Targeted genomic sequencing with probe capture for discovery and surveillance of coronaviruses in bats

Kevin S. Kuchinski[1,2], Kara D. Loos[3,4], Danae M. Suchan[3,4], Jennifer N. Russell[3,4], Ashton N. Sies[3,4], Charles Kumakamba[5], Francisca Muyembe[5], Placide Mbala Kingebeni[5,6], Ipos Ngay Lukusa[5], Frida N'Kawa[5], Joseph Atibu Losoma[5], Maria Makuwa[5,7], Amethyst Gillis[8,9], Matthew LeBreton[10], James A. Ayukekbong[11,12], Corina Monagin[8,13], Damien O. Joly[11,14], Karen Saylors[7,8], Nathan D. Wolfe[8], Edward M. Rubin[8], Jean J. Muyembe Tamfum[6], Natalie A. Prystajecky[1,2], David J. McIver[11,15], Christian E. Lange[7,11], Andrew D.S. Cameron*[3,4]

[1] University of British Columbia, Department of Pathology and Laboratory Medicine, Vancouver, British Columbia, Canada

[2] British Columbia Centre for Disease Control, Public Health Laboratory, Vancouver, British Columbia, Canada

[3] University of Regina, Department of Biology, Faculty of Science, University of Regina, Regina, Saskatchewan, Canada

[4] University of Regina, Institute for Microbial Systems and Society, Faculty of Science, University of Regina, Regina, Saskatchewan, Canada

[5] Metabiota Inc., Kinshasa, Democratic Republic of the Congo

[6] Institut National de Recherche Biomédicale, Kinshasa, Democratic Republic of the Congo

[7] Labyrinth Global Health Inc., St. Petersburg, USA

[8] Metabiota Inc., San Francisco, USA

[9] Development Alternatives, Inc., Washington DC, USA

[10] Mosaic, Yaoundé, Cameroon

[11] Metabiota Inc., Nanaimo, Canada

[12] Southbridge Care, Cambridge, Ontario, Canada

[13] One Health Institute, School of Veterinary Medicine, University of California, Davis, California, USA

[14] Nyati Health Consulting, Nanaimo, British Columbia, Canada

[15] Institute for Global Health Sciences, University of California, San Francisco, California, USA

* Correspondence: andrew.cameron@uregina.ca

ABSTRACT

Public health emergencies like SARS, MERS, and COVID-19 have prioritized surveillance of zoonotic coronaviruses, resulting in extensive genomic characterization of coronavirus diversity in bats. Sequencing viral genomes directly from animal specimens remains a laboratory challenge, however, and most bat coronaviruses have been characterized solely by PCR amplification of small regions from the best-conserved gene. This has resulted in limited phylogenetic resolution and left viral genetic factors relevant to threat assessment undescribed.

In this study, we evaluated whether a technique called hybridization probe capture can achieve more extensive genome recovery from surveillance specimens. Using a custom panel of 20,000 probes, we captured and sequenced coronavirus genomic material in 21 swab specimens collected from bats in the Democratic Republic of the Congo. For 15 of these specimens, probe capture recovered more genome sequence than had been previously generated with standard amplicon sequencing protocols, providing a median 6.1-fold improvement (ranging up to 69.1-fold). Probe capture data also identified five novel *alpha*- and *betacoronaviruses* in these specimens, and their full genomes were recovered with additional deep sequencing. Based on these experiences, we discuss how probe capture could be effectively operationalized alongside other sequencing technologies for high-throughput, genomics-based discovery and surveillance of bat coronaviruses.

INTRODUCTION

Orthocoronavirinae, commonly known as coronaviruses (CoVs), are a diverse subfamily of RNA viruses that infect a broad range of mammals and birds [Corman 2018, Ye 2020, Ruiz-Aravena 2021]. Since the 1960s, four endemic human CoVs have been identified as common causes of mild respiratory illnesses [Corman 2018, Ye 2020]. In the past two decades, additional CoV threats have emerged, most notably SARS-CoV, MERS-CoV, and SARS-CoV-2, causing severe disease, public health emergencies, and global crises [Drosten 2003, Zaki 2012, Hu 2015, Corman 2018, Ye 2020, Zhou 2020]. These spill-overs have established CoVs alongside influenza A viruses as important zoonotic pathogens and pandemic threats. Indeed, evolving perceptions of CoV risk have led to speculation that some historical pandemics have been mis-attributed to influenza, and they may have in fact been the spill-overs of now-endemic human CoVs [Vijgen 2005, Corman 2018, Brüssow 2021].

Emerging CoV threats have motivated extensive viral discovery and surveillance activities at the interface between humans, livestock, and wildlife [Drexler 2014, Frutos 2021, Geldenhuys 2021]. Many of these activities have focused on bats (order *Chiroptera*). They are the second-most diverse order of mammals, following rodents, and they are a vast reservoir of CoV diversity [Drexler 2014, Hu 2015, Frutos 2021, Geldenhuys 2021, Ruiz-Aravena 2021]. Bats have been implicated in the emergence of SARS-CoV, MERS-CoV, SARS-CoV-2 and, less recently, the endemic human CoVs NL63 and 229E [Li 2005, Pfefferle 2009, Tong 2009, Huynh 2012, Corman 2015, Hu 2015, Yang 2015, Tao 2017, Ye 2020, Zhou 2020, Ruiz-Aravena 2021].

Genomic sequencing has been instrumental for characterizing CoV diversity and potential zoonotic threats, but recovering viral genomes directly from animal specimens remains a laboratory challenge. Host tissues and microbiota contribute excessive background genomic material to specimens, diluting viral genome fragments and vastly increasing the sequencing depth required for target detection and accurate genotyping. Consequently, laboratory methods for targeted enrichment of viral genome material have been necessary for practical, high-throughput sequencing of surveillance specimens [Houldcroft 2017, Fitzpatrick 2021].

There are two major paradigms for targeted enrichment of genomic material. The first, called amplicon sequencing, uses PCR to amplify target genomic material. It is comparatively straightforward and sensitive, but PCR chemistry limits amplicon length and relies on the presence of specific primer sites across diverse taxa [Houldcroft 2017, Fitzpatrick 2021]. In practice, extensive genomic divergence within viral taxa often constrains amplicon locations to the most conserved genes, limiting phylogenetic resolution [Drexler 2014, Li 2020]. This also hinders characterization of viral genetic factors relevant for threat assessment like those encoding determinants of host range, tissue tropism, and virulence. These kinds of targets are often hypervariable due to strong evolutionary pressures from host adaptation and immune evasion, and consequently they do not have well-conserved locations for PCR primers. Due to these limitations, studies of CoV diversity have been almost exclusively based on small regions of the relatively conserved RNA-dependent RNA polymerase (RdRp) gene [Drexler 2014, Geldenhuys 2021].

The second major paradigm for enriching viral genomic material is called hybridization probe capture. This method uses longer nucleotide oligomers to anneal and immobilize complementary target genomic fragments while background material is washed away. Probes are

typically 80 to 120 nucleotides in length, making them more tolerant of sequence divergence and nucleotide mismatches than PCR primers [Brown 2016]. Probe panels are also highly scalable, allowing for the simultaneous capture of thousands to millions of target sequences. This has made them popular for applications where diverse and hypervariable viruses are targeted, but they have only been occasionally used to attempt sequencing of bat CoVs [Lim 2019, Li 2020].

In this study, we evaluated hybridization probe capture for enriching CoV genomic material in oral and rectal swabs previously collected from bats. We designed a custom panel of 20,000 hybridization probes targeting the known diversity of bat coronavirus. This panel was applied to 21 swab specimens collected in the Democratic Republic of the Congo (DRC), in which novel CoVs had been previously characterized by partial RdRp sequencing using standard amplicon methods [Kumakamba 2021]. We compared the extent of genome recovery by probe capture and amplicon sequencing, and we used probe capture data in conjunction with deep metagenomic sequencing to characterize full genomes for five novel *alpha-* and *betacoronaviruses*. Based on these experiences, we discuss how probe capture could be effectively operationalized alongside other targeted sequencing technologies for high-throughput, genomics-based discovery and surveillance of bat coronaviruses.

MATERIALS AND METHODS

Additional details for the following materials and methods are provided in Supplemental 1.

Bat swab specimens and partial RdRP sequences: Rectal and oral swabs were collected between August 2015 and June 2018 in different locations in DRC from bats that were either captured and released or that were for sale in local markets [Kumakamba 2021]. Swabs were collected into individual 2.0 ml screw-top cryotubes containing 1.5 ml of either Universal Viral Transport Medium (BD) or Trizol® (Invitrogen), stored in liquid nitrogen for transport as soon as practical and later transferred into -80°C freezers. CoV screening involved two consensus PCR assays targeting the RNA-dependant RNA polymerase (RdRp) performed in Kinshasa, DRC, and commercial Sanger sequencing of amplicons [Quan 2010, Watanabe 2010, Kumakamba 2021]. Bat species were identified by ecologists in the field and verified using a PCR targeting the Cytochrome B gene [Townzen 2008]. 21 unique specimens were shipped to Canada: 15 as RNA extracts only, 2 as unextracted swabs in transport medium, and 4 as both

previously extracted RNA and unextracted swabs in transport medium. Swabs in transport medium were re-extracted upon receipt using the Invitrogen TRIzol Reagent (#15596026) following the manufacturer's protocol. RNA concentration and RNA Integrity Number (RIN) for all RNA extracts were measured using the Agilent BioAnalyzer 2100 instrument with the RNA 6000 Nano kit.

