN 6
1. Visual display of taxonomic classifications
2. Contigs for genetic analysis
B. VirusSeeker (Newbler assembly)
Raw Reads
Trimming
Read-stitching and Assembly
Quality filter
Viral database
BLASTN_VDB
BLASTX_VDB
MegaBLAST_NT
BLASTN_NT
BLASTX_NR
VirusSeeker AssignmentReport
Proportion of reads and contigs assigned to viral families

Figure 32. Bioinformatic pipelines. Panel A shows the modular, in-house pipeline that was developed using publicly available tools. Panel B shows the steps of VirusSeeker (209).

wide variability in genome sizes present within a metagenomic sample (54). The value was multiplied by 1000 to ease the processing burden in R, as the original normalized values were too low to process using the desired tools. Our normalization approach required manual research of virus genome length and, while time consuming, enabled a fair comparison among viruses detected within the same sample. As the field of metagenomics continues to grow and generate copious amounts of data, appropriate normalization approaches will be crucial for data analysis and interpretation.

Exploration of the lesser dawn bat virome

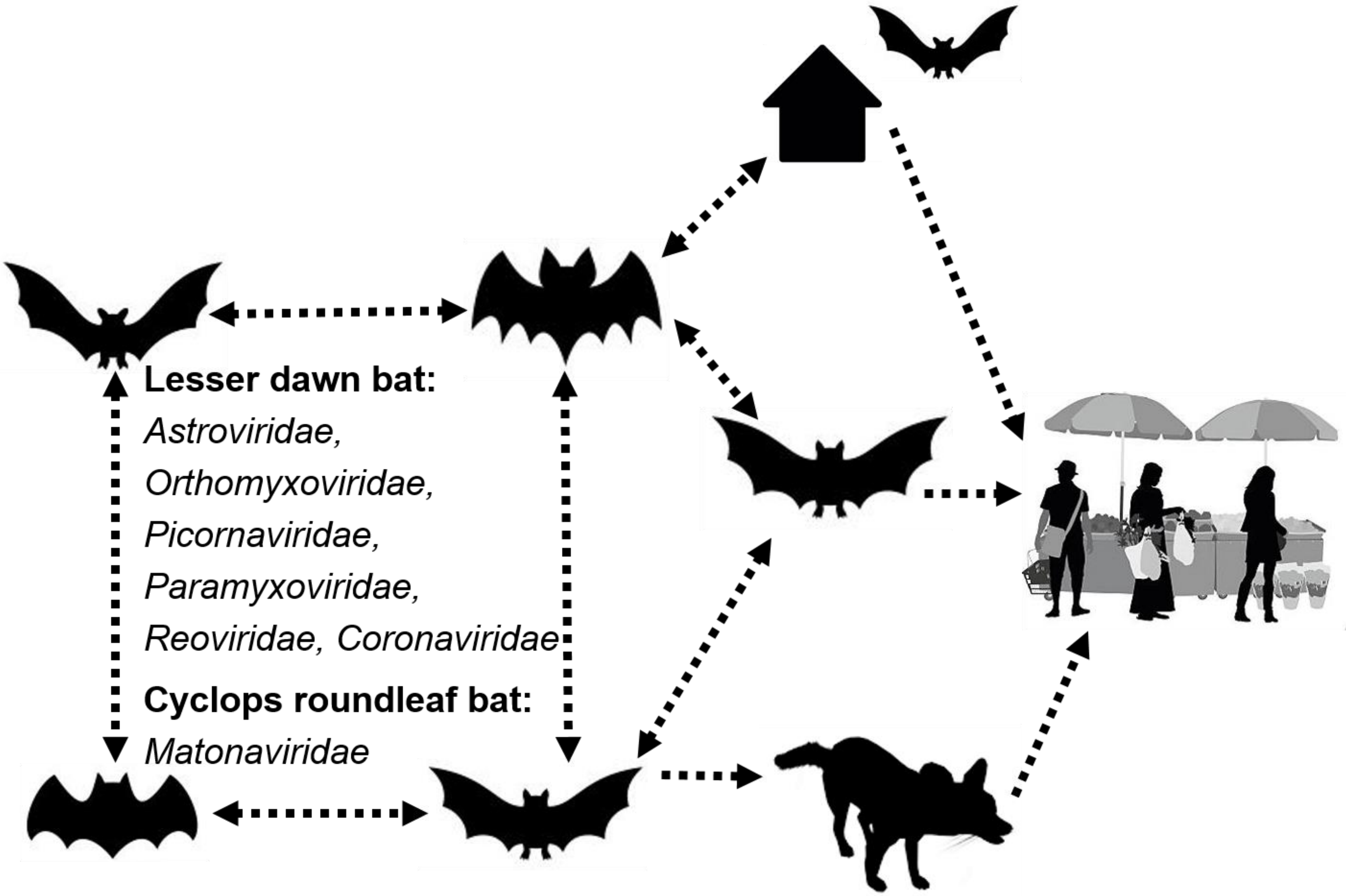
RNA virome characterization

To investigate a previously unknown virome, this thesis work utilized an unbiased, discovery-based approach. While computationally burdensome and requiring detailed manual analysis, such an approach provides the opportunity to uncover previously unknown disease-causing progenitor viruses in the wild prior to the occurrence of an outbreak. These approaches are atypical because biosurveillance is usually biased toward previously-known viruses in response to an outbreak of human disease. For example, the ancestral virus to human SARS-CoV, bat SARS-CoV, was detected in bats through multiple stages of surveillance. The virus was sequenced but not isolated from fecal samples, and genetically similar bat-borne viruses have been sequenced and/or isolated since (110). By limiting viral discovery efforts to urgent outbreak response or to surveillance for viral relatives of known human pathogens, we introduce serious bias into research questions about viruses of pandemic potential. This limitation can be somewhat alleviated by the use of unbiased methods.

Although it is important to survey wild bats to gain a better understanding of circulating viruses, this type of study has limited ability to track individual bats or guarantee the evaluation of the same bats at later times. To our knowledge, our study is the first longitudinal study of virome dynamics in a single captive colony of old world bats. We were able to normalize and track viral abundance at the individual bat level and at the colony level over a period of eighteen months, an advantage of a longitudinal study that is not afforded by snapshot studies in the wild. Through the design of this study, in which variables such as migration were removed, we were able to determine which viruses persisted within the colony by identifying viruses that became undetectable over time and those that were detectable across four or more time points. However, it is possible that viruses that became undetectable long-term were still replicating at levels below the limit of detection for our assays.

Typical biosurveillance approaches focus on predicted intersections of viral transmission (Figure 33), which include locations where synanthropic adaptation has occurred (i.e. bats roosting in human dwellings or business fronts), cases of probable bat-human interaction (i.e. the consumption of smoked bat meat or commercial bat guano farming), and indirect interaction through intermediate amplifying hosts (i.e. masked palm civets or swine). The samples evaluated in Chapters 3, 4 and 5 focused on a different intersection, inter-bat virus transmission, by sampling from colonies of bats. This work reported novel findings in the following viruses of interest for two species of bats: *Astroviridae, Orthomyxoviridae, Picornaviridae, Paramyxoviridae, Reoviridae, Coronaviridae* (lesser dawn bats) and *Matonaviridae* (cyclops roundleaf bats). By better understanding viruses that circulate among bats and by continuing to share sequences across research institutions, we may provide advantages to future outbreak investigators such as clues about potential near neighbors to future outbreak strains.

Figure 33. Targets for biosurveillance. Typical biosurveillance begins with a clue from an outbreak that is gathered by tracing contacts of infected cases. Beginning at the right of the figure, using a marketplace as an example, this process usually works backward to identify possible interactions between wildlife reservoirs and humans or intermediate hosts. Cataloging viruses that persist in bat colonies could help responders to future outbreaks determine where to direct response efforts as soon as the virus is detected. If the host reservoir or intermediate host is quickly traceable, then public health interventions can be executed more efficiently. Important viruses that were detected in Singapore (lesser dawn bat) and have particular relevance to public health include influenza A virus, novel bat filovirus, bat mumps paramyxovirus and Rousettus bat coronavirus GCCDC1. A rubella-like virus, Ruhugu virus, was discovered at the study site in Uganda (cyclops round leaf bat).


Lesser dawn bat:
Astroviridae,
Orthomyxoviridae,
Picornaviridae,
Paramyxoviridae,
Reoviridae, Coronaviridae
Cyclops roundleaf bat:
Matonaviridae

Persistent viruses in the captive bat colony

Given the observations documented in the longitudinal bat virome study, it is reasonable to conclude that several zoonotic-related viruses are capable of long-term persistence in bat colonies. While it is useful to know that these viruses persist, and to consider the potential for cross-species transmission, it is critical to investigate the true tropism of circulating bat viruses for other mammalian cells. As we generate more data about viral evolution, and machine learning tools continue to be developed using accurate and rich data, it may be possible in the future to extrapolate the spillover risk for newly discovered viruses circulating in wild bats, given knowledge of the viral sequences.

While colony-level persistence was detected in viruses belonging to several families, zoonotic agent-related viruses were detected sporadically at the individual bat-level. This study utilized swabs from the exterior of the bat (head and body) to evaluate the virus population. We observed that, while swabs of individual bats were not sufficiently representative of colony-level infection, the combined data among each of the four swab types was reflective of colony-level infection. It is possible that we did not capture all persistent viruses due to the nature of our noninvasive sampling study design; however, we hypothesize that the findings from external and shedder sites are of particular relevance because these viruses are most readily transferred beyond an individual bat's body. We interpret our findings to indicate that the viruses we detected most frequently may possess a propensity for spillover due to their comparatively higher recurrence in shedder sites, at abundance sufficient for detection by our methods. Taken together, we conclude that noninvasive surveillance methods that target the body of bats detect viruses shed by individual bats, while also representing viral populations circulating within the entire colony. External swabs could be informative targets for sample collection.

Non-zoonotic agent-related viruses and unclassifiable dark matter

Detection rates vary by host taxonomy and there is room for exploration of factors that influence this challenge. Viruses are typically detected in bats via swab, blood, or tissue collection (206). In the case of variable detection of genomic evidence of viruses in oral swabs, it could be possible that the oral cavity of a fruit or nectar bat could provide a different niche (influenced by factors such as pH (45)) in which viruses could be detected as compared to the oral cavity of an insectivorous bat. Additionally, an individual bat's level of aggression and immunological background may influence its shedding patterns. Therefore, we hypothesize that the discovery rate of viruses among bat species varies by diet and behavior of those species, and may not necessarily solely reflect host restriction for certain viruses.

Through this work, an estimated 293 Mbp of sequence data was unclassifiable and a portion of that is likely viral. These data consist of high-quality reads that lack a known sequenced neighbor. While this work did not directly address viral dark matter, it is possible that these data could be classified at a future time as public databases grow. Approaches to interrogate dark matter have been discussed in the literature and it is a growing topic of interest (92). For data generated by this thesis project, clustering dark matter in large datasets that encompass sequence reads from an entire colony could reveal common sequences and provide clues to identify novel viruses. This approach, in conjunction with cross-assembly, would mimic the methodology used to discover crAssphage, a bacteriophage that is now known to be ubiquitous in the human gut (46).

We hypothesize that analysis of dark sequence matter in rectal swabs may reveal a bat crAssphage that holds similarities to human crAssphage.

DISCOVERY OF NEW VIRUSES

Following the longitudinal analysis of the lesser dawn and cyclops roundleaf bat virome, we identified specific viral targets of interest for biosurveillance. In the case of the lesser dawn bats, these include persisting viruses that have been shown to cross the species barrier, such as *Coronaviridae, Reoviridae* and *Paramyxoviridae*. By studying the virome of cyclops roundleaf bats, we identified a novel rubella-like virus. Human rubella virus can cause severe congenital birth defects and miscarriage and until now the only known host was humans; therefore it is important to further investigate this newly discovered near-neighbor. Ruhugu virus should be a target of biosurveillance in cyclops roundleaf bats and possible co-roosting bat and squirrel species. The threat of widespread rubella-like viruses posed by an animal reservoir has not yet been elucidated, and further investigation should be prioritized toward their potential for pathogenicity.