Probe panel design and reference sequence coverage assessments: All available bat CoV sequences were downloaded from NCBI GenBank on October 4, 2020. A custom panel of 20,000 hybridization probes was designed from these sequences using the ProbeTools package (v0.0.5) [Kuchinski 2022]. Probe coverage of reference sequences was also assessed *in silico* using ProbeTools. The final panel (Supplemental 2) was synthesized by Twist Bioscience (San Francisco, CA, USA).

Library construction, probe capture, and sequencing of captured libraries: Sequencing libraries were constructed using the NEBNext Ultra II RNA Library Prep with Sample Purification Beads kit (E7775), then libraries were barcoded with unique dual indices from the NEBNext Multiplex Oligos for Illumina kit (E6440). Libraries were pooled together, then the pool was captured twice sequentially by our custom probe panel with the Twist Bioscience Fast Hybridization kit (#100964), Universal Blockers (#100578), Binding and Purification Beads (#100983), and Fast Wash Buffers (#100971). Probe captured libraries were sequenced on an Illumina MiSeq instrument using V2 300 cycle reagent kits (#MS-102-2002). Index hops were filtered using HopDropper (v0.0.3) (https://github.com/KevinKuchinski/HopDropper).

Control specimens were prepared by spiking 100,000 copies of a synthetic control oligo into 200 ng of Invitrogen Human Reference RNA (#QS0639). The control oligo was manufactured by Integrated DNA Technologies (Coralville, IA, USA) as a dsDNA gBlock with a known artificial sequence created by the authors. Probes targeting the control oligo were included in the custom capture panel. Control specimens were prepared into libraries alongside bat specimens from the same reagent master mixes, and they were included in the same pool for probe capture. Detection and enrichment of the control oligo sequence in control specimen libraries was used as a positive control for library construction and probe capture. Absence of control oligo sequences in bat specimen libraries and absence of bat CoV sequences in control

specimen libraries were used as a negative control for contamination and as a positive control for index hop removal by HopDropper.

De novo assembly of contigs from captured reads: coronaSPAdes (v3.15.0) was used to assemble contigs *de novo* from probe captured MiSeq data [Meleshko 2021]. CoV contigs were identified using BLASTn (v2.5.0) against a local database composed of all *coronaviridae* sequences in GenBank available as of October 11, 2021 [Camacho 2009].

Alignment of reads and contigs to bat CoV reference sequences: Probe captured reads were mapped to selected reference sequences using bwa mem (0.7.17-r1188), then alignments were filtered, sorted, and indexed using samtools (v1.11) [Li 2009a, Li 2009b]. Depth and extent of read coverage were determined with bedtools genomecov (v2.30.0) [Quinlan 2010]. Contig coverage was determined by aligning contigs to reference sequences with BLASTn (v2.5.0) and extracting subject start and subject end coordinates [Camacho 2009].

Deep metagenomic sequencing of uncaptured libraries and generation of complete viral genomes: Selected specimens were sequenced on an Illumina HiSeq X instrument by the Michael Smith Genome Sciences Centre (Vancouver, BC, Canada). Reads were assembled and scaffolded into draft genomes with coronaSPAdes (v3.15.3) [Meleshko 2021]. HiSeq reads were mapped to draft genomes using bwa mem (v0.7.17-r1188), then alignments were filtered, sorted, and indexed using samtools (v1.11) [Li 2009a, Li 2009b]. Variants were called with bcftools mpileup and call (v1.9), then variants were applied to draft genomes with bcftools consensus (v1.9) to generate final complete genomes [Danecek 2021].

Phylogenetic analysis of novel spike gene sequences: Novel spike genes were translated from complete genomes then queried against all translated *coronaviridae* spike sequences in GenBank using BLASTp (v2.5.0) [Camacho 2009]. For each genus, novel spike genes from study specimens were combined with the 25 closest-matching GenBank spike sequences and all spike sequences available in RefSeq. Multiple sequence alignments were conducted with clustalw (v2.1), then phylogenetic trees were constructed from aligned sequences using PhyML (v3.3.20190909) [Thompson 1994, Guindon 2005].

RESULTS

Custom hybridization probe panel provided broad coverage *in silico* of known bat CoV diversity: To begin this study, we designed a custom panel of hybridization probes targeting known bat CoV diversity. We obtained 4,852 bat CoV genomic sequences from GenBank, used them to design a custom panel of 20,000 probe sequences, then assessed *in silico* how extensively these reference sequences were covered by our custom panel (Figure 1A). For 90% of these bat CoV sequences, the custom panel covered at least 94.32% of nucleotide positions. We also evaluated probe coverage for the subset of these sequences representing full-length bat CoV genomes (Figure 1B), and 90% of these targets had at least 98.73% of their nucleotide positions covered. These results showed broad probe coverage of known bat CoV diversity at the time the panel was designed.

Probe capture provided more extensive genome recovery than previous amplicon sequencing for most specimens: We used our custom panel to assess probe capture recovery of CoV material in 25 metagenomic sequencing libraries. We prepared these libraries from a retrospective collection of 21 bat oral and rectal swabs that had been collected in DRC between 2015 and 2018. These swabs had been collected as part of the PREDICT project, a large-scale United States Agency for International Development (USAID) Emerging Pandemic Threats initiative that has collected over 20,000 animal specimens from 20 CoV hotspot countries [*e.g.* Anthony 2017, Lacroix 2017, Nziza 2020, Valitutto 2020, Ntumvi 2022]. Most libraries (n=19) were prepared from archived RNA that had been previously extracted from these specimens, although some libraries (n=6) were prepared from RNA that was freshly extracted from archived primary specimens (Table 1). CoVs had been previously detected in these specimens with PCR assays by Quan *et al.* (2010) and Watanabe *et al.* (2010). Sanger sequencing of these amplicons by Kumakamba *et al.* (2021) had generated partial RdRp sequences of 286 or 387 nucleotides, which had been used to assign these specimens to four novel phylogenetic groups of *alpha-* and *betacoronaviruses* (Table 1).

We captured CoV genomic material in these metagenomic bat swab libraries with our custom probe panel then performed genomic sequencing. To assess CoV recovery, we began with a strategy that would be suitable for automated bioinformatic analysis in high-throughput

surveillance settings: sequencing reads from probe captured libraries were assembled *de novo* into contigs, then CoV sequences were identified by locally aligning contigs against a database of CoV reference sequences. In total, 113 CoV contigs were recovered from 17 of 25 libraries. We compared contig lengths to the partial RdRp amplicons that been previously generated for these specimens (Figure 2A). The protocol by Watanabe *et al.* had generated 387 nucleotide-long partial RdRp sequences, but median contig size with probe capture for these specimens was 696 nucleotides (IQR: 453 to 1,051 nucleotides, max: 19,601 nucleotides). The protocol by Quan *et al.* had generated 286 nucleotide-long partial RdRp sequences, but median contig size with probe capture for these specimens was 602 nucleotides (IQR: 423 to 1,053 nucleotides, max: 4,240 nucleotides). Overall, 107 contigs (93.8%) were longer than the partial RdRp sequence previously generated for their specimen by standard amplicon sequencing protocols, demonstrating the capacity of probe capture to recover larger contiguous fragments of CoV genome sequence.

Next, we used assembly size metrics to assess the extent to which these contigs represented complete genomes. The median total assembly size was 1,724 nucleotides (IQR: 0 to 5,834 nucleotides), while median assembly N50 size was 533 nucleotides (IQR: 0 to 908 nucleotides) (Figure 2B). This assembly size-based assessment of genome completeness had limitations, however. Some assembly sizes may have been understated by genome regions with comparatively low read coverage that failed to assemble. Conversely, other assembly sizes may have been overstated by redundant contigs resulting from forked assembly graphs, either due to genetic variation within the intrahost viral population or due to polymerase errors introduced during library construction and probe capture. For instance, the total assembly size for library CDAB0217R-PRE was 33,195 nucleotides, exceeding the length of the longest known CoV genome (Figure 2C). Another limitation of this analysis was that these assembly metrics provided no indication of which regions of the genome had been recovered.

To address these limitations, we also applied a reference sequence-based strategy. We used the contigs to identify the best available CoV reference sequences for each of the four novel phylogenetic groups to which these specimens had been assigned. Sequencing reads from captured libraries were directly mapped to these reference sequences and the contigs we had assembled *de novo* were also locally aligned to them (Fig 3 and S1-S4). Based on these read mappings and contig alignments, we calculated for each library a breadth of reference sequence

recovery, *i.e.* the number of nucleotide positions in the reference sequence covered by either mapped sequencing reads or contigs (Figure 4A).

The median breadth of reference sequence recovery for all libraries was 2,376 nucleotides (IQR: 306 to 9,446 nucleotides). Most libraries (48%) represented specimens from phylogenetic group Q-Alpha-4, which had a median reference sequence recovery of 6,497 nucleotides (IQR: 733 to 9,802 nucleotides, max: 12,673 nucleotides). Phylogenetic group W-Beta-3 also accounted for a substantial fraction of libraries (32%), and although median reference sequence recovery was lower than for Q-Alpha-4 (2,427 nucleotides), W-Beta-3 provided the libraries with the most extensive reference sequence recoveries (IQR: 780 to 19,286 nucleotides, max: 26,755 nucleotides). As a simple way to quantify differences in recovery of CoV genome sequence between probe capture and amplicon sequencing, we calculated the ratio between the breadth of reference sequence recovery and the length of the previously generated partial RdRp amplicon sequence for each library (Figure 4B). The median ratio was 6.1-fold (IQR: 0.8-fold to 33.0-fold), reaching a maximum of 69.1-fold. Probe capture recovery was greater for 18 of 25 libraries (72%), representing 15 of 21 specimens (71%).