Our sample collection incorporated external and internal swab sites for virus detection. We observed that external swabs were equally useful representations of colony-level infection as internal swabs. In future surveillance, it may be more practical to collect a large number of less invasive, external swabs to evaluate circulating viruses in a bat colony, than to catch a smaller number of bats, but perform more invasive internal swabbing. A less invasive sample collection could make amenable the creation of more remote surveillance approaches. Notably, external swab collection would likely be more informative when performed on gregarious bats than on more solitary species such as *H. cyclops* that exhibit behaviors with distinctly fewer bat-to-bat interactions.

The findings of this work are relevant to bat surveillance and eradication efforts in regions surrounding Singapore and Kibale National Park in Uganda. The range of each bat species expands beyond the sampling region and the detected viruses may have wider prevalence across Southeast Asia, including China (lesser dawn bats), and across equatorial Africa (cyclops roundleaf bats). Rousettus bat coronavirus GCCDC1, for example, was previously detected exclusively in Yúnnan Province, China. However, this study detected the virus with high prevalence among lesser dawn bats in Singapore, some 2500 km to the south. The more we can fill insufficiencies by cataloging the circulation of known viruses, the better we can predict future viral spread and emergence.

Our capabilities to predict which viruses will spill over from bats to humans still require substantial additional data and virus discovery research. Knowledge of near neighbors provides a more informed starting place from which to investigate future outbreaks. The SARS-CoV-2 (NC_045512) outbreak is a timely example of how our efforts could improve future outbreak response. This novel virus spread to four countries within weeks of detection in December 2019. The virus is most closely related to bat SARS-CoV and human SARS-CoV. There is 96% shared identity among the SARS-CoV-2 strain and a bat SARS-CoV strain (111; 212). Outbreak response is influenced by the knowledge that many cases were associated with the Wuhan seafood market in central China, yet despite one month of searching, there are few clues as to the origin and amplifying host of the virus. If we knew more about the present circulation of viruses in the wild, we might identify other viral sequences that sit more closely on the phylogenetic tree to SARS-CoV-2. Additional clues from sequence data may allow us to predict an outbreak's source, including a host reservoir, and provide a rational starting place for surveying amplifying intermediate or incidental hosts and target an intervention. Work similar to this thesis could potentially help future outbreak responses by providing valuable information about where to intervene or search for disease-causing progenitor viruses. These results indicate that biosurveillance should focus on colony-level infections within bat reservoirs, rather than snapshots of data collected from individual bats.

CONCLUSION

This project contributed massive quantities of raw sequence data, classified viruses from two species of bats, the discovery of multiple zoonotic agent-related viruses and a comprehensive pipeline to efficiently process complex metagenomic samples. The raw sequence data and corresponding metadata can be downloaded via NCBI's SRA (PRJNA494391, PRJNA561193). Additionally, knowledge gained through the characterization of our biosurveillance probe panel has been useful in determining the extent of binding tolerance that our existing panel may exhibit for the detection of emerging and re-emerging viruses such as the novel virus, SARS-CoV-2. The findings of this work can be used to build upon our understanding of the virosphere and further develop existing biosurveillance methods.

APPENDIX A: Relevant novel bat virus discoveries

PARAMYXOVIRUSES

Paramyxoviruses have negative-sense, single-stranded RNA genomes and include some of the most significant mammalian viruses such as mumps virus, measles virus and distemper virus (43). Unclassifiable genomic sequences related to paramyxovirus L and N genes were previously reported in fecal samples collected from lesser dawn bats in Singapore (124). Additionally, the L gene (KC599257) of a henipa-related virus was detected in lesser dawn bat fecal samples in China (207). To our knowledge, no full-length paramyxovirus genomes detected in lesser dawn bats have been published.

Novel bat paramyxovirus

A near-neighbor to Canine distemper virus (CDV) was detected at one time point in 2016 by read mapping to Canine distemper virus reference AF378705. Various distemper viruses exist within the morbillivirus genus and cause disease in animals such as dogs and other canines, seals, dolphins and pigs. CDV infection in canines results in severe immune suppression but, although human infection is possible, the virus is highly host-adapted and does not result in human disease (37). One morbillivirus-related paramyxovirus, Bat paramyxovirus (NCBI Taxonomy ID: 1300978), was identified in Brazillian neotropical vampire bats and shares higher amino acid identity with canine morbilliviruses than human-associated morbilliviruses (43). When reads were mapped to CDV with parameters set to require greater than 95% identity, one sample sequenced by both shotgun and targeted enrichment sequencing was found to contain many reads that map with high confidence (Samples 3 and 16, **Table 12**). Furthermore, 10,591,403 of 43,022,904 trimmed reads from targeted-enrichment data from Sample 3 rectal swab mapped to CDV reference AF378705 with mapping parameters requiring 75% identity or higher (consensus shown in **Figure 34**).

Morbillivirus genomes are approximately 15 kb and encode for eight proteins: N, C, V, P, M, F, H and L. RNA editing produces the V protein and leaky scanning of the P mRNA produces the C protein. Reads covered 40.3% of the genome, including each of the eight protein-encoding genes. Amino acid identity ranged from 87-100%.

Our findings are the second report of a morbillivirus relative detected in bats. Deeper sequencing and recovery of a full genome sequence or isolate of the virus could provide informative data with regard to functional similarity to CDV. Given the knowledge that a near neighbor to CDV is circulating in lesser dawn bats, it would be relevant for public health officials to have the data necessary to extrapolate risk of spillover to other mammals.

Novel bat mumps virus

Human mumps virus causes a common and vaccine-preventable childhood disease. A study of antigenic relatedness between human mumps virus and related bat-borne mumps viruses, genus *Rubulavirus*, suggested that they belong to one serogroup (43). Furthermore, the shared amino acid identity between bat mumps virus (HQ660095) and human mumps virus is above 89.5% in all genes.

A novel bat mumps virus was detected in 46 swabs collected from the captive lesser dawn bat colony in April, July, October 2016 and May 2017. Contigs from one body swab collected from bat 7633EDB in April 2016 cover 13653 nt of 15378 nt (88.8%) of the published bat mumps virus reference (HQ660095) with 54.23% identity at

Sample	Source	Sequencing method	Avg. fold coverage	% Covered	Covered bases	Plus reads	Minus reads	Total reads
1	Isolate	Enriched	0.0191	1.1472	180	1	1	2
2	Oral		0.2677	14.0153	2199	17	11	28
3	Rectal		76144.65	30.7776	4829	3978610	3986121	7964731
4	Body		0.2008	13.4799	2115	11	10	21
5	Head		0.4876	18.4576	2896	23	28	51
16	Rectal	Shotgun	5.1147	27.9669	4388	267	268	535

Table 12. Results from samples with sequence data that mapped to CDV reference AF378705 requiring 95% identity or higher. Stringent read mapping results indicate that Sample 3, taken from the same rectal swab as Sample 16, contains a large number of reads that map to the reference with high stringency.

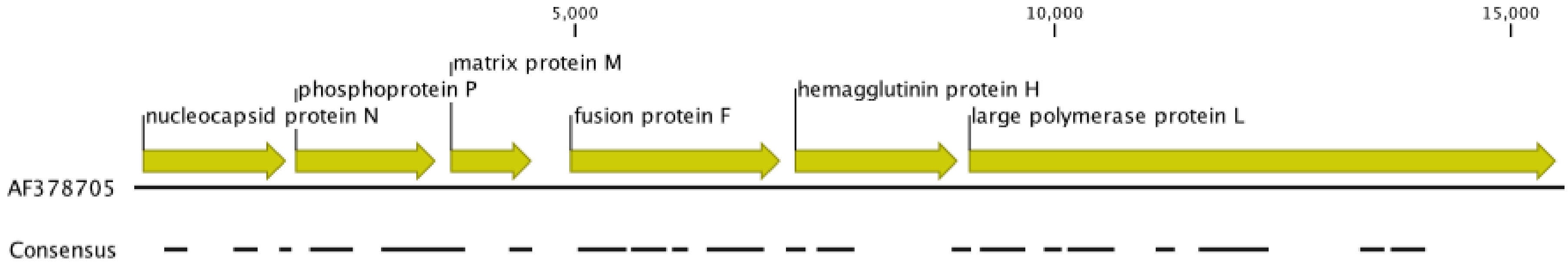


Figure 34. Read mapping of trimmed reads to CDV reference. 10,591,403/43,022,904 trimmed reads from Sample 3, a rectal swab, mapped to CDV reference AF378705 with mapping parameters: 0.9 length fraction and 75% identity. The consensus shown covers 40% of the CDV reference genome.

Table 13. Prevalence of novel bat mumps virus at each time point

Date	**# bats sampled**	**# bats positive**	**% bats positive**	**# swabs sequenced**	**# swabs positive**	**% swabs positive**
Apr-16	18	18	100.00%	41	26	63.41%
Jul-16	19	16	84.21%	28	21	75.00%
Oct-16	20	20	100.00%	75	53	70.67%
Jan-17	11	2	18.18%	14	2	14.29%
May-17	15	1	6.67%	21	1	4.76%
Sep-17	13	8	61.54%	27	11	40.74%

the nucleotide level. The prevalence of this virus at each time point is summarized in Table 13.

Novel bat mumps virus was not detected at each of the six time points and did not persist beyond May 2017 in shedder sites. It is unknown whether the virus replication fell below the limit of detection for our assay or the virus was cleared from the colony. In a study of the natural history of infection by measuring morbilli-related paramyxovirus RNA, a *Myotis myotis* breeding colony was found to shed viral RNA at a constant concentration over three observation periods spanning the years 2008 through 2010 (43). This report indicates the propensity for viral persistence of morbilli-related PVs in bats.

Our discovery of bat mumps virus in Southeast Asia provides further evidence of the ecological relationship between bats and mammalian paramyxoviruses, although a bat mumps rubulavirus yet remains to be isolated. Sequence data for bat mumps virus has been generated from spleen (43), kidney (128), oral, rectal, head and body swabs (this work). Further investigation of similar sample types may expand the knowledge of the evolutionary relationship between bats and mammalian paramyxoviruses.

NOVEL BAT FILOVIRUS

Ebolavirus is a negative-sense, single-stranded RNA filovirus that causes a severe viral hemorrhagic fever in humans for which there is currently no standard effective treatment. Ebola virus, which has previously caused devastating outbreaks in West Africa, is currently causing a major outbreak in the Democratic Republic of the Congo (84). Old World fruit bats are thought to be natural reservoirs for filoviruses and several novel bat filoviruses have recently been discovered in China and Africa (63; 202). In response to concern for spillover of filoviruses from bats to humans, modelling was performed by the EcoHealth Alliance to predict the location of bats that host filoviruses. This work determined the most likely geographic location to be Southeast Asia and *Eonycteris spelaea*, an understudied Old World fruit bat, was identified as a potential host species (70). Serological (e.g. indirect) evidence of filovirus infection in this species of bats has been published, but filovirus isolation from these bats has yet to be reported (98). Through this study, we have discovered genetic evidence of a filovirus in bat swabs. This evidence was identified in three distinct swab samples through a combination of unbiased high throughput sequencing and targeted enrichment sequencing.

Filovirus reads were detected in one rectal swab collected in April 2016, one body swab collected in September 2017 and one body swab collected in May 2017. Reads from the body swab collected in May 2017, Swab 340, were used for phylogenetic analysis. Phylogenetic analysis of the novel bat filovirus demonstrates that the unclassified bat filovirus is most closely related to Mengla dianlovirus, which was recently isolated from Rousettus bats in China (Figure 35) (202). A maximum likelihood phylogeny using a consensus sequence of reads from a body swab was generated using MEGA7 based on the General Time Reversible model (94) (130).

An unclassifiable filovirus L gene was detected in lesser dawn bats in China in RNA from homogenized spleen and lung tissue, reported in 2017 (203). Our discovery of genomic sequence in multiple bats is confirmatory of the first report and provides genomic evidence of a filovirus in shedder sites (rectal swab) of lesser dawn bats, indicating a potential route of dispersal for the virus. Limited conclusions can be drawn

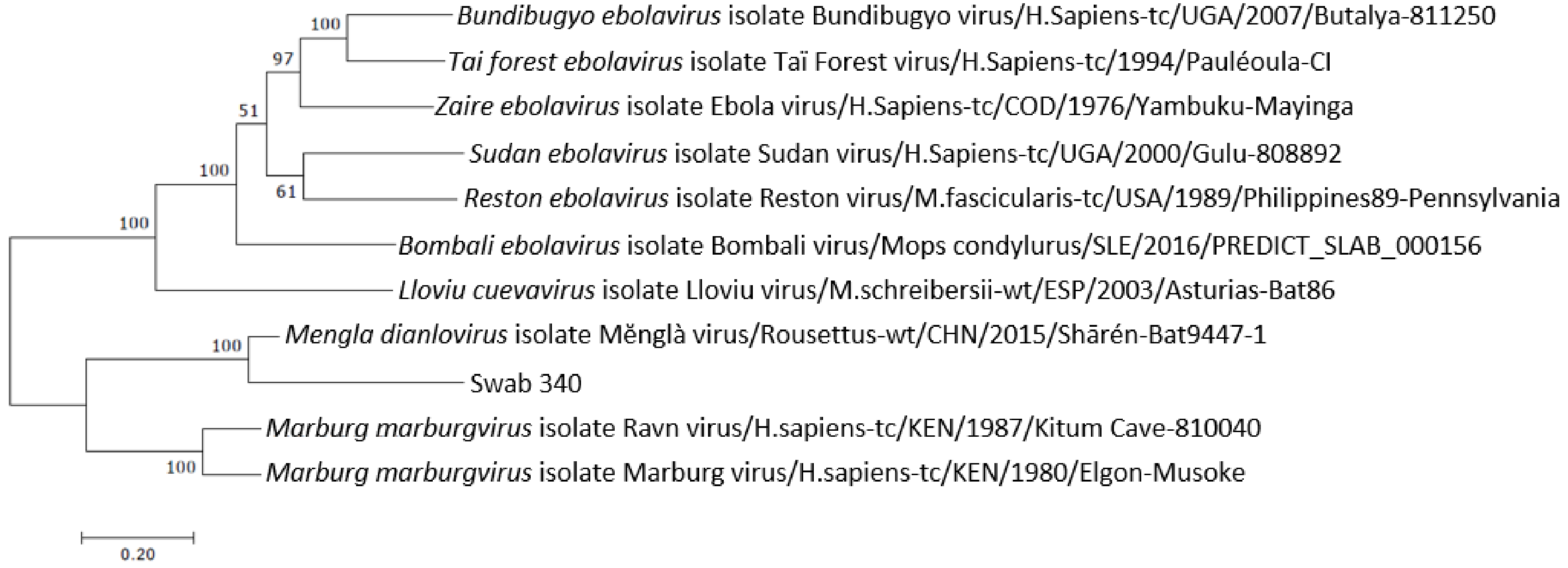

100
97
51
100
61
100
100
100
Bundibugyo ebolavirus isolate Bundibugyo virus/H.Sapiens-tc/UGA/2007/Butalya-811250
Tai forest ebolavirus isolate Taï Forest virus/H.Sapiens-tc/1994/Pauléoula-CI
Zaire ebolavirus isolate Ebola virus/H.Sapiens-tc/COD/1976/Yambuku-Mayinga
Sudan ebolavirus isolate Sudan virus/H.Sapiens-tc/UGA/2000/Gulu-808892
Reston ebolavirus isolate Reston virus/M.fascicularis-tc/USA/1989/Philippines89-Pennsylvania
Bombali ebolavirus isolate Bombali virus/Mops condylurus/SLE/2016/PREDICT_SLAB_000156
Lloviu cuevavirus isolate Lloviu virus/M.schreibersii-wt/ESP/2003/Asturias-Bat86
Mengla dianlovirus isolate Měnglà virus/Rousettus-wt/CHN/2015/Shārén-Bat9447-1
Swab 340
Marburg marburgvirus isolate Ravn virus/H.sapiens-tc/KEN/1987/Kitum Cave-810040
Marburg marburgvirus isolate Marburg virus/H.sapiens-tc/KEN/1980/Elgon-Musoke
0.20

Figure 35. Phylogeny of bat filovirus detected in May 2017

without an isolate or complete genome sequence. One approach to close the knowledge gap is to expand our target enrichment panel to include the recently discovered Mengla dianlovirus, Bombali ebolavirus and the sequences discovered through our work. Mengla dianlovirus was first reported in March 2019 and isolated from homogenized tissues from a *Rousettus* bat in China (202). Bombali ebolavirus was discovered in RNA extract from oral and rectal swabs of *Chaerephonpumilus* and *Mops condylurus* bats in Sierra Leone (63). The detection of a filovirus in rectal and body swabs is consistent with the literature and it is possible that deeper sequencing could provide more informative data. As we have detected only partial genomic sequences thus far, there is not enough sequence data for us to draw conclusions with regard to the similarity of this virus to filoviruses that infect humans; more information is required to understand the potential for this virus to pose a threat.

NOVEL BAT INFLUENZA A VIRUS

Influenza A virus (IAV) is an enveloped virus with a negative-sense, segmented RNA genome. It is notorious for crossing species barriers and capable of causing pandemics (185). IAV has previously been detected in South American bats (171; 172) and most recently in Egyptian *Rousettus aegyptiacus* (85). As more surveillance is performed globally, it is possible that a larger geographic range of bat-borne influenza viruses may be discovered and that the capacity for reassortment of bat-borne *Orthomyxoviruses* with known human influenza A viruses will be understood (198). To our knowledge, our virome study is the first to report the detection of IAV in Southeast Asian bats. IAV was detected in 24 swabs but no full-length segments were recovered, preventing unequivocal strain typing. Similar to other zoonotic-related viruses, IAV was not detected in the same bat's oral or rectal swabs at subsequent time points. This could reflect cycles of transmission among the colony or possibly persistent virus replication and shedding below the limit of detection for our assays. The low number of reads and incomplete coverage of all eight segments of IAV is consistent with the dearth of bat-borne influenza virus data.

There was not enough coverage of the NA gene to perform phylogenetic analyses, however it was possible to construct a maximum likelihood phylogeny using consensus sequences of reads aligning to HA. The phylogeny in Figure 36 was inferred using MEGA7 based on the General Time Reversible model (94) (130). The highest log likelihood was -618, likely due to the lack of coverage for the HA gene in data from the selected swabs.

The four representative swabs in the maximum likelihood phylogeny, Swabs 14, 55, 205 and 390 were collected from distinct bats and do not cluster together. Swab 14 falls on the branch with H1-type IAVs, Swab 55 and 205 fall on the branch with type H18-type IAVs, and Swab 390 falls on a branch with H5-type IAVs. It is possible that different results could be inferred with data yielding greater coverage across the HA gene, but without deeper sequencing we cannot unequivocally type the IAV circulating among lesser dawn bats in Singapore.

Future studies could pursue isolation of IAV from lesser dawn bats either from oral or rectal samples. Alternatively, hybridization-based target enrichment sequencing using probes designed against bat-borne IAV strains could mitigate the problem of low breadth of coverage across each gene segment. Clear typing of HA for this IAV could be