Probe capture recovery limited by *in vitro* sensitivity: No CoV sequences were recovered from 4 of 25 libraries (representing 3 specimens), despite partial RdRp sequences being obtained from them previously. Furthermore, probe capture did not yield any complete CoV genomes, and many specimens displayed scattered and discontinuous reference sequence coverage (Figures S1-S4). We considered two explanations for this result. First, CoV material in these libraries may not have been completely captured because they were not targeted by any probe sequences in the panel. Second, CoV material in these specimens may not have been incorporated into the sequencing libraries due to factors limiting *in vitro* sensitivity, *e.g.* low prevalence of viral genomic material; sub-optimal nucleic acid concentration and integrity in archived RNA and primary specimens; and library preparation reaction inefficiencies.

First, we assessed *in vitro* sensitivity. To exclude missing probe coverage as a confounder in this analysis, we evaluated recovery of the previously sequenced partial RdRp amplicons. Since their sequences were known, we could assess probe coverage *in silico* and demonstrate whether these targets were covered by the panel. All partial RdRp amplicons had at least 95.3% of their nucleotide positions covered by the probe panel (Figure 5A), but this did not translate

into extensive recovery. For 12 of 25 libraries, no part of the partial RdRp sequence was recovered, and full/nearly-full recovery (>95%) of the partial RdRp sequence was achieved for only 7 of 25 libraries (Figure 5A). These results demonstrated that genome recovery had been limited by factors other than probe panel inclusivity.

Next, we examined nucleic acid concentration and integrity, two specimen characteristics associated with successful library preparation. Median RNA Integrity Number (RIN) values and RNA concentrations for these specimens were low: 1.1 and 14 ng/µl respectively, as was expected from archived material (Figure 5B). To assess the impact of RIN and RNA concentration on probe capture recovery, we compared these specimen characteristics against breadth of reference sequence recovery from the corresponding libraries (Figure 5B). Weak monotonic relationships were observed, with lower RNA concentration and lower RIN values generally leading to worse genome recovery. This relationship was significant for RNA concentration ($p=0.045$, Spearman's rank correlation), but not for RNA integrity despite trending towards significance ($p=0.053$, Spearman's rank correlation). These weak associations suggested additional factors hindered recovery, *e.g.* low prevalence of viral material or missing probe coverage for genomic regions outside the partial RdRp target. Missing probe coverage is considered in the next section. Prevalence of viral material was not practical to consider as there are no established pan-CoV methods for quantifying genome copies in RNA specimens, a limitation that would also preclude attempts to triage surveillance specimens based on viral abundance in high-throughput settings.

Inclusivity of custom probe panel against CoV taxa in study specimens: Next, we considered if blind spots in the probe panel had contributed to incomplete genome recovery from these specimens. This inquiry suffered a counterfactual problem: to assess whether the CoV taxa in our specimens were fully covered by our probe panel, we would need their complete genome sequences. We did not have their full genome sequences, however, because the probes did not recover them. Instead, we evaluated probe coverage of the reference sequences assigned to each phylogenetic group, assuming they were the available CoV sequences most similar to those in our specimens.

Probe coverage was nearly complete for all reference sequences (Figure 6). Nonetheless, reference sequence recovery did not exceed 92.3% for any of these libraries, and complete spike

genes were conspicuously absent (Figure 3, S1-S4). This included specimens like CDAB0203R-PRE, CDAB0217R-PRE, and CDAB0492R-PRE where recovery was otherwise extensive and contiguous, suggesting genomic material was sufficiently abundant and intact for sensitive library construction. These results indicated the presence of CoVs similar to Bat coronavirus CMR704-P12 and *Chaerephon* bat corornavirus/Kenya/KY22/2006, except with novel spike genes that diverged from the spike genes of these reference sequences and all other CoVs described in GenBank.

Recovery of complete genome sequences from five novel bat *alpha-* and *betacoronaviruses*: Analysis of our probe capture data confirmed the presence of several novel coronaviruses in these specimens, as had been previously determined by Kumakamba *et al.* (2021). Our results also suggested the CoVs in these specimens contained spike genes that were highly divergent from any others that have been previously described. This led us to perform deep metagenomic sequencing on select specimens to attempt recovery of complete CoV genomes. We selected the following nine specimens, either due to extensive recovery by probe capture (indicating comparatively abundant and intact viral genomic material) or to ensure representation of the four novel phylogenetic groups: CDAB0017RSV, CDAB0040RSV, CDAB0174R, CDAB0203R, CDAB0217R, CDAB0113RSV, CDAB0491R, and CDAB0492R.

Complete genomes were only recovered from 5 specimens: CDAB0017RSV, CDAB0040RSV, CDAB0203R, CDAB0217R, and CDAB0492R. The abundance of CoV genomic material in these 5 specimens was estimated by mapping reads from uncaptured libraries to the complete genome sequence that we recovered. On-target rates, *i.e.* the percentage of total reads mapping to the CoV genome, were calculated (Figure 7A). These ranged from 0.003% to 0.064%, revealing the extremely low abundance of viral genomic material present in these swabs. Considering these were the most successful libraries, these results highlighted that low prevalence of viral genomic material is one challenging characteristic of swab specimens.

We also used the complete genome sequences that we recovered to assess how effectively probe capture enriched target genomic material in these specimens. Valid reads from probe captured libraries were mapped to the complete genomes from their corresponding specimens. On-target rates for captured libraries ranged from 11.3% to 45.1% of valid reads (Figure 7B).

Due to insufficient library material remaining after probe capture, new libraries had been made for deep metagenomic sequencing. Consequently, we did not pair on-target rates for these libraries to calculate fold-enrichment values. Instead, we compared mean on-target rates for the deep-sequenced unenriched metagenomic libraries (0.029% mean on-target) against the original probe captured libraries (29.6% mean on-target); we observed a 1,020-fold difference between these means, with the probe captured on-target rates significantly higher ($p<0.001$, t-test on 2 independent means). These results confirmed effective enrichment by probe capture of CoV material present in these libraries.

Phylogenetic analysis of novel spike gene sequences: Novel spike gene sequences were translated from the complete genomes we had recovered, then these were compared to spike protein sequences from other CoVs in GenBank. Spike protein sequences from specimens CDAB0017RSV and CDAB0040RSV formed a monophyletic clade, as did those from specimens CDAB0203R and CDAB0217R, reflecting their membership in partial RdRp-based phylogenetic groups W-Beta-2 and W-Beta-3 respectively (Figure 8). These novel spike proteins also grouped with spike protein sequences from three *betacoronaviruses* in GenBank: HQ728482.1, MG693168.1, and NC_048212.1 (Figure 8). The spike protein sequence from specimen CDAB0492R, the lone Q-Alpha-4 representative, grouped with spikes from two *alphacoronaviruses* in GenBank: HQ728486.1 and MZ081383.1 (Figure 9).

Pairwise global alignments of amino acid sequences were conducted between these novel spike genes and the spike genes from GenBank with which they grouped phylogenetically. Alignments completely covered all novel spike sequences, but they were all less than 76.5% identical and less than 85.7% positive (Table 2). We compared host species and geographic collection locations for our study specimens and the phylogenetically related spike sequences. Only specimens CDAB0203R and CDAB0217R were collected from the same bat species as their closest spike protein matches in GenBank (*Eidolon helvum*). Other specimens were detected in bat genera different from their closest GenBank match. All study specimens were collected from the DRC, but their closest GenBank matches were collected from diverse locales, including neighbouring Kenya, Cameroon in West Africa, and Yunnan province in China. Taken together, these low alignment scores, disparate host species, and dispersed collection locations suggested these viruses belong to extensive but hitherto poorly characterized taxa of CoV.

We also conducted pairwise global alignments of nucleotide sequences. This was done to confirm that probe capture had been hindered by divergence of these novel spike genes from their closest matches in GenBank, which we had used to design our custom panel. For specimen CDAB0017RSV, sequence similarity was so low that no alignment was generated for the spike gene. Nucleotide alignments for the other specimens were all incomplete (18% to 83% coverage of the novel spike sequence) with low nucleotide identities (71.5% to 84.6%).

DISCUSSION

This study highlights the potential for probe capture to recover greater extents of CoV genome compared to standard amplicon sequencing methods. In discovery and surveillance applications, this would permit characterization of CoV genomes outside of the constrained partial RdRp regions that are typically described, enabling additional phylogenetic resolution among specimens with similar partial RdRp sequences. Recovering more extensive fragments from diverse regions of the genome would also provide additional genetic sequence to compare against reference sequences in databases like GenBank and RefSeq. This could permit more confident identification of known threats and better assessment of virulence and potential spill-over from novel CoVs. Sequences from additional genome regions could also be used to identify CoVs where recombination has occurred, which is increasing recognized as a potential hallmark of zoonotic CoVs [Hu 2015, Corman 2018, Ye 2020, Ruiz-Aravena 2021].

This study also showed the usefulness of probe capture for identifying specimens that warrant the expense of deep metagenomic sequencing for more extensive characterization. The genomic regions missed by the probe panel can provide as much insight into viral novelty as the sequences that are recovered. In this study, failure to capture complete spike gene sequences, even from libraries with otherwise extensive coverage, was successfully used to predict the presence of novel spike genes. Furthermore, contiguity across recovered regions can be used to evaluate abundance and intactness of viral genomic material, identifying specimens where deep metagenomic sequencing is likeliest to succeed. This is valuable when targeting higher taxonomic levels where methods for directly quantifying viral genome copies are hindered by the same genomic variability that constrains amplicon sequencing.