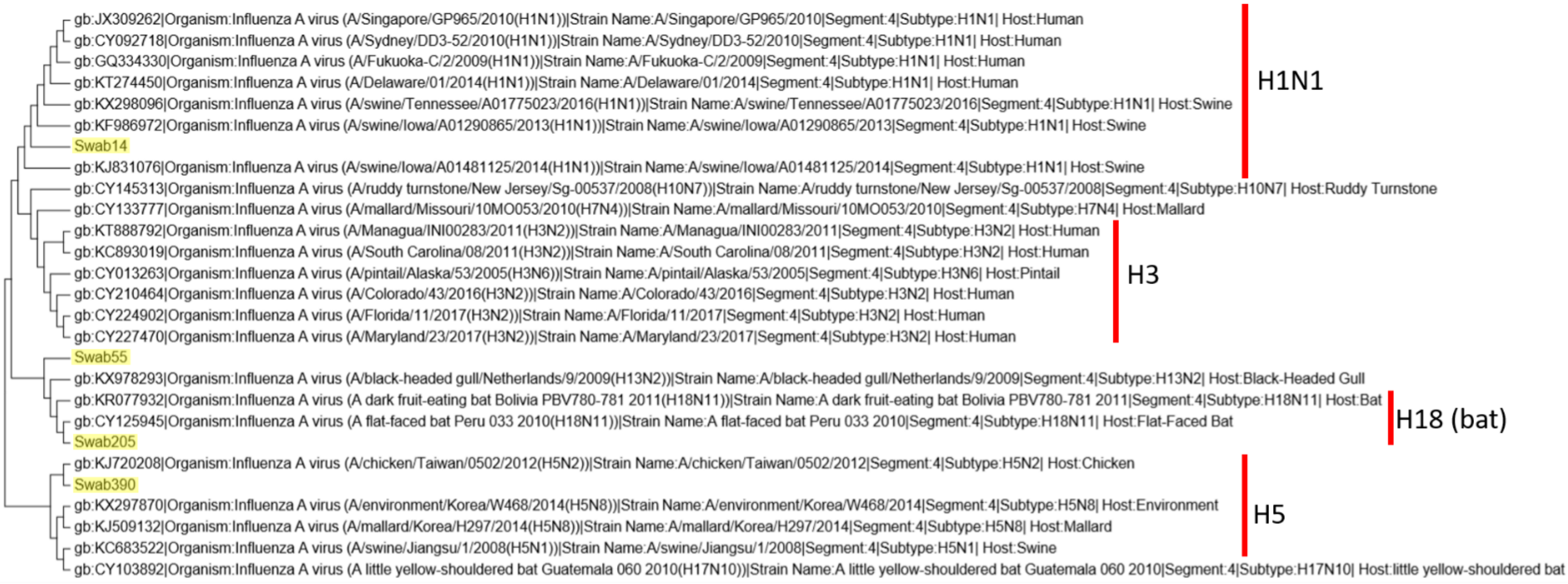

gb:JX309262|Organism:Influenza A virus (A/Singapore/GP965/2010(H1N1))|Strain Name:A/Singapore/GP965/2010|Segment:4|Subtype:H1N1| Host:Human
gb:CY092718|Organism:Influenza A virus (A/Sydney/DD3-52/2010(H1N1))|Strain Name:A/Sydney/DD3-52/2010|Segment:4|Subtype:H1N1| Host:Human
gb:GQ334330|Organism:Influenza A virus (A/Fukuoka-C/2/2009(H1N1))|Strain Name:A/Fukuoka-C/2/2009|Segment:4|Subtype:H1N1| Host:Human
gb:KT274450|Organism:Influenza A virus (A/Delaware/01/2014(H1N1))|Strain Name:A/Delaware/01/2014|Segment:4|Subtype:H1N1| Host:Human
gb:KX298096|Organism:Influenza A virus (A/swine/Tennessee/A01775023/2016(H1N1))|Strain Name:A/swine/Tennessee/A01775023/2016|Segment:4|Subtype:H1N1| Host:Swine
gb:KF986972|Organism:Influenza A virus (A/swine/Iowa/A01290865/2013(H1N1))|Strain Name:A/swine/Iowa/A01290865/2013|Segment:4|Subtype:H1N1| Host:Swine
Swab14
gb:KJ831076|Organism:Influenza A virus (A/swine/Iowa/A01481125/2014(H1N1))|Strain Name:A/swine/Iowa/A01481125/2014|Segment:4|Subtype:H1N1| Host:Swine
H1N1
gb:CY145313|Organism:Influenza A virus (A/ruddy turnstone/New Jersey/Sg-00537/2008(H10N7))|Strain Name:A/ruddy turnstone/New Jersey/Sg-00537/2008|Segment:4|Subtype:H10N7| Host:Ruddy Turnstone
gb:CY133777|Organism:Influenza A virus (A/mallard/Missouri/10MO053/2010(H7N4))|Strain Name:A/mallard/Missouri/10MO053/2010|Segment:4|Subtype:H7N4| Host:Mallard
gb:KT888792|Organism:Influenza A virus (A/Managua/INI00283/2011(H3N2))|Strain Name:A/Managua/INI00283/2011|Segment:4|Subtype:H3N2| Host:Human
gb:KC893019|Organism:Influenza A virus (A/South Carolina/08/2011(H3N2))|Strain Name:A/South Carolina/08/2011|Segment:4|Subtype:H3N2| Host:Human
gb:CY013263|Organism:Influenza A virus (A/pintail/Alaska/53/2005(H3N6))|Strain Name:A/pintail/Alaska/53/2005|Segment:4|Subtype:H3N6| Host:Pintail
gb:CY210464|Organism:Influenza A virus (A/Colorado/43/2016(H3N2))|Strain Name:A/Colorado/43/2016|Segment:4|Subtype:H3N2| Host:Human
gb:CY224902|Organism:Influenza A virus (A/Florida/11/2017(H3N2))|Strain Name:A/Florida/11/2017|Segment:4|Subtype:H3N2| Host:Human
gb:CY227470|Organism:Influenza A virus (A/Maryland/23/2017(H3N2))|Strain Name:A/Maryland/23/2017|Segment:4|Subtype:H3N2| Host:Human
H3
Swab55
gb:KX978293|Organism:Influenza A virus (A/black-headed gull/Netherlands/9/2009(H13N2))|Strain Name:A/black-headed gull/Netherlands/9/2009|Segment:4|Subtype:H13N2| Host:Black-Headed Gull
gb:KR077932|Organism:Influenza A virus (A dark fruit-eating bat Bolivia PBV780-781 2011(H18N11))|Strain Name:A dark fruit-eating bat Bolivia PBV780-781 2011|Segment:4|Subtype:H18N11| Host:Bat
gb:CY125945|Organism:Influenza A virus (A flat-faced bat Peru 033 2010(H18N11))|Strain Name:A flat-faced bat Peru 033 2010|Segment:4|Subtype:H18N11| Host:Flat-Faced Bat
Swab205
H18 (bat)
gb:KJ720208|Organism:Influenza A virus (A/chicken/Taiwan/0502/2012(H5N2))|Strain Name:A/chicken/Taiwan/0502/2012|Segment:4|Subtype:H5N2| Host:Chicken
Swab390
gb:KX297870|Organism:Influenza A virus (A/environment/Korea/W468/2014(H5N8))|Strain Name:A/environment/Korea/W468/2014|Segment:4|Subtype:H5N8| Host:Environment
gb:KJ509132|Organism:Influenza A virus (A/mallard/Korea/H297/2014(H5N8))|Strain Name:A/mallard/Korea/H297/2014|Segment:4|Subtype:H5N8| Host:Mallard
gb:KC683522|Organism:Influenza A virus (A/swine/Jiangsu/1/2008(H5N1))|Strain Name:A/swine/Jiangsu/1/2008|Segment:4|Subtype:H5N1| Host:Swine
H5
gb:CY103892|Organism:Influenza A virus (A little yellow-shouldered bat Guatemala 060 2010(H17N10))|Strain Name:A little yellow-shouldered bat Guatemala 060 2010|Segment:4|Subtype:H17N10| Host:little yellow-shouldered bat

Figure 36. Topology of Influenza A virus segment 4 consensus sequences.

pertinent to public health, as H18-mediated cell entry has been reported to use MHC-II molecules and could indicate broad tropism (87). NA-like proteins may be dispensable for viral replication of the H18N11 bat-borne strains, therefore a relevant next step should be to determine the HA type of the IAV circulating in lesser dawn bats.

METHODS

Bat Paramyxovirus

High throughput sequencing data was generated using 2x75 bp read lengths on the Illumina HiSeq platform by Danielle Anderson at Duke NUS for 10 swabs collected from the captive colony in 2016. Information about the samples was blinded and 5 samples were designated as enriched (samples 1-5). The samples were trimmed and filtered for quality using bbduk (19). Reads were mapped to the *Eonycteris spelaea* genome and bacterial database then discarded. The remaining reads were assembled using metaSPAdes and reads were mapped back to contigs (132). Taxonomic classifications were determined by DIAMOND using the BlastX algorithm against RefSeq viral protein sequences from NCBI (3; 18). Results were visualized using MEGAN 6 (81). VirusSeeker was used as orthogonal confirmation for taxonomic assignments (209).

Read mapping was also performed using bbsplit (19) with a reference index consisting of Rousettus bat coronavirus GCCDC1 (NC_030886), Canine distemper virus (AF378705) and a scaffolded mumps virus reference that was generated from sequence data from a body swab collected in April 2016 from bat 7633EDB. Bbsplit uses an algorithm that determines mapping agreement based on the ratio of each read’s alignment score to the maximum alignment core (where 100% of bases match the reference). A ratio cutoff of 0.9, which is greater than 95% identity across the read, was used.

Novel bat mumps virus, novel bat filovirus and novel influenza A virus

Methods as described in Chapter 3 were utilized. In brief, contigs assembled by metaSPAdes (132), stitched reads from VirusSeeker (209) and consensus sequences from read mapping were utilized to assemble as much of the novel virus genome as possible using CLC Genomics Workbench V11(QIAGEN Bioinformatics; Redwood City, CA).

The International Committee for the Taxonomy of Viruses (ICTV) defines viral species as “a monophyletic group of viruses whose properties can be distinguished from those of other species by multiple criteria” and identity cutoffs vary by genus and are determined by natural and experimental host range, cell and tissue tropism, pathogenicity, vector specificity, antigenicity and the degree of relatedness of genomes (82). In this study, cutoffs of 90% or higher at the amino acid level and 95% or higher at the nucleotide level for the definition of a viral strain were used in viral classification. Sequences that were not sufficiently related to known species were classified using the following naming convention: “novel [name of viral family] virus.”

The evolutionary history of a novel filovirus and influenza-like virus detected in the sample was inferred by using the Maximum Likelihood method based on the General Time Reversible model (130) with 100 bootstrap replications.

APPENDIX B: The lesser dawn bat microbiota

This thesis is focused on zoonotic agent-associated viruses in bats; however, due to the unbiased nature of shotgun sequencing and relatively low multiplexing utilized for the longitudinal bat virome study, it is possible to evaluate the full microbiota of the lesser dawn bat using the data published by this thesis work. To explore the complete microbial richness of our shotgun sequence data, classification of all domains of life was performed using Burrows-Wheeler Aligner (BWA) for read-based analysis (107). We also investigated the richness and complexity of samples by evaluating the virome data for patterns of co-occurrence of viruses within individual swabs. In this section, we discuss preliminary results and propose possible avenues to extend these microbiota analyses.

MICROBIOTA CLASSIFICATION

To gain understanding of the most prevalent domains of life detected by shotgun sequencing of head, body, oral and rectal swabs, taxonomy classification was performed using BWA (107) against NCBI's non-redundant database (nr). Quality-controlled, host-removed reads were obtained as previously described and subsampled to 500,000 reads using seqtk (1); read-based taxonomy classification was performed by EDGE Software (108) using BWA against RefSeq. Heatmaps were generated to visualize abundance of classified taxa across the whole colony using ggplot2 in R (167).

The abundance of the top 40 microbial species is represented across each swab in **Figure 37**. Overall, microbial abundance values increased in 2016 time points following the bats' introduction to captivity. We observe that, as expected, the microbiota increased in richness following the introduction of 14 new, wild-caught bats in the summer of 2017.

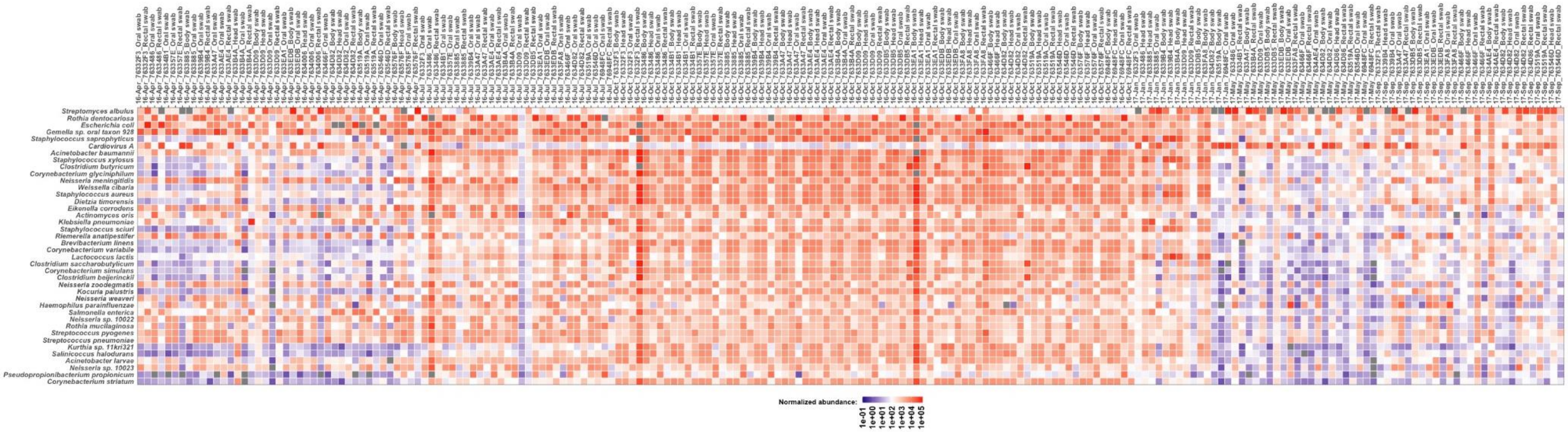


Figure 37. Read-based taxonomic classification was performed for subsampled read sets of 500,000 using BWA. The 40 most abundant species are represented along the y-axis and each swab is described along the top x-axis.

PEARSON ANALYSES FOR CO-OCCURRENCE OF VIRUSES

We were interested in assessing the influence of viral coinfection on the bat microbiota by examining shotgun data that was generated by our virome study. While we did not address the entire microbiota in coinfection analyses, we leveraged our viral classification data to begin to answer questions about coinfection by studying viral co-occurrence. To address distribution of viruses across samples and evaluate individual samples for trends in viral co-occurrence, we calculated Pearson correlation coefficient values for each sample. Pearson correlation coefficient is a classical method by which to measure the statistical association between two variables, based on covariance. One possible pitfall of this approach is that an analysis like this will not necessarily detect complex microbial interactions (103).

We hypothesized that shedder sites (oral and rectal) are more restrictive of viral diversity due to competition among replicating microbes, ultimately resulting in the correlation of shedding for a dominant set of viruses in oral and rectal swabs. In other words, competition would result in a dominant set of viruses in each niche. By logic, we expected to identify a higher total correlation among viruses in carrier sites (head and body) representing the skin outside of a bat. Viruses cannot replicate in bat fur yet are found in carrier sites, and so we based this hypothesis on the assumption that the exterior of the bat may act as a mechanical vector to viruses shed by other bats.

To perform Pearson Correlation analysis, the lower triangle of a correlation matrix for viruses detected in shedder (oral, rectal) and carrier (head, body) swabs was generated using R base (141; 167). Results were reordered by Pearson correlation coefficient value and were defined as significant if the Pearson correlation value was less than -0.6 (negative correlation) or more than 0.6 (positive correlation). Pearson correlation values range from -1 to 1, and a value less than -0.6 or more than 0.6 indicates a strong linear relationship among the correlation between the abundance of viruses. The visualization program ggplot2 in R and Graphpad Prism 7 were used to generate figures.

Through analyzing Pearson correlation coefficient values, we identified that viruses detected in carrier swabs had fewer negative associations (less than -0.6) than viruses detected in shedder swabs (**Figure 38A, 38B**). This finding could support our hypothesis that viruses detected in shedder sites are more reflective of dominant viruses than the viruses detected in external sites.

In addition to negative correlations found among viruses in shedder swabs (**Figure 38A**), several intriguing positive correlations among viruses were noted. Novel bat paramyxovirus was positively correlated with influenza B virus. Mumps virus was positively correlated with influenza A virus (IAV) and Rousettus bat coronavirus GCCDC1, and unclassified bat rotavirus was positively correlated with unclassified filovirus. Statistical associations could be followed up by evaluating how similar this co-occurrence is found in other metagenomic data or by examining microbial interactions *in vitro*.

Several zoonotic agent-related taxa found in carrier swabs had multiple negative correlations with dietary viruses. For example, IAV was negatively correlated with an unclassified *Betaflexiviridae*, Papaya ringspot virus and unclassified sobemovirus. We hypothesize that this is observed, not because there is a biological inhibition of dietary viruses due to the presence of IAV, but rather that IAV is shed in greater abundance as compared to dietary viruses that may not replicate within the bat. This caveat may be an artifact of sequencing, in that longer virus genomes or more abundant populations can drown out the signal of less abundant viruses or those with shorter length genomes and make them less likely to be detected.

In carrier swabs (**Figure 38B**), unclassified bat rotavirus, unclassified bat orthoreovirus, mumps virus and novel bat paramyxovirus were each positively correlated with the others listed. The large number of positively-correlated viruses in carrier swabs is consistent with our hypothesis that head and body swabs are representative of viral populations that are present in the colony as a whole. Therefore, if both viruses are present in the colony, we would expect to consistently find them in bat head and body swabs across the colony. Taken together, we conclude that noninvasive surveillance methods that target the bodies of bats not only detect viruses shed within the colony, but also represent viral populations dispersed throughout the entire colony. As previously discussed, this finding could inform future sample collection.

OTHER POSSIBLE FUTURE ANALYSES

Bacterial OTU-based analysis among time points

Reads-based analysis of all domains of life as discussed in the beginning of this appendix should be followed-on by at least one orthogonal method to validate findings. One approach to evaluate changes in the microbiota with a specific focus on bacteria is to classify bacterial operational taxonomic units (OTU) and analyze differences among swabs and time points. OTUs utilize a marker gene as a proxy for species association and are frequently used to classify bacteria in complex metagenomic samples. In the same way that we were able to make colony-level observations of viral abundance over time, it would be informative to evaluate the bacterial OTU abundance over time in the captive bat colony. One caveat of this approach for our study design is that RNA sequencing was performed, and therefore it is possible that breadth of coverage across DNA genome OTUs might vary among samples.

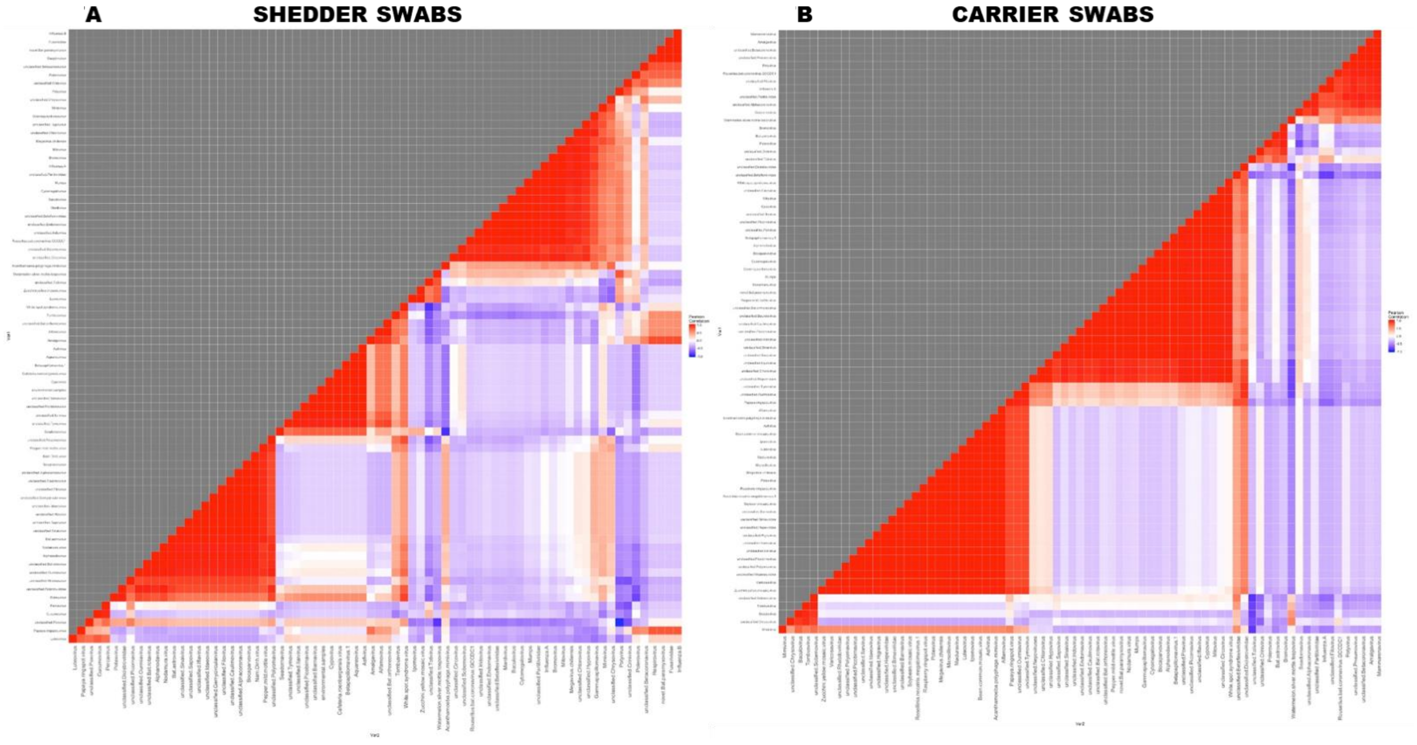


Figure 38. Viral community profiling using Pearson correlation coefficient analysis of unbiased shotgun sequencing data. Samples were analyzed together as shedder (oral, rectal) or carrier (head, body) swabs. Each graph shows the taxa along both axes, with the lower triangle representing Pearson correlate coefficients for each co-occurring pair of viruses (negative coefficient -1, blue; positive coefficient 1, red).

Clustering viral dark matter

As previously discussed, a portion of each sequencing run could not be classified due to the lack of a near neighbor in publicly-available sequence databases. It is important to save these data for the occasion when near neighbors or new technologies become available, allowing the reads or contigs to be classified. One potential approach to use those data is to cluster unclassifiable contigs using CD-HIT to identify commonalities among a dataset (109). Parameters such as nucleotide identity threshold can be adjusted in CD-HIT to cluster contigs or raw reads based on nucleotide similarity to each other. While this approach will not classify new organisms, it can help begin the process of genome discovery by enabling the identification of similar contigs that did not assemble into a single, long contig. Through manual analysis of contigs within clusters, new genomes or parts of new genomes could be discerned. Further analyses of these identified novel genomes could include annotation, functional prediction and comparison to distant sequenced relatives. This process requires intense manual analysis and experience, but could result in the discovery of very distant viruses. We hypothesize that applying these methods to contigs assembled from bat rectal swabs would uncover a phage genome similar to human crAssphage. The more viral genomes we uncover, the better we will understand the ecological context of the virosphere. The significance of expanding our knowledge of viruses touches on applications in both human medicine and environmental conservation.

ERROR RATE CALCULATIONS

Before all analysis, sequencing error rates within both shotgun and enriched samples were evaluated using bbmap (19). The range of error was 0.34-1.65% with an average of 0.738%. The MiSeq data error rate was higher than NextSeq data, which is consistent with the literature (115). Raw data is shown in Table 14.

In addition to error rate, we evaluated the effect of the quality control portion of our pipeline on read counts per sequencing batch (Figure 39). Low-quality reads with a Q-score less than 20, and reads that were identified to have an exact duplicated match, were removed using clumpify.sh, leaving a smaller and cleaner dataset to be trimmed and filtered for quality (1; 19). The "cleaned" datasets consisted of less than 20 million reads per batch of 24 samples. The total read count per batch was relatively uniform after discarding low quality reads.

Table 14. Raw data for PhiX error rates of both trimmed and untrimmed datasets obtained from NextSeq and MiSeq batches.

Sequencing platform	**Batch**	**Untrimmed reads**		**Trimmed reads**		**Average error after trimming (phiX reads only)**	**Average error**
		R1	**R2**	**R1**	**R2**		
NextSeq	1	0.644	1.34	0.422	0.613	0.51725	0.47073333
NextSeq	2	0.7	1.37	0.46	0.588	0.52395	
NextSeq	3	0.74	1.196	0.495	0.564	0.4902	
NextSeq	4	0.809	1.377	0.513	0.64	0.57685	
NextSeq	5	0.822	1.343	0.434	0.398	0.41575	
NextSeq	6	1.201	0.674	0.598	0.352	0.4748	
NextSeq	7	0.616	0.984	0.515	0.59	0.5525	
NextSeq	8	0.436	0.512	0.355	0.342	0.34845	
NextSeq	9	0.41	0.458	0.349	0.325	0.33685	
MiSeq	1	2.546	2.168	2.506	0.349	1.4277	0.9392625
MiSeq	2	2503	3.702	1.527	0.273	0.8997	
MiSeq	3	2.634	1.342	2.922	0.357	1.63935	
MiSeq	4	2.15	1.358	2.268	0.303	1.28545	
MiSeq	5	1.017	0.81	1.288	0.249	0.76865	
MiSeq	6	1.06	1.415	1.255	0.266	0.76055	
MiSeq	7	1.1	1.051	1.373	0.273	0.82255	
MiSeq	8	1.19	1.281	1.145	0.312	0.72835	
MiSeq	9	1.196	0.973	1.195	0.276	0.7355	
MiSeq	10	1.172	0.964	1.219	0.261	0.74015	
MiSeq	11	1.499	1.47	1.082	0.284	0.683	
MiSeq	12	1.397	1.589	1.204	0.357	0.7802	

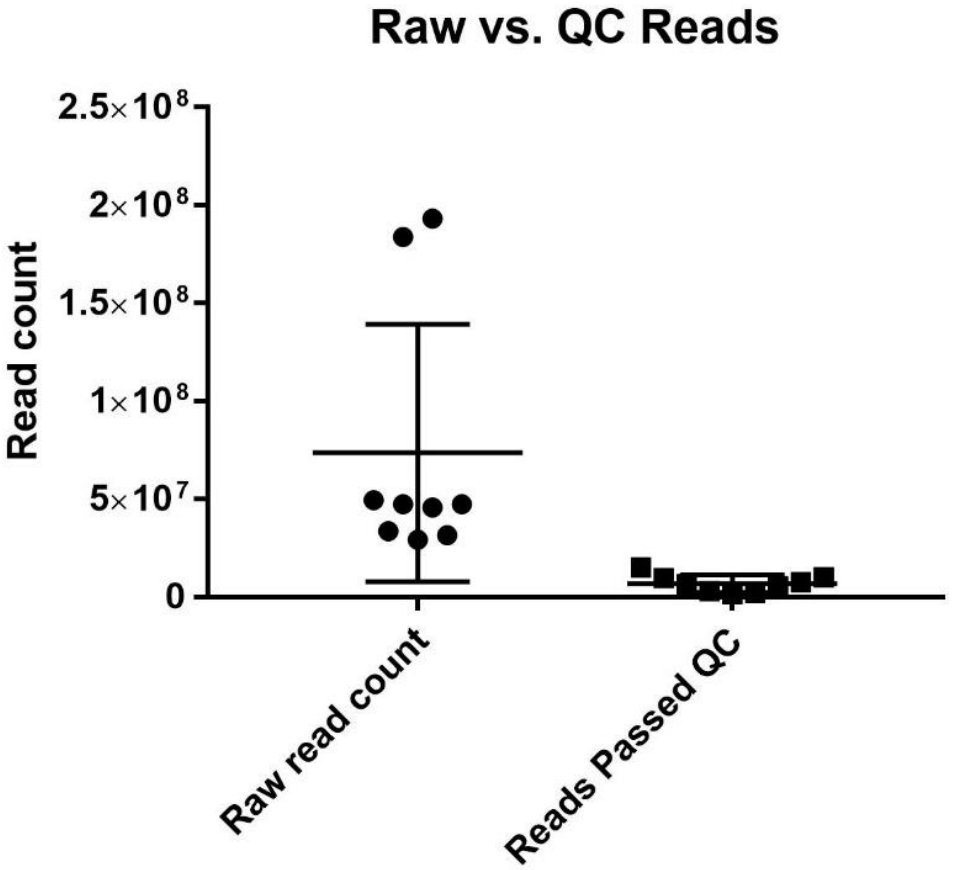


Figure 39. Quality control of reads minimizes inter-run variability. The total read count per batch was normalized by discarding low-quality reads.