This study also revealed two important limitations for probe capture in CoV discovery and surveillance applications. The first, which appeared to be the most limiting in this study, is

the *in vitro* sensitivity of this method. Probe capture must be performed on already constructed metagenomic sequencing libraries. The library construction process involves numerous sequential biochemical reactions and bead clean-ups, where inefficiencies result in compounding losses of input material. Combined with the low prevalence of viral genomic material in swab specimens, these loses of input material can lead to the presence of incomplete viral genomes in sequencing libraries and stochastic recovery during probe capture. Amplicon sequencing does not suffer the same attrition because enrichment occurs as the first step of the process, allowing library construction to occur on abundant amplicon input material. Further work optimizing metagenomic library construction protocols could be done to improve sensitivity for probe capture. Also, this study relied on archived material in suboptimal condition, so better results could be expected from fresh surveillance specimens.

The second limitation highlighted by this work is the challenge of designing hybridization probes from available reference sequences for poorly characterized taxa. Currently, the extent of human knowledge about bat CoV diversity remains limited, especially across hypervariable genes like spike, and it seems impossible to design a broadly inclusive pan-bat CoV probe panel at this moment. As recently as 2017, it was observed that only 6% of CoV sequences in GenBank were from bats, while the remaining 94% of sequences concentrated on a limited number of known human and livestock pathogens [Anthony 2017]. The vastness of CoV diversity that remains to be characterized is evident by the continuing high rate of novel CoV discovery by research studies and surveillance programs, this current work included [*e.g.* Tao 2017, Wang 2017, Markotter 2019, Wang 2019, Nziza 2020, Valitutto 2020, Kumakamba 2021, Shapiro 2021, Tan 2021, Wang 2021, Zhou 2021, Alkhovsky 2022, Ntumvi 2022].

Fortunately, probe capture is highly adaptable and existing panels can be easily supplemented with additional probes as new CoV taxa are described. For instance, the genomes recovered in this study could be used to design supplemental probes for re-capturing existing specimens as well as for future projects with new specimens. Improved recovery would be especially expected for projects returning to similar geographic regions targeting similar bat populations. These probe design limitations are also only a meaningful impediment for CoV discovery, specifically the gold standard recovery of complete genomes, as surveillance activities do not require recovery of the entire genome to adequately detect known pathogenic threats. Furthermore, extensive sequencing of zoonotic CoV taxa that have already emerged has

435 provided abundant reference sequences for probe design geared towards genomic detection of
436 these known pathogenic threats.
437 Our results lead us to conclude that probe capture amounts to a trade-off; sensitivity
438 limitations mean that CoV sequence recovery may occur less frequently than with amplicon
439 sequencing, but when it does succeed, CoV sequences may be more extense and more diverse.
440 Likewise, probe panel designs may not be broadly inclusive enough to recover complete
441 genomes in all cases, but the sequencing depth required – and thus the cost per specimen – to
442 attempt recovery will be fractional compared to untargeted methods. Consequently, probe
443 capture is not a replacement for amplicon sequencing or deep metagenomic sequencing, but a
444 complementary method to both.
445 Based on these observations, we propose that the most effective CoV discovery and
446 surveillance programs will combine amplicon sequencing, probe capture, and deep metagenomic
447 sequencing. The simplicity, sensitivity, and affordability of amplicon sequencing makes it well-
448 suited for initial screening. This method also requires the least laboratory infrastructure, much of
449 which already exists in surveillance hotspots at facilities with extensive experience and
450 established track records of success. Screening by amplicon sequencing would enable direct
451 phylogenetic comparisons between specimens across consistent genomic loci and enable a
452 preliminary assessment of threat and novelty. This screening would also identify CoV-positive
453 specimens warranting further study, limiting the number of specimens to be transported to more
454 specialized laboratories with probe capture and deep sequencing capacity.
455 Probe capture on select CoV-positive specimens would be valuable for potentially
456 acquiring additional sequence information which could refine assessments of threat and novelty.
457 As new CoVs are characterized and probe panel designs are expanded, recovery of host range
458 and virulence factors by probe capture would steadily increase.
459 Finally, probe capture results would be used to identify interesting specimens warranting
460 the expense of deep metagenomic sequencing. It would also be used to triage specimens based
461 on the abundance and intactness of viral genomic material inferred from the probe capture
462 results. Deep sequencing would allow for the most extensive characterization and evaluation of
463 novel CoV genomes, especially for hypervariable host range and virulence factors like spike
464 gene. It would also provide novel sequences for updating probe panel designs. Deploying these

methods in conjunction, with each used to its strength, would enable highly effective genomics-based discovery and surveillance for bat CoVs.

DATA AVAILABILITY

The sequence data from this study is available at National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) as BioProject PRJNA823716. The assembled coronavirus genomes are available at GenBank with following accession numbers: ON313743 (CDAB0017RSV); ON313744 (CDAB0040RSV); ON313745 (CDAB0203R); ON313746 (CDAB0217R); ON313747 (CDAB0492R).

ACKNOWLEDGEMENTS

The authors would like to thank: members of the Institute for Microbial Systems and Society, Caroline Cameron and David Alexander for helpful discussions; the government of the Democratic Republic of the Congo for the permission to conduct this study and the late Prime Mulembakani for his invaluable contribution to the success of this work; Guy Midingi Sepolo, Joseph Fair, Bradley Schneider, Anne Rimoin, Nicole Hoff and other members of the PREDICT consortium (https://ohi.vetmed.ucdavis.edu/programs-projects/predict-project/authorship) for their support.

FUNDING

This study was made possible by funding from Genome Prairie COVID-19 Rapid Regional Response (COV3R) and the Saskatchewan Health Research Foundation COV3R Partnership grants. This study was made possible partially thanks to the generous support of the American people through the United States Agency for International Development (USAID) Emerging Pandemic Threats PREDICT program (cooperative agreement number AID-OAA-A-14-00102). The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government.

Conflict of interest statement. None declared.

REFERENCES

Alkhovsky S, Lenshin S, Romashin A, et al. SARS-like Coronaviruses in Horseshoe Bats (Rhinolophus spp.) in Russia, 2020. Viruses. 2022;14(1):113. Published 2022 Jan 9. doi:10.3390/v14010113

Anthony SJ, Johnson CK, Greig DJ, et al. Global patterns in coronavirus diversity. Virus Evol. 2017;3(1):vex012. Published 2017 Jun 12. doi:10.1093/ve/vex012

Brown JR, Roy S, Ruis C, et al. Norovirus Whole-Genome Sequencing by SureSelect Target Enrichment: a Robust and Sensitive Method. J Clin Microbiol. 2016;54(10):2530-2537. doi:10.1128/JCM.01052-16

Brüssow H, Brüssow L. Clinical evidence that the pandemic from 1889 to 1891 commonly called the Russian flu might have been an earlier coronavirus pandemic. Microb Biotechnol. 2021;14(5):1860-1870. doi:10.1111/1751-7915.13889

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421.

Corman VM, Baldwin HJ, Tateno AF, et al. Evidence for an Ancestral Association of Human Coronavirus 229E with Bats. J Virol. 2015;89(23):11858-11870. doi:10.1128/JVI.01755-15

Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and Sources of Endemic Human Coronaviruses. Adv Virus Res. 2018;100:163-188. doi:10.1016/bs.aivir.2018.01.001

Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008. doi:10.1093/gigascience/giab008

Drexler JF, Corman VM, Drosten C. Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS. Antiviral Res. 2014;101:45-56. doi:10.1016/j.antiviral.2013.10.013

525 Drosten C, Günther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M,
526 Kolesnikova L, Fouchier RA, Berger A, Burguière AM, Cinatl J, Eickmann M, Escriou N,
527 Grywna K, Kramme S, Manuguerra JC, Müller S, Rickerts V, Stürmer M, Vieth S, Klenk HD,
528 Osterhaus AD, Schmitz H, Doerr HW. Identification of a novel coronavirus in patients with
529 severe acute respiratory syndrome. N Engl J Med. 2003 May 15;348(20):1967-76. doi:
530 10.1056/NEJMoa030747. Epub 2003 Apr 10. PMID: 12690091.
531
532 Fitzpatrick AH, Rupnik A, O'Shea H, Crispie F, Keaveney S, Cotter P. High Throughput
533 Sequencing for the Detection and Characterization of RNA Viruses. Front Microbiol.
534 2021;12:621719. Published 2021 Feb 22. doi:10.3389/fmicb.2021.621719
535
536 Frutos R, Serra-Cobo J, Pinault L, Lopez Roig M, Devaux CA. Emergence of Bat-Related
537 Betacoronaviruses: Hazard and Risks. Front Microbiol. 2021;12:591535. Published 2021 Mar
538 15. doi:10.3389/fmicb.2021.591535
539
540 Geldenhuys M, Mortlock M, Epstein JH, Pawęska JT, Weyer J, Markotter W. Overview of Bat
541 and Wildlife Coronavirus Surveillance in Africa: A Framework for Global Investigations.
542 Viruses. 2021;13(5):936. Published 2021 May 18. doi:10.3390/v13050936
543
544 Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online--a web server for fast maximum
545 likelihood-based phylogenetic inference. Nucleic Acids Res. 2005;33(Web Server issue):W557-
546 W559. doi:10.1093/nar/gki352
547
548 Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome
549 sequencing. Nat Rev Microbiol. 2017;15(3):183-192. doi:10.1038/nrmicro.2016.182
550
551 Hu B, Ge X, Wang LF, Shi Z. Bat origin of human coronaviruses. Virol J. 2015;12:221.
552 Published 2015 Dec 22. doi:10.1186/s12985-015-0422-1
553
554 Huynh J, Li S, Yount B, et al. Evidence supporting a zoonotic origin of human coronavirus strain
555 NL63. J Virol. 2012;86(23):12816-12825. doi:10.1128/JVI.00906-12