REFERENCES

1. 2019. seqtk. *https://github.com/lh3/seqtk*

2. Abernathy E, Chen MH, Bera J, Shrivastava S, Kirkness E, *et al.* 2013. Analysis of whole genome sequences of 16 strains of rubella virus from the United States, 1961-2009. Virol J 10:32

3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-10

4. Amman BR, Jones ME, Sealy TK, Uebelhoer LS, Schuh AJ, *et al.* 2015. Oral shedding of Marburg virus in experimentally infected Egyptian fruit bats (Rousettus aegyptiacus). J Wildl Dis 51:113-24

5. Amman BR, Nyakarahuka L, McElroy AK, Dodd KA, Sealy TK, *et al.* 2014. Marburgvirus resurgence in Kitaka Mine bat population after extermination attempts, Uganda. Emerg Infect Dis 20:1761-4

6. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. 370

7. Anthony SJ, Johnson CK, Greig DJ, Kramer S, Che X, *et al.* 2017. Global patterns in coronavirus diversity. Virus Evol 3:vex012

8. Bakker KM, Rocke TE, Osorio JE, Abbott RC, Tello C, *et al.* 2019. Fluorescent biomarkers demonstrate prospects for spreadable vaccines to control disease transmission in wild bats. Nat Ecol Evol 3:1697-704

9. Banerjee A, Kulcsar K, Misra V, Frieman M, Mossman K. 2019. Bats and Coronaviruses. Viruses 11(1):41.

10. Banyard AC, Evans JS, Luo TR, Fooks AR. 2014. Lyssaviruses and bats: emergence and zoonotic threat. Viruses 6:2974-90

11. Behura SK, Severson DW. 2013. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol Rev Camb Philos Soc 88:49-61

12. Benkert P, Biasini M, Schwede T. 2011. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27:343-50

13. Bennett AJ, Bushmaker T, Cameron K, Ondzie A, Niama FR, *et al.* 2019. Diverse RNA viruses of arthropod origin in the blood of fruit bats suggest a link between bat and arthropod viromes. Virology 528:64-72

14. Bharadwaj SD, Sahay RR, Yadav PD, Dhanawade S, Basu A, *et al.* 2018. Acute encephalitis with atypical presentation of rubella in family cluster, India. Emerg Infect Dis 24:1923-5

15. Biasini M. 2015. PV - JavaScript Protein Viewer v. 1.8.1.

16. Blackley DJ, Wiley MR, Ladner JT, Fallah M, Lo T, *et al.* 2016. Reduced evolutionary rate in reemerged Ebola virus transmission chains. Sci Adv 2:e1600378

17. Brown JR, Roy S, Ruis C, Yara Romero E, Shah D, *et al.* 2016. Norovirus Whole-Genome Sequencing by SureSelect Target Enrichment: a Robust and Sensitive Method. J Clin Microbiol 54:2530-7

18. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59-60

19. Bushnell B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner.

20. Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T. 2006. Bats: important reservoir hosts of emerging viruses. Clin Microbiol Rev 19:531-45

21. Callahan JD, Wu SJ, Dion-Schultz A, Mangold BE, Peruski LF, *et al.* 2001. Development and evaluation of serotype- and group-specific fluorogenic reverse transcriptase PCR (TaqMan) assays for dengue virus. Journal of clinical microbiology 39:4119-24

22. Carlson CJ, Zipfel CM, Garnier R, Bansal S. 2019. Global estimates of mammalian viral diversity accounting for host sharing. Nat Ecol Evol 3:1070-5

23. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, *et al.* 2018. The Global Virome Project. Science 359:872-4

24. Charif D, Thioulouse J, Lobry JR, Perriere G. 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21:545-7

25. Chen L, Liu B, Yang J, Jin Q. 2014. DBatVir: the database of bat-associated viruses. Database (Oxford) 2014:bau021

26. Chionh YT, Cui J, Koh J, Mendenhall IH, Ng JHJ, *et al.* 2019. High basal heat-shock protein expression in bats confers resistance to cellular heat/oxidative stress. Cell Stress Chaperones 24:835-49

27. Cooper LZ, Krugman S. 1966. Diagnosis and management: congenital rubella. Pediatrics 37:335-8

28. Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, *et al.* 2014. Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. J Virol 88:11297-303

29. Corman VM, Rasche A, Baronti C, Aldabbagh S, Cadar D, *et al.* 2016. Assay optimization for molecular detection of Zika virus. Bull World Health Organ 94:880-92

30. Cruz CD, Torre A, Troncos G, Lambrechts L, Leguia M. 2016. Targeted full-genome amplification and sequencing of dengue virus types 1-4 from South America. J Virol Methods 235:158-67

31. Cui J, Li F, Shi ZL. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181-92

32. Cummings MJ, Tokarz R, Bakamutumaho B, Kayiwa J, Byaruhanga T, *et al.* 2018. Precision surveillance for viral respiratory pathogens: virome capture sequencing for the detection and genomic characterization of severe acute respiratory infection in Uganda. Clin Infect Dis 68(7):1118-25.

33. Cunningham AA, Daszak P, Wood JLN. 2017. One Health, emerging infectious diseases and wildlife: two decades of progress? Philos Trans R Soc Lond B Biol Sci 372(1725):20160167

34. Daszak P, Cunningham AA, Hyatt AD. 2000. Emerging infectious diseases of wildlife--threats to biodiversity and human health. Science 287:443-9

35. Davidson I, Silva RF. 2008. Creation of diversity in the animal virus world by inter-species and intra-species recombinations: lessons learned from poultry viruses. Virus Genes 36:1-9

36. de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, *et al.* 2013. Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. J Virol 87:7790-2

37. de Vries RD, Duprex WP, de Swart RL. 2015. Morbillivirus infections: an introduction. Viruses 7:699-706

38. Decher J, Fahr J. 2005. Hipposideros cyclops. Mammalian Species:1(763):1-7.

39. Dehority WN, Eickman MM, Schwalm KC, Gross SM, Schroth GP, *et al.* 2017. Complete genome sequence of a KI polyomavirus isolated from an otherwise healthy child with severe lower respiratory tract infection. J Med Virol 89:926-30

40. Dixon P. 2003. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science 14.6:927-30

41. Doan T, Wilson MR, Crawford ED, Chow ED, Khan LM, *et al.* 2016. Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. Genome Med 8(1):90

42. Dovih P, Laing ED, Chen Y, Low DHW, Ansil BR, *et al.* 2019. Filovirus-reactive antibodies in humans and bats in Northeast India imply zoonotic spillover. PLoS Negl Trop Dis 13:e0007733

43. Drexler JF, Corman VM, Muller MA, Maganga GD, Vallo P*, et al.* 2012. Bats host major mammalian paramyxoviruses. Nat Commun 3:796

44. DuBois RM, Vaney MC, Tortorici MA, Kurdi RA, Barba-Spaeth G, *et al.* 2013. Functional and evolutionary insight from the crystal structure of rubella virus protein E1. Nature 493:552-6

45. Dumont ER. 1997. Salivary pH and buffering capacity in frugivorous and insectivorous bats. Journal of mammalogy 78:1210-9

46. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, *et al.* 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5:4498

47. Dutilh BE, Schmieder R, Nulton J, Felts B, Salamon P, *et al.* 2012. Reference-independent comparative metagenomics using cross-assembly: crAss. Bioinformatics 28:3225-31

48. Edmunds WJ, Gay NJ, Kretzschmar M, Pebody RG, Wachmann H, Network EPES-e. 2000. The pre-vaccination epidemiology of measles, mumps and rubella in Europe: implications for modelling studies. Epidemiol Infect 125:635-50

49. Faye O, Faye O, Diallo D, Diallo M, Weidmann M, Sall AA. 2013. Quantitative real-time PCR detection of Zika virus and evaluation with field-caught mosquitoes. Virol J 10:311

50. Fenner F, Pereira HG, Porterfield JS, Joklik WK, Downie AW. 1974. Family and generic names for viruses approved by the International Committee on Taxonomy of Viruses, June 1974. Intervirology 3:193-8

51. Francis C, Rosell-Ambal, G., Tabaranza, B., Carino, P., Helgen, K., Molur, S. & Srinivasulu, C. 2008. The IUCN Red List of Threatened Species 2008. e.T7787A12850087

52. Freer G, Maggi F, Pifferi M, Di Cicco ME, Peroni DG, Pistello M. 2018. The Virome and Its Major Component, Anellovirus, a Convoluted System Molding Human Immune Defenses and Possibly Affecting the Development of Asthma and Respiratory Diseases in Childhood. Front Microbiol 9:686

53. French RK, Holmes EC. 2019. An Ecosystems Perspective on Virus Evolution and Emergence. Trends Microbiol. 3(3):165-175

54. Frey KG, Bishop-Lilly KA. 2015. Next-generation sequencing for pathogen detection and identification. In Methods in Microbiology, 42:525-54: Elsevier. Number of 525-54 pp.

55. Frey KG, Herrera-Galeano JE, Redden CL, Luu TV, Servetas SL, *et al.* 2014. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. BMC Genomics 15(1):96

56. Frey TK. 1997. Neurological aspects of rubella virus infection. Intervirology 40:167-75

57. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150-2

58. Gentile G, Micozzi A. 2016. Speculations on the clinical significance of asymptomatic viral infections. Clin Microbiol Infect 22:585-8

59. Geoghegan JL, Duchene S, Holmes EC. 2017. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. PLoS Pathog 13:e1006215

60. Geoghegan JL, Pirotta V, Harvey E, Smith A, Buchmann JP, *et al.* 2018. Virological Sampling of Inaccessible Wildlife with Drones. Viruses 10(6):300

61. Geoghegan JL, Senior AM, Di Giallonardo F, Holmes EC. 2016. Virological factors that increase the transmissibility of emerging human viruses. Proc Natl Acad Sci U S A 113:4170-5

62. Gohl DM, Magli A, Garbe J, Becker A, Johnson DM, *et al.* 2019. Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification. Genome Biol 20(1):1-7

63. Goldstein T, Anthony SJ, Gbakima A, Bird BH, Bangura J, *et al.* 2018. The discovery of Bombali virus adds further support for bats as hosts of ebolaviruses. Nat Microbiol 3:1084-9

64. Gonzalez JM, Gomez-Puertas P, Cavanagh D, Gorbalenya AE, Enjuanes L. 2003. A comparative sequence analysis to revise the current taxonomy of the family Coronaviridae. Arch Virol 148:2207-35

65. Grant GB, Masresha BG, Moss WJ, Mulders MN, Rota PA, *et al.* 2019. Accelerating measles and rubella elimination through research and innovation - findings from the Measles & Rubella Initiative research prioritization process, 2016. Vaccine 37:5754-61

66. Grant GB, Reef SE, Patel M, Knapp JK, Dabbagh A. 2017. Progress in Rubella and Congenital Rubella Syndrome Control and Elimination - Worldwide, 2000-2016. MMWR Morb Mortal Wkly Rep 66:1256-60

67. Gregory AC, Zayed AA, Conceicao-Neto N, Temperton B, Bolduc B, *et al.* 2019. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell 177:1109-23 e14

68. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, *et al.* 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature 546:401-5

69. Halpin K, Hyatt AD, Fogarty R, Middleton D, Bingham J, *et al.* 2011. Pteropid bats are confirmed as the reservoir hosts of henipaviruses: a comprehensive experimental study of virus transmission. Am J Trop Med Hyg 85:946-51

70. Han BA, Schmidt JP, Alexander LW, Bowden SE, Hayman DT, Drake JM. 2016. Undiscovered Bat Hosts of Filoviruses. PLoS Negl Trop Dis 10:e0004815

71. Hassell JM, Begon M, Ward MJ, Fevre EM. 2017. Urbanization and Disease Emergence: Dynamics at the Wildlife-Livestock-Human Interface. Trends Ecol Evol 32:55-67

72. Hayes S, Mahony J, Nauta A, van Sinderen D. 2017. Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. Viruses 9(6):127

73. Hayman DT, Bowen RA, Cryan PM, McCracken GF, O'Shea TJ, *et al.* 2013. Ecology of zoonotic infectious diseases in bats: current knowledge and future directions. Zoonoses Public Health 60:2-21

74. Heaney LR. 1998. A synopsis of the mammalian fauna of the Philippine Islands. Fieldiana. Zoology. New Series 88:1-61

75. Ho T, Tzanetakis IE. 2014. Development of a virus detection and discovery pipeline using next generation sequencing. Virology 471-473:54-60

76. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35:518-22

77. Hobman TC, Gillam S. 1989. In vitro and in vivo expression of rubella virus glycoprotein E2: the signal peptide is contained in the C-terminal region of capsid protein. Virology 173:241-50

78. Hodgkin J. 2001. Genome Size. 865

79. Hu B, Zeng LP, Yang XL, Ge XY, Zhang W, *et al.* 2017. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog 13:e1006698

80. Huang C, Liu WJ, Xu W, Jin T, Zhao Y, *et al.* 2016. A Bat-Derived Putative Cross-Family Recombinant Coronavirus with a Reovirus Gene. PLoS Pathog 12:e1005883

81. Huson DH, Beier S, Flade I, Gorska A, El-Hadidi M, *et al.* 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol 12:e1004957

82. ICTV. 2018b. The International Code of Virus Classification and Nomenclature. International Committee on Taxonomy of Viruses. 2018. Archives of virology 163(9):2601-31

83. Illumina. 2014. TruSeq RNA Access Library Prep Guide. In: Cold Spring Harbor Protocols. 31 March 2014 edn; 2014

84. Ilunga Kalenga O, Moeti M, Sparrow A, Nguyen VK, Lucey D, Ghebreyesus TA. 2019. The Ongoing Ebola Epidemic in the Democratic Republic of Congo, 2018-2019. N Engl J Med 381(4):373-83

85. Kandeil A, Gomaa MR, Shehata MM, El Taweel AN, Mahmoud SH, *et al.* 2019. Isolation and Characterization of a Distinct Influenza A Virus from Egyptian Bats. J Virol 93(2):e01059-18

86. Kaneko H, Iida T, Ishiko H, Ohguchi T, Ariga T, *et al.* 2009. Analysis of the complete genome sequence of epidemic keratoconjunctivitis-related human adenovirus type 8, 19, 37 and a novel serotype. J Gen Virol 90:1471-6

87. Karakus U, Thamamongood T, Ciminski K, Ran W, Gunther SC, *et al.* 2019. MHC class II proteins mediate cross-species entry of bat influenza viruses. Nature 567:109-12

88. Kassambara A, Mundt F. 2017. Package ‘factoextra’. Extract and visualize the results of multivariate data analyses 76

89. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845-58

90. Kolodny O, Weinberg M, Reshef L, Harten L, Hefetz A, *et al.* 2019. Coordinated change at the colony level in fruit bat fur microbiomes through time. Nat Ecol Evol 3:116-24

91. Korber B. 2000. Computational Analysis of HIV Molecular Sequences. Kluwer Academic Publishers:55-72

92. Krishnamurthy SR, Wang D. 2017. Origins and challenges of viral dark matter. Virus Res 239:136-42

93. Kuhn JH, Wolf YI, Krupovic M, Zhang YZ, Maes P, *et al.* 2019. Classify viruses - the gain is worth the pain. Nature 566:318-20

94. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870-4

95. Kunz TH, Braun de Torrez E, Bauer D, Lobova T, Fleming TH. 2011. Ecosystem services provided by bats. Ann N Y Acad Sci 1223:1-38

96. Kunz TH, Fenton MB, editors. 2005. Bat ecology. University of Chicago Press p4

97. Lacroix A, Duong V, Hul V, San S, Davun H, *et al.* 2017. Genetic diversity of coronaviruses in bats in Lao PDR and Cambodia. Infect Genet Evol 48:10-8

98. Laing ED, Mendenhall IH, Linster M, Low DHW, Chen Y, *et al.* 2018. Serologic Evidence of Fruit Bat Exposure to Filoviruses, Singapore, 2011-2016. Emerg Infect Dis 24:114-7

99. Lambert N, Strebel P, Orenstein W, Icenogle J, Poland GA. 2015. Rubella. Lancet 385:2297-307

100. Lanciotti RS, Kosoy OL, Laven JJ, Velez JO, Lambert AJ, *et al.* 2008. Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. Emerg Infect Dis 14:1232-9

101. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-9

102. Lau SK, Poon RW, Wong BH, Wang M, Huang Y, *et al.* 2010. Coexistence of different genotypes in the same bat and serological characterization of Rousettus bat coronavirus HKU9 belonging to a novel Betacoronavirus subgroup. J Virol 84:11385-94

103. Layeghifard M, Hwang DM, Guttman DS. 2017. Disentangling Interactions in the Microbiome: A Network Perspective. Trends Microbiol 25:217-28

104. Lefeuvre P, Martin DP, Elena SF, Shepherd DN, Roumagnac P, Varsani A. 2019. Evolution and ecology of plant viruses. Nat Rev Microbiol 17:632-44

105. Leguia M, Cruz CD, Felices V, Torre A, Troncos G, *et al.* 2017. Full-genome amplification and sequencing of Zika viruses using a targeted amplification approach. J Virol Methods 248:77-82

106. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. 2011. The sequence read archive. Nucleic Acids Res 39:D19-21

107. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-60

108. Li PE, Lo CC, Anderson JJ, Davenport KW, Bishop-Lilly KA, *et al.* 2017. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. Nucleic Acids Res 45:67-80

109. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-9

110. Li W, Shi Z, Yu M, Ren W, Smith C, *et al.* 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310:676-9

111. Lu R, Zhao X, Li J, Niu P, Yang B, *et al.* 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet 395(10224):565-574

112. Luis AD, Hayman DT, O'Shea TJ, Cryan PM, Gilbert AT, *et al.* 2013. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proc Biol Sci 280:20122753

113. Luke T, Bennett RS, Gerhardt DM, Burdette T, Postnikova E, *et al.* 2018. Fully Human Immunoglobulin G From Transchromosomic Bovines Treats Nonhuman Primates Infected With Ebola Virus Makona Isolate. J Infect Dis 218(suppl_5):S636-48

114. Luo Y, Li B, Jiang RD, Hu BJ, Luo DS, *et al.* 2018. Longitudinal Surveillance of Betacoronaviruses in Fruit Bats in Yunnan Province, China During 2009-2016. Virol Sin 33:87-95

115. Manley LJ, Ma D, Levine SS. 2016. Monitoring Error Rates In Illumina Sequencing. J Biomol Tech 27:125-8

116. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, *et al.* 2015. Molecular Evidence of Sexual Transmission of Ebola Virus. N Engl J Med 373:2448-54

117. Maton WG. 1815. Some account of a rash liable to be mistaken for scarlatina. Medical Transactions of the Royal College of Physicians 5:149-65

118. McLaren MR, Willis AD, Callahan BJ. 2019. Consistent and correctable bias in metagenomic sequencing experiments. Elife 8;e46923

119. Mehand MS, Al-Shorbaji F, Millett P, Murgue B. 2018. The WHO R&D Blueprint: 2018 review of emerging infectious diseases requiring urgent research and development efforts. Antiviral Res 159:63-7

120. Menachery VD, Yount BL, Jr., Debbink K, Agnihothram S, Gralinski LE, *et al.* 2015. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med 21:1508-13

121. Mendenhall IH, Borthwick S, Neves ES, Low D, Linster M, *et al.* 2016. Identification of a Lineage D Betacoronavirus in Cave Nectar Bats (*Eonycteris spelaea*) in Singapore and an Overview of Lineage D Reservoir Ecology in SE Asian Bats. Transbound Emerg Dis 64(6):1790-800

122. Mendenhall IH, Borthwick S, Neves ES, Low D, Linster M, *et al.* 2017. Identification of a Lineage D Betacoronavirus in Cave Nectar Bats (*Eonycteris spelaea*) in Singapore and an Overview of Lineage D Reservoir Ecology in SE Asian Bats. Transbound Emerg Dis 64:1790-800

123. Mendenhall IH, Skiles MM, Neves ES, Borthwick SA, Low DHW, *et al.* 2017. Influence of age and body condition on astrovirus infection of bats in Singapore: An evolutionary and epidemiological analysis. One Health 4:27-33

124. Mendenhall IH, Wen DLH, Jayakumar J, Gunalan V, Wang L, *et al.* 2019. Diversity and Evolution of Viral Pathogen Community in Cave Nectar Bats (*Eonycteris spelaea*). Viruses 11(3):250

125. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, *et al.* 2017. Zika virus evolution and spread in the Americas. Nature 546(7658):411-5

126. Metsky HC, Siddle KJ, Gladden-Young A, Qu J, Yang DK, *et al.* 2019. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. Nat Biotechnol 37:160-8

127. Metsky HC, Siddle KJ, Gladden-Young A, Qu J, Yang DK, *et al.* 2018. Capturing diverse microbial sequence with comprehensive and scalable probe design. In bioRxiv (preprint)

128. Mortlock M, Dietrich M, Weyer J, Paweska JT, Markotter W. 2019. Co-Circulation and Excretion Dynamics of Diverse Rubula- and Related Viruses in Egyptian Rousette Bats from South Africa. Viruses 11(1):37

129. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, *et al.* 2016. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 17:53

130. Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. Oxford university press. Chapter 9.

131. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-74

132. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824-34

133. O'Flaherty BM, Li Y, Tao Y, Paden CR, Queen K, *et al.* 2018. Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. Genome Res 28:869-77

134. Oba M, Tsuchiaka S, Omatsu T, Katayama Y, Otomaru K, *et al.* 2018. A new comprehensive method for detection of livestock-related pathogenic viruses using a target enrichment system. Biochem Biophys Res Commun 495:1871-7

135. Obameso JO, Li H, Jia H, Han M, Zhu S, *et al.* 2017. The persistent prevalence and evolution of cross-family recombinant coronavirus GCCDC1 among a bat population: a two-year follow-up. Sci China Life Sci 60:1357-63

136. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. 2017. Host and viral traits predict zoonotic spillover from mammals. Nature 546:646-50

137. Paixao ES, Teixeira MG, Rodrigues LC. 2018. Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases. BMJ Glob Health 3:e000530

138. Panning M, Grywna K, van Esbroeck M, Emmerich P, Drosten C. 2008. Chikungunya fever in travelers returning to Europe from the Indian Ocean region, 2006. Emerg Infect Dis 14:416-22

139. Parker MT. 2016. An Ecological Framework of the Human Virome Provides Classification of Current Knowledge and Identifies Areas of Forthcoming Discovery. Yale J Biol Med 89:339-51

140. Paskey AC, Frey KG, Schroth G, Gross S, Hamilton T, Bishop-Lilly KA. 2019. Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples. BMC Genomics 20:155

141. Pearson K. 1895. Notes on Regression and Inheritance in the Case of Two Parents Proceedings of the Royal Society of London, 58, 240-242.