Kuchinski KS, Duan J, Himsworth C, Hsiao W, Prystajecky NA. 2022. ProbeTools: Designing hybridization probes for targeted genomic sequencing of diverse and hypervariable viral taxa. bioRxiv doi: https://doi.org/10.1101/2022.02.24.481870

Kumakamba C, Niama FR, Muyembe F, et al. Coronavirus surveillance in wildlife from two Congo basin countries detects RNA of multiple species circulating in bats and rodents. PLoS One. 2021;16(6):e0236971. Published 2021 Jun 9. doi:10.1371/journal.pone.0236971

Lacroix A, Duong V, Hul V, et al. Genetic diversity of coronaviruses in bats in Lao PDR and Cambodia. Infect Genet Evol. 2017;48:10-18. doi:10.1016/j.meegid.2016.11.029

Li B, Si HR, Zhu Y, et al. Discovery of Bat Coronaviruses through Surveillance and Probe Capture-Based Next-Generation Sequencing [published correction appears in mSphere. 2020 Mar 18;5(2):]. mSphere. 2020;5(1):e00807-19. Published 2020 Jan 29. doi:10.1128/mSphere.00807-19

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, Wang LF. Bats are natural reservoirs of SARS-like coronaviruses. Science. 2005; Oct 28;310(5748):676-9. doi: 10.1126/science.1118391. Epub 2005 Sep 29. PMID: 16195424.

Lim XF, Lee CB, Pascoe SM, et al. Detection and characterization of a novel bat-borne coronavirus in Singapore using multiple molecular approaches. J Gen Virol. 2019;100(10):1363-1374. doi:10.1099/jgv.0.001307

Markotter W, Geldenhuys M, Jansen van Vuren P, et al. Paramyxo- and Coronaviruses in Rwandan Bats. Trop Med Infect Dis. 2019;4(3):99. Published 2019 Jul 2. doi:10.3390/tropicalmed4030099

Meleshko D, Hajirasouliha I, Korobeynikov A. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. Bioinformatics. 2021; Aug 18:btab597. doi: 10.1093/bioinformatics/btab597. Epub ahead of print. PMID: 34406356.

Ntumvi NF, Ndze VN, Gillis A, Diffo JLD, Tamoufe U, Takuo JM, Mouiche MMM, Nwobegahay J, LeBreton M, Rimoin AW, Schneider BS, Monagin C, McIver DJ, Roy S, Ayukekbong JA, Saylors K, Joly DO, Wolfe ND, Rubin EM, Lange CE, Wildlife in Cameroon Harbor Diverse Coronaviruses Including Many Closely Related to Human Coronavirus 229E, Virus Evolution, 2022; veab110, https://doi.org/10.1093/ve/veab110

Nziza J, Goldstein T, Cranfield M, et al. Coronaviruses Detected in Bats in Close Contact with Humans in Rwanda. Ecohealth. 2020;17(1):152-159. doi:10.1007/s10393-019-01458-8

Pfefferle S, Oppong S, Drexler JF, et al. Distant relatives of severe acute respiratory syndrome coronavirus and close relatives of human coronavirus 229E in bats, Ghana. Emerg Infect Dis. 2009;15(9):1377-1384. doi:10.3201/eid1509.090224

Quan PL, Firth C, Street C, et al. Identification of a severe acute respiratory syndrome coronavirus-like virus in a leaf-nosed bat in Nigeria. mBio. 2010;1(4):e00208-10. doi:10.1128/mBio.00208-10

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033

Ruiz-Aravena M, McKee C, Gamble A, et al. Ecology, evolution and spillover of coronaviruses from bats [published online ahead of print, 2021 Nov 19] [published correction appears in Nat Rev Microbiol. 2022 Jan 13;:]. Nat Rev Microbiol. 2021;1-16. doi:10.1038/s41579-021-00652-2

Shapiro JT, Mollerup S, Jensen RH, et al. Metagenomic Analysis Reveals Previously Undescribed Bat Coronavirus Strains in Eswatini. Ecohealth. 2021;18(4):421-428. doi:10.1007/s10393-021-01567-3

Tan CS, Noni V, Sathiya Seelan JS, Denel A, Anwarali Khan FA. Ecological surveillance of bat coronaviruses in Sarawak, Malaysian Borneo. BMC Res Notes. 2021;14(1):461. Published 2021 Dec 20. doi:10.1186/s13104-021-05880-6

Tao Y, Shi M, Chommanard C, et al. Surveillance of Bat Coronaviruses in Kenya Identifies Relatives of Human Coronaviruses NL63 and 229E and Their Recombination History. J Virol. 2017;91(5):e01953-16. Published 2017 Feb 14. doi:10.1128/JVI.01953-16

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673-4680. doi:10.1093/nar/22.22.4673

Tong S, Conrardy C, Ruone S, et al. Detection of novel SARS-like and other coronaviruses in bats from Kenya. Emerg Infect Dis. 2009;15(3):482-485. doi:10.3201/eid1503.081013

Townzen JS, Brower AV, Judd DD. Identification of mosquito bloodmeals using mitochondrial cytochrome oxidase subunit I and cytochrome b gene sequences. Med Vet Entomol. 2008;22(4):386-93. doi: 10.1111/j.1365-2915.2008.00760.x. PMID: 19120966.

Valitutto MT, Aung O, Tun KYN, et al. Detection of novel coronaviruses in bats in Myanmar. PLoS One. 2020;15(4):e0230802. Published 2020 Apr 9. doi:10.1371/journal.pone.0230802

Vijgen L, Keyaerts E, Moës E, et al. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J Virol. 2005;79(3):1595-1604. doi:10.1128/JVI.79.3.1595-1604.2005

Wang L, Fu S, Cao Y, et al. Discovery and genetic analysis of novel coronaviruses in least horseshoe bats in southwestern China. Emerg Microbes Infect. 2017;6(3):e14. Published 2017 Mar 29. doi:10.1038/emi.2016.140

Wang N, Luo C, Liu H, et al. Characterization of a New Member of Alphacoronavirus with Unique Genomic Features in Rhinolophus Bats. Viruses. 2019;11(4):379. Published 2019 Apr 24. doi:10.3390/v11040379

Wang N, Luo CM, Yang XL, et al. Genomic Characterization of Diverse Bat Coronavirus HKU10 in Hipposideros Bats. Viruses. 2021;13(10):1962. Published 2021 Sep 29. doi:10.3390/v13101962

Watanabe S, Masangkay JS, Nagata N, et al. Bat coronaviruses and experimental infection of bats, the Philippines. Emerg Infect Dis. 2010;16(8):1217-1223. doi:10.3201/eid1608.100208

Yang XL, Hu B, Wang B, et al. Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. J Virol. 2015;90(6):3253-3256. Published 2015 Dec 30. doi:10.1128/JVI.02582-15

Ye ZW, Yuan S, Yuen KS, Fung SY, Chan CP, Jin DY. Zoonotic origins of human coronaviruses. Int J Biol Sci. 2020;16(10):1686-1697. Published 2020 Mar 15. doi:10.7150/ijbs.45472

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012 Nov 8;367(19):1814-20. doi: 10.1056/NEJMoa1211721. Epub 2012 Oct 17. Erratum in: N Engl J Med. 2013 Jul 25;369(4):394. PMID: 23075143.

677

678 Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of
679 probable bat origin. Nature. 2020;579(7798):270-273. doi:10.1038/s41586-020-2012-7

680

681 Zhou H, Ji J, Chen X, et al. Identification of novel bat coronaviruses sheds light on the
682 evolutionary origins of SARS-CoV-2 and related viruses. Cell. 2021;184(17):4380-4391.e14.
683 doi:10.1016/j.cell.2021.06.008

684

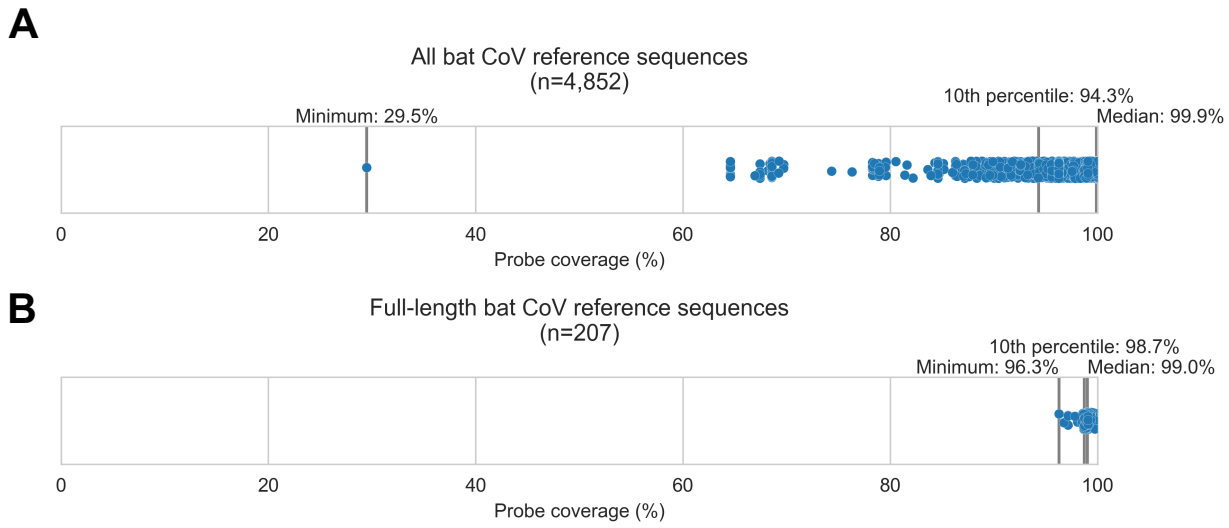

Figure 1: Custom hybridization probe panel provided broadly inclusive coverage of known bat coronavirus diversity *in silico*. Bat CoV sequences were obtained by downloading all available *alphacoronavirus*, *betacoronavirus*, and unclassified *coronaviridae* and *coronavirinae* sequences from GenBank on Oct 4, 2020 and searching for bat-related keywords in sequence headers. A custom panel of 20,000 probes was designed to target these sequences using the *makeprobes* module in the ProbeTools package. The ProbeTools *capture* and *stats* modules were used to assess probe coverage of bat CoV reference sequences. **A)** Each bat CoV sequence is represented as a dot plotted according to its probe coverage, *i.e.* the percentage of its nucleotide positions covered by at least one probe in the custom panel. **B)** The same analysis was performed on the subset of sequences representing full-length genomes (>25 kb in length).

Table 1: Bat specimens and sequencing libraries analyzed in this study. Collaborators Kumakamba *et al.* provided archived RNA previously extracted from 19 oral and rectal swabs along with 6 archived oral and rectal swab specimens, which were newly extracted with Trizol reagent upon receipt. Swabs had been collected in the Democratic Republic of the Congo between 2015 and 2018. Kumakamba *et al.* (2021) generated partial sequences from the RNA-dependent RNA polymerase gene using amplicon sequencing protocols by Quan *et al.* (2010) and Watanabe *et al.* (2010), which were used to assign these specimens to four novel phylogenetic groups of *alpha-* and *betacoronaviruses*.

Specimen ID	Library ID	Host	Swab type	RNA extraction method	Phylogenetic group
CDAB0017RSV	CDAB0017RSV-PRE	*Micropteropus pusillus*	Rectal	Previously-extracted	W-Beta-2
CDAB0040R	CDAB0040R-PRE	*Myonycteris sp.*	Rectal	Previously-extracted	W-Beta-2
CDAB0040RSV	CDAB0040RSV-PRE	*Myonycteris sp.*	Rectal	Previously-extracted	W-Beta-2
CDAB0305R	CDAB0305R-PRE	*Micropteropus pusillus*	Rectal	Previously-extracted	W-Beta-2
CDAB0146R	CDAB0146R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0158R	CDAB0158R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0160R	CDAB0160R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0173R	CDAB0173R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0174R	CDAB0174R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0203R	CDAB0203R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0212R	CDAB0212R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0217R	CDAB0217R-PRE	*Eidolon helvum*	Rectal	Previously-extracted	W-Beta-3
CDAB0113RSV	CDAB0113RSV-PRE	*Hipposideros cf. ruber*	Rectal	Previously-extracted	W-Beta-4
CDAB0486R	CDAB0486R-PRE	*Chaerephon sp.*	Rectal	Previously-extracted	Q-Alpha-4
CDAB0488R	CDAB0488R-PRE	*Mops condylurus*	Rectal	Previously-extracted	Q-Alpha-4
CDAB0488R	CDAB0488R-TRI	*Mops condylurus*	Rectal	Trizol re-extraction	Q-Alpha-4
CDAB0491R	CDAB0491R-PRE	*Mops condylurus*	Rectal	Previously-extracted	Q-Alpha-4
CDAB0491R	CDAB0491R-TRI	*Mops condylurus*	Rectal	Trizol re-extraction	Q-Alpha-4
CDAB0492R	CDAB0492R-PRE	*Mops condylurus*	Rectal	Previously-extracted	Q-Alpha-4
CDAB0492R	CDAB0492R-TRI	*Mops condylurus*	Rectal	Trizol re-extraction	Q-Alpha-4
CDAB0494O	CDAB0494O-TRI	*Mops condylurus*	Oral	Trizol re-extraction	Q-Alpha-4
CDAB0494R	CDAB0494R-PRE	*Mops condylurus*	Rectal	Previously-extracted	Q-Alpha-4
CDAB0494R	CDAB0494R-TRI	*Mops condylurus*	Rectal	Trizol re-extraction	Q-Alpha-4
CDAB0495O	CDAB0495O-PRE	*Mops condylurus*	Oral	Previously-extracted	Q-Alpha-4
CDAB0495R	CDAB0495R-TRI	*Mops condylurus*	Rectal	Trizol re-extraction	Q-Alpha-4

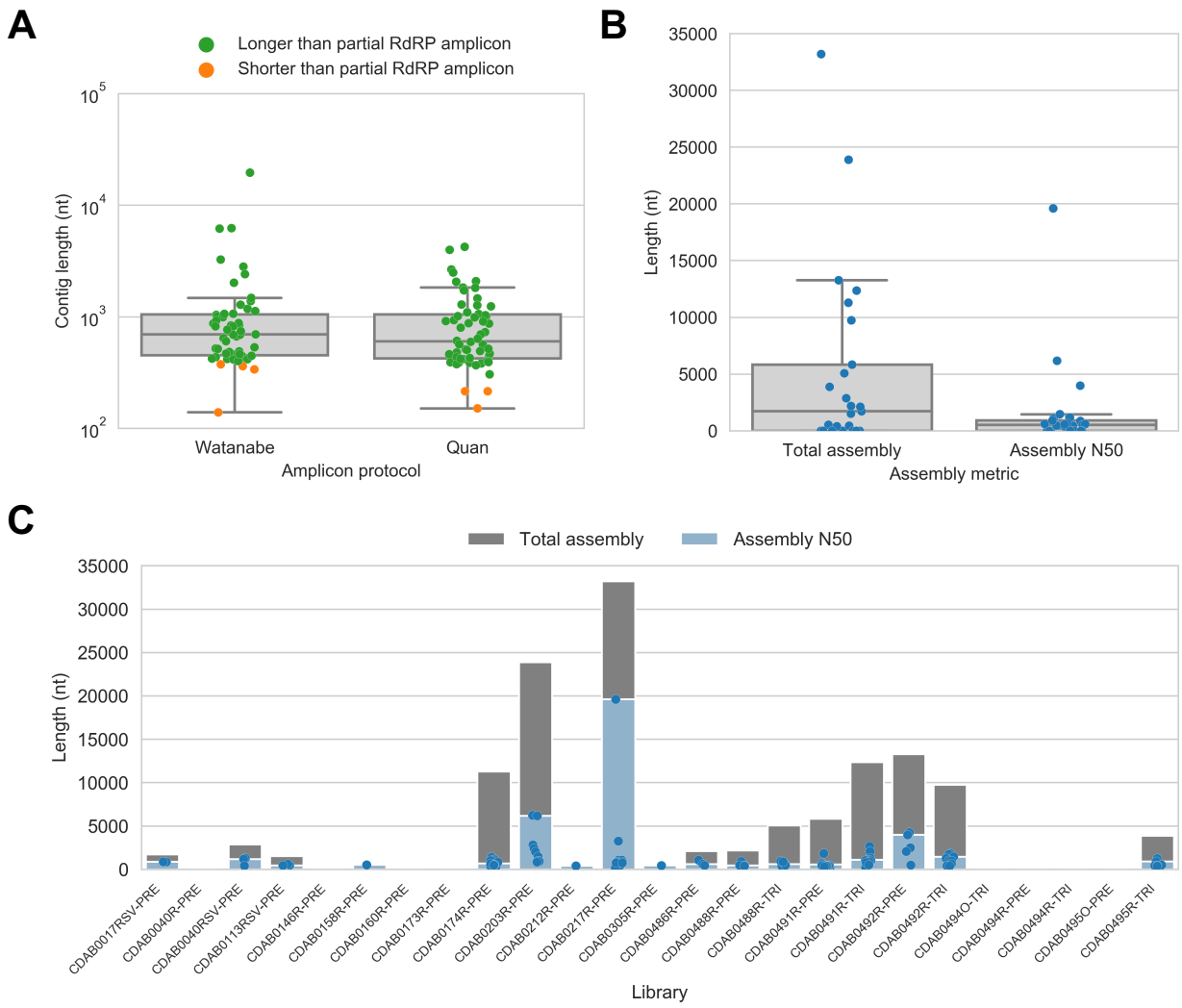

Figure 2: *De novo* assembly of probe captured libraries yielded more genome sequence than standard amplicon sequencing methods for most specimens. Reads from probe captured libraries were assembled *de novo* with coronaSPAdes, and coronavirus contigs were identified by local alignment against a database of all *coronaviridae* sequences in GenBank. **A)** The size distribution of contigs from all libraries is shown. Dots are coloured to indicate whether the length of the contig exceeded partial RNA-dependent RNA polymerase (RdRP) gene amplicons previously sequenced from these specimens. **B)** Total assembly size and assembly N50 distributions for all libraries. **C)** Each contig is represented as a dot plotted according to its length. Assembly N50 sizes and total assembly sizes are indicated by the height of their bars.