142. Peterson AT, Carroll DS, Mills JN, Johnson KM. 2004. Potential mammalian filovirus reservoirs. Emerg Infect Dis 10:2073-81

143. Phelps KL, Hamel L, Alhmoud N, Ali S, Bilgin R, *et al.* 2019. Bat Research Networks and Viral Surveillance: Gaps and Opportunities in Western Asia. Viruses 11(3):240

144. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, *et al.* 2012. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 40:D593-8

145. Plowright RK, Eby P, Hudson PJ, Smith IL, Westcott D, *et al.* 2015. Ecological dynamics of emerging bat virus spillover. Proc Biol Sci 282:20142124

146. Plowright RK, Foley P, Field HE, Dobson AP, Foley JE, *et al.* 2011. Urban habituation, ecological connectivity and epidemic dampening: the emergence of Hendra virus from flying foxes (Pteropus spp.). Proc Biol Sci 278:3703-12

147. Plowright RK, Peel AJ, Streicker DG, Gilbert AT, McCallum H, *et al.* 2016. Transmission or Within-Host Dynamics Driving Pulses of Zoonotic Viruses in Reservoir-Host Populations. PLoS Negl Trop Dis 10:e0004796

148. Poon LL, Chu DK, Chan KH, Wong OK, Ellis TM, *et al.* 2005. Identification of a novel coronavirus in bats. J Virol 79:2001-9

149. Pustowoit B, Liebert UG. 1998. Predictive value of serological tests in rubella virus infection during pregnancy. Intervirology 41:170-7

150. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, *et al.* 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 12:1261-76

151. Rascovan N, Duraisamy R, Desnues C. 2016. Metagenomics and the Human Virome in Asymptomatic Individuals. Annu Rev Microbiol 70:125-41

152. Rubing Chen SM, Andres Merits, Bethany Bolling, Farooq Nasar, Lark L. Coffey, Ann Powers, Scott C. Weaver, Donald Smith, Peter Simmonds and Stuart Siddell. 2018. Create a new family Matonaviridae to include the genus Rubivirus, removed from the family *Togaviridae*. Approved ICTV Report 2018.013S.R.Matonaviridae

153. Ruiz Silva M, Aguilar Briseno JA, Upasani V, van der Ende-Metselaar H, Smit JM, Rodenhuis-Zybert IA. 2017. Suppression of chikungunya virus replication and differential innate responses of human peripheral blood mononuclear cells during co-infection with dengue virus. PLoS Negl Trop Dis 11:e0005712

154. Salmier A, Tirera S, de Thoisy B, Franc A, Darcissac E, *et al.* 2017. Virome analysis of two sympatric bat species (Desmodus rotundus and Molossus molossus) in French Guiana. PLoS One 12:e0186943

155. Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, *et al.* 2015. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. Hum Mutat 36:903-14

156. Sanitá Lima M, Smith DR. 2017. Don't just dump your data and run: Authors should submit as much experimental information as possible when uploading sequence data. EMBO reports 18:2087-9

157. Scheuch M, Hoper D, Beer M. 2015. RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinformatics 16:69

158. Schountz T, Baker ML, Butler J, Munster V. 2017. Immunological Control of Viral Infections in Bats and the Emergence of Viruses Highly Pathogenic to Humans. Front Immunol 8:1098

159. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, *et al.* 2016. Redefining the invertebrate RNA virosphere. Nature 540(7634):539-43

160. Smith AT, Xie Y, Hoffmann RS, Lunde D, MacKinnon J, *et al.* 2010. A guide to the mammals of China. Princeton University Press pp 327-349

161. Solonenko SA, Ignacio-Espinoza JC, Alberti A, Cruaud C, Hallam S, *et al.* 2013. Sequencing platform and library preparation choices impact viral metagenomes. BMC Genomics 14:320

162. Sozhamannan S, Holland MY, Hall AT, Negron DA, Ivancich M, *et al.* 2015. Evaluation of Signature Erosion in Ebola Virus Due to Genomic Drift and Its Impact on the Performance of Diagnostic Assays. Viruses 7:3130-54

163. Stoesser G, Sterk P, Tuli MA, Stoehr PJ, Cameron GN. 1997. The EMBL Nucleotide Sequence Database. Nucleic Acids Res 25:7-14

164. Su S, Wong G, Shi W, Liu J, Lai ACK, *et al.* 2016. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol 24:490-502

165. Subudhi S, Rapin N, Misra V. 2019. Immune System Modulation and Viral Persistence in Bats: Understanding Viral Spillover. Viruses 11(2):192

166. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. 2019. Choice of assembly software has a critical impact on virome characterisation. Microbiome 7:12

167. Team RC. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing

168. Tennekes M. 2018. tmap: Thematic Maps in R. Journal of Statistical Software 84:1-39

169. Thomas SP, Suthers RA. 1972. The physiology and energetics of bat flight. Journal of Experimental Biology 57:317-35

170. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. 2009. Laboratory procedures to generate viral metagenomes. Nat Protoc 4:470-83

171. Tong S, Li Y, Rivailler P, Conrardy C, Castillo DA, *et al.* 2012. A distinct lineage of influenza A virus from bats. Proc Natl Acad Sci U S A 109:4269-74

172. Tong S, Zhu X, Li Y, Shi M, Zhang J, *et al.* 2013. New world bats harbor diverse influenza A viruses. PLoS Pathog 9:e1003657

173. Tyrrell DA, Bynoe ML. 1966. Cultivation of viruses from a high proportion of patients with colds. Lancet 1:76-7

174. van Boheemen S, de Graaf M, Lauber C, Bestebroer TM, Raj VS, *et al.* 2012. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. MBio 3(6):e00473-12

175. van der Schaar HM, Rust MJ, Waarts BL, van der Ende-Metselaar H, Kuhn RJ, *et al.* 2007. Characterization of the early events in dengue virus cell entry by biochemical assays and single-virus tracking. J Virol 81:12019-28

176. van Duijl-Richter MKS, Blijleven JS, van Oijen AM, Smit JM. 2015. Chikungunya virus fusion properties elucidated by single-particle and bulk approaches. J Gen Virol 96:2122-32

177. Vilgalys R. 2003. Taxonomic misidentification in public DNA databases. New Phytologist 160:4-5

178. Voigt CC, Kingston T. 2016. Bats in the Anthropocene. In Bats in the Anthropocene: Conservation of Bats in a Changing World, ed. CC Voigt, T Kingston:1-9. Cham: Springer International Publishing. Number of 1-9 pp.

179. Walker JM. 2005. The proteomics protocols handbook. Humana Press pp 571-607

180. Wang L-f, Cowled C. 2015. Bats and viruses: a new frontier of emerging infectious diseases. John Wiley & Sons

181. Wang LF, Anderson DE. 2019. Viruses in bats and potential spillover to animals and humans. Curr Opin Virol 34:79-89

182. Wang WK, Sung TL, Tsai YC, Kao CL, Chang SM, King CC. 2002. Detection of dengue virus replication in peripheral blood mononuclear cells from dengue virus type 2-infected patients by a reverse transcription-real-time PCR assay. J Clin Microbiol 40:4472-8

183. Watanabe S, Masangkay JS, Nagata N, Morikawa S, Mizutani T, *et al.* 2010. Bat coronaviruses and experimental infection of bats, the Philippines. Emerg Infect Dis 16:1217-23

184. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, *et al.* 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296-W303

185. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. Microbiol Rev 56:152-79

186. Wen M, Ng JHJ, Zhu F, Chionh YT, Chia WN, *et al.* 2018. Exploring the genome and transcriptome of the cave nectar bat *Eonycteris spelaea* with PacBio long-read sequencing. Gigascience 7(10):giy116

187. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag

188. Wimsatt W. 2012. Biology of bats. Elsevier pp 11

189. Wolf YI, Kazlauskas D, Iranzo J, Lucia-Sanz A, Kuhn JH, *et al.* 2018. Origins and Evolution of the Global RNA Virome. MBio 9(6):e02329-18

190. Wong AH, Cheng PK, Lai MY, Leung PC, Wong KK, *et al.* 2012. Virulence potential of fusogenic orthoreoviruses. Emerg Infect Dis 18:944-8

191. Woo PC, Huang Y, Lau SK, Yuen KY. 2010. Coronavirus genomics and bioinformatics analysis. Viruses 2:1804-20

192. Woo PC, Lau SK, Lam CS, Lai KK, Huang Y, *et al.* 2009. Comparative analysis of complete genome sequences of three avian coronaviruses reveals a novel group 3c coronavirus. J Virol 83:908-17

193. Woo PC, Lau SK, Lam CS, Lau CC, Tsang AK, *et al.* 2012. Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J Virol 86:3995-4008

194. Woo PC, Lau SK, Li KS, Tsang AK, Yuen KY. 2012. Genetic relatedness of the novel human group C betacoronavirus to Tylonycteris bat coronavirus HKU4 and Pipistrellus bat coronavirus HKU5. Emerg Microbes Infect 1:e35

195. Woo PC, Wang M, Lau SK, Xu H, Poon RW, *et al.* 2007. Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. J Virol 81:1574-85

196. Woolhouse ME, Brierley L, McCaffery C, Lycett S. 2016. Assessing the Epidemic Potential of RNA and DNA Viruses. Emerg Infect Dis 22:2037-44

197. Woolhouse ME, Gowtage-Sequeria S. 2005. Host range and emerging and reemerging pathogens. Emerg Infect Dis 11:1842-7

198. Wu Y, Wu Y, Tefsen B, Shi Y, Gao GF. 2014. Bat-derived influenza-like viruses H17N10 and H18N11. Trends Microbiol 22:183-91

199. Wylezich C, Papa A, Beer M, Hoper D. 2018. A Versatile Sample Processing Workflow for Metagenomic Pathogen Detection. Sci Rep 8:13108

200. Wylie TN, Wylie KM, Herter BN, Storch GA. 2015. Enhanced virome sequencing using targeted sequence capture. Genome Res 25:1910-20

201. Wynne JW, Wang LF. 2013. Bats and viruses: friend or foe? PLoS Pathog 9:e1003651

202. Yang XL, Tan CW, Anderson DE, Jiang RD, Li B, *et al.* 2019. Characterization of a filovirus (Mengla virus) from Rousettus bats in China. Nat Microbiol 4:390-5

203. Yang XL, Zhang YZ, Jiang RD, Guo H, Zhang W, *et al.* 2017. Genetically Diverse Filoviruses in Rousettus and Eonycteris spp. Bats, China, 2009 and 2015. Emerg Infect Dis 23:482-6

204. Yang Y, Walls SD, Gross SM, Schroth GP, Jarman RG, Hang J. 2018. Targeted Sequencing of Respiratory Viruses in Clinical Specimens for Pathogen Identification and Genome-Wide Analysis. Methods Mol Biol 1838:125-40

205. Yinda CK, Ghogomu SM, Conceicao-Neto N, Beller L, Deboutte W, *et al.* 2018. Cameroonian fruit bats harbor divergent viruses, including rotavirus H, bastroviruses, and picobirnaviruses using an alternative genetic code. Virus Evol 4:vey008

206. Young CC, Olival KJ. 2016. Optimizing Viral Discovery in Bats. PLoS One 11:e0149237

207. Yuan L, Li M, Li L, Monagin C, Chmura AA, *et al.* 2014. Evidence for retrovirus and paramyxovirus infection of multiple bat species in china. Viruses 6:2138-54

208. Zhang YZ, Shi M, Holmes EC. 2018. Using Metagenomics to Characterize an Expanding Virosphere. Cell 172:1168-72

209. Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, *et al.* 2017. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology 503:21-30

210. Zhou P, Fan H, Lan T, Yang XL, Shi WF, *et al.* 2018. Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. Nature 556:255-8

211. Zhou P, Tachedjian M, Wynne JW, Boyd V, Cui J, *et al.* 2016. Contraction of the type I IFN locus and unusual constitutive expression of IFN-alpha in bats. Proc Natl Acad Sci U S A 113:2696-701

212. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, *et al.* 2020. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv (preprint)