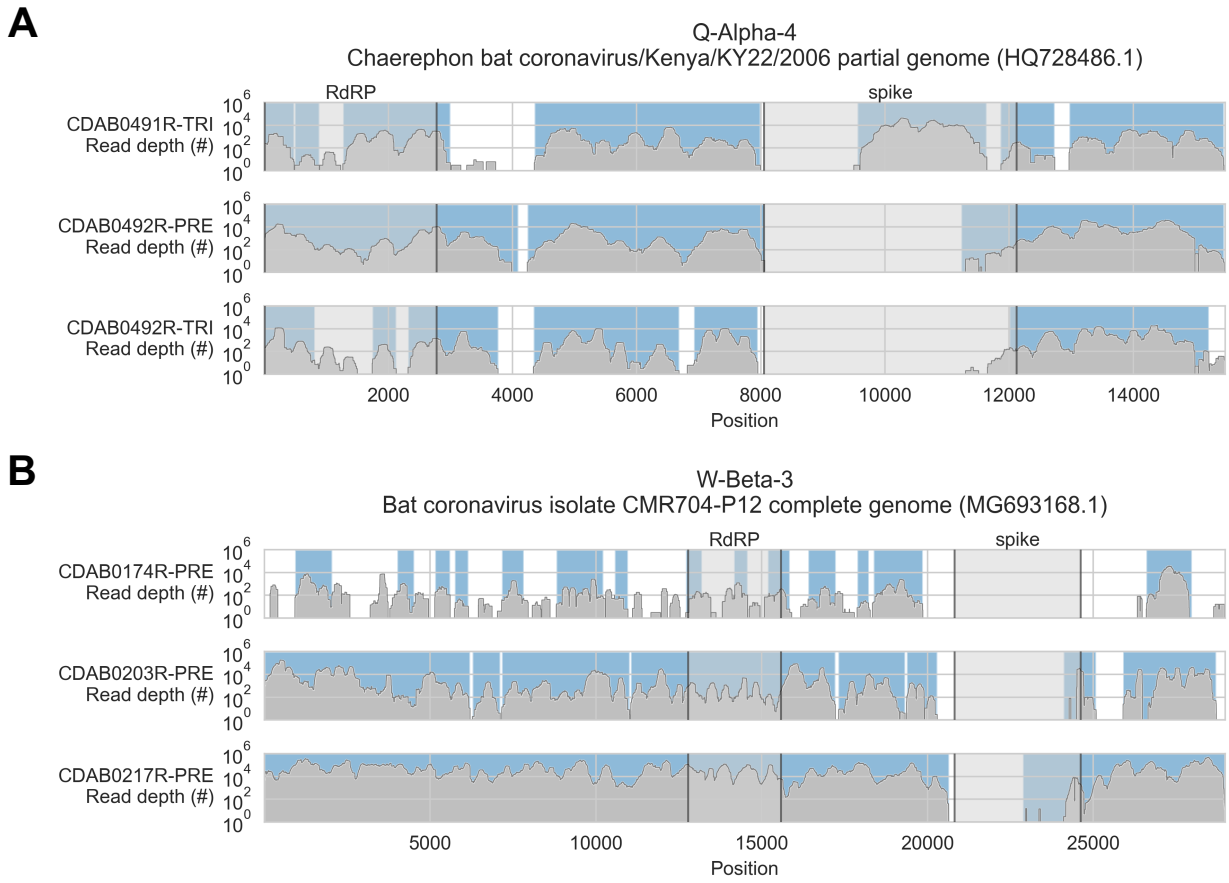

Figure 3: Coverage of reference sequences by probe captured libraries was used to assess extent and location of recovery. Reference sequences were chosen for each previously identified phylogenetic group (indicated in panel titles). Coverage of these reference sequences was determined by mapping reads and aligning contigs from probe captured libraries. Dark grey profiles show depth of read coverage along reference sequences. Blue shading indicates spans where contigs aligned. The locations of spike and RNA-dependent RNA polymerase (RdRP) genes are indicated in each reference sequence and shaded light grey. This figure shows the 6 libraries with the most extensive reference sequence coverage. Similar plots are provided as Figures S1-S4 for all libraries where any coronavirus sequence was recovered.

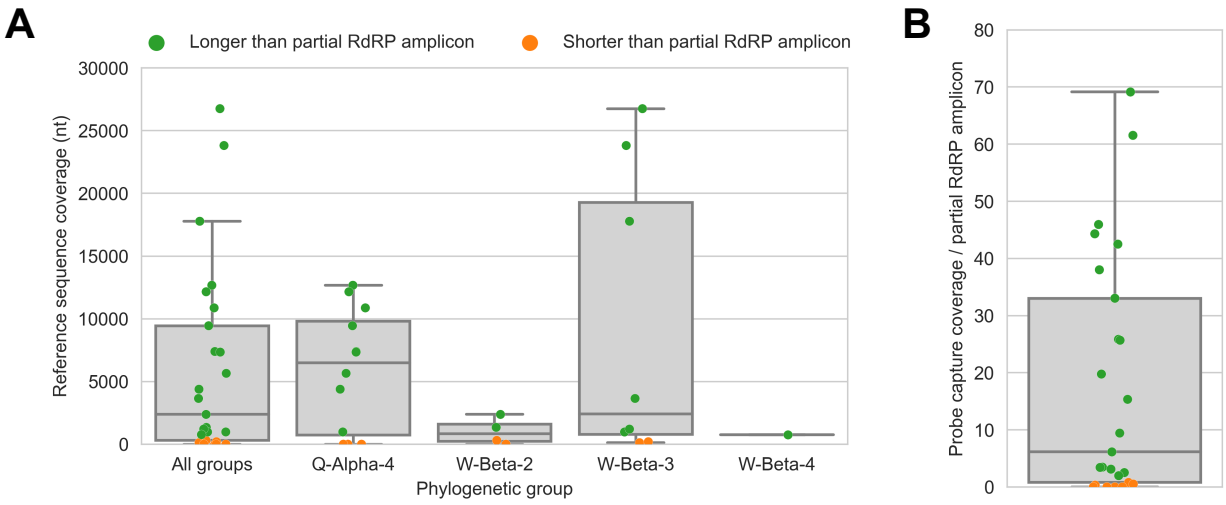

Figure 4: Probe captured libraries provided more extensive coverage of reference genomes than standard amplicon sequencing protocols for most specimens. Reference sequences were selected for the previously identified phylogenetic groups to which these specimens had been assigned by Kumakamba *et al.* (2020). **A)** Coverage of these reference sequences was determined by mapping reads and aligning contigs from probe captured libraries. Each library is represented as a dot, and dots are coloured according to whether reference sequence coverage exceeded the length of the partial RNA-dependent RNA polymerase (RdRP) gene sequence that had been previously generated by amplicon sequencing. **B)** The number of reference sequence positions covered by probe captured libraries was divided by the length of the partial RdRP amplicon sequences from these specimens. This provided the fold-difference in recovery between probe capture and standard amplicon sequencing methods.

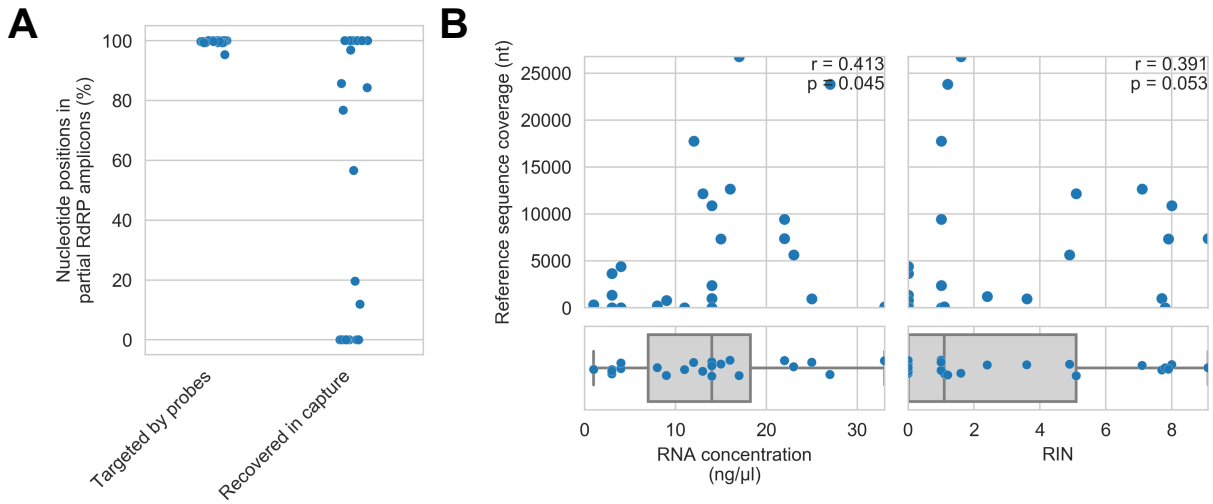

Figure 5: Recovery of CoV genomic material was limited *in vitro* by method sensitivity. A) Sensitivity was assessed by evaluating recovery of partial RNA-dependent RNA polymerase (RdRp) gene regions that had been previously sequenced in these specimens by amplicon sequencing. Probe coverage of partial RdRp sequences was assessed *in silico* to exclude insufficient probe design as an alternate explanation for incomplete recovery of these targets. **B)** Library input RNA from these specimens had low RNA concentrations and RNA integrity numbers (RINs). The impact of these specimen characteristics on recovery by probe capture (as measured by reference sequence coverage) was assessed using Spearman's rank correlation (test results stated in plots). An outlier was omitted from this analysis: RNA concentration for specimen CDAB0160R was recorded as 190 ng/µl, a value 4.7 SDs from the mean of the distribution.

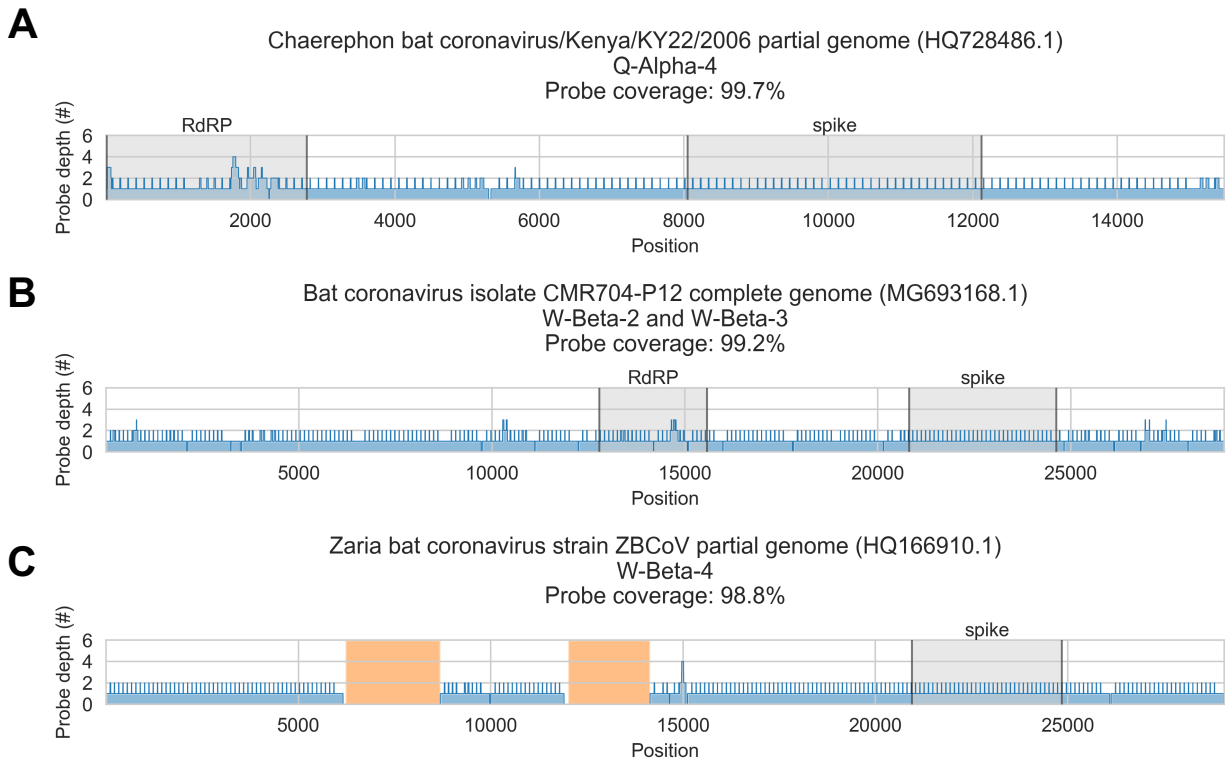

Figure 6: *In silico* assessment of probe panel coverage for reference genomes. Reference sequences were chosen for each previously identified phylogenetic group (indicated in panel titles). Blue profiles show the number of probes covering each nucleotide position along the reference sequence. Probe coverage, *i.e.* the percentage of nucleotide positions covered by at least one probe, is stated in panel titles. Ambiguity nucleotides (Ns) are shaded in orange, and these positions were excluded from the probe coverage calculations. The locations of spike and RNA-dependent RNA polymerase (RdRP) genes are indicated in each reference sequence (where available) and shaded grey.

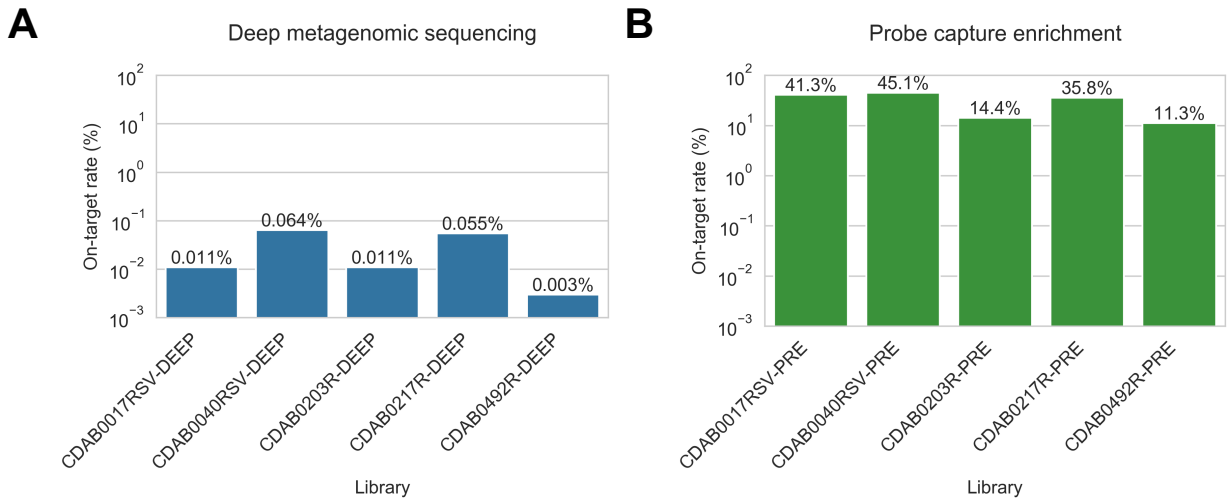

Figure 7: CoV genomic material was low abundance in swab specimens but effectively enriched by probe capture. A) Reads from uncaptured, deep metagenomic sequenced libraries were mapped to complete genomes recovered from these specimens to assess abundance of CoV genomic material. On-target rate was calculated as the percentage of total reads mapping that mapped to the CoV genome sequence. **B)** Reads from probe captured libraries were also mapped to assess enrichment and removal of background material. Most libraries used for probe capture (-PRE and -TRI) had insufficient volume remaining for deep metagenomic sequencing, so new libraries were prepared (-DEEP) from the same specimens.

Figure 8: Phylogenetic tree of translated spike gene sequences from *alphacoronaviruses.* Spike sequences are coloured according to whether they were from study specimens (blue), human CoVs (red), RefSeq (black), or GenBank (grey). Only the 25 closest-matching spike sequences from GenBank were included, as determined by blastp bitscores. GenBank and RefSeq accession numbers are provided in parentheses. The scale bar measures amino acid substitutions per site.

Figure 9: Phylogenetic tree of translated spike gene sequences from *betacoronaviruses.* Spike sequences are coloured according to whether they were from study specimens (blue), human CoVs (red), RefSeq (black), or GenBank (grey). Only the 25 closest-matching spike sequences from GenBank were included, as determined by blastp bitscores. GenBank and RefSeq accession numbers are provided in parentheses. The scale bar measures amino acid substitutions per site.

Table 2: Alignments between translated spike sequences from study specimens and phylogenetically proximate entries from GenBank and RefSeq. Alignments were conducted with blastp. Reference sequence host and collection location were obtained from GenBank entry summaries.

Specimen	Specimen host	Reference sequence GenBank accession number	Reference sequence host	Reference sequence collection location	Alignment query coverage (%)	Alignment identity (%)	Alignment positivity (%)
CDAB0492R	*Mops condylurus*	HQ728486.1	*Chaerephon sp.*	Kenya	100	71.2	80.1
CDAB0492R	*Mops condylurus*	MZ081383.1	*Chaerephon plicatus*	Yunnan, China	100	65.8	77.5
CDAB0017RSV	*Micropteropus pusillus*	HQ728482.1	*Eidolon helvum*	Kenya	99	76.5	85.7
CDAB0017RSV	*Micropteropus pusillus*	MG693168.1	*Eidolon helvum*	Cameroon	99	63.7	77.7
CDAB0040RSV	*Myonycteris sp.*	HQ728482.1	*Eidolon helvum*	Kenya	99	75.9	84.7
CDAB0040RSV	*Myonycteris sp.*	MG693168.1	*Eidolon helvum*	Cameroon	99	64.4	77.7
CDAB0203R	*Eidolon helvum*	HQ728482.1	*Eidolon helvum*	Kenya	100	73.7	85.3
CDAB0203R	*Eidolon helvum*	MG693168.1	*Eidolon helvum*	Cameroon	100	65.6	78.8
CDAB0217R	*Eidolon helvum*	HQ728482.1	*Eidolon helvum*	Kenya	100	73.5	85.1
CDAB0217R	*Eidolon helvum*	MG693168.1	*Eidolon helvum*	Cameroon	100	65.2	79.0

803 **Table 3: Nucleotide alignments between novel spike genes from study specimens and**
804 **phylogenetically related sequences from GenBank and RefSeq.** Alignments were conducted
805 with blastn. Discontinuous alignments are represented as multiple lines in the table, *e.g.*
806 CDAB0217R vs MG693168.1.

Specimen	Reference sequence GenBank accession number	Alignment query coverage (%)	Alignment identity (%)
CDAB0492R	HQ728486.1	60	81.0
CDAB0492R	MZ081383.1	18	71.5
CDAB0040RSV	HQ728482.1	83	75.4
CDAB0203R	HQ728482.1	78	75.5
CDAB0203R	MG693168.1	45	76.6
CDAB0217R	HQ728482.1	71	76.0
CDAB0217R	MG693168.1	47	75.7
CDAB0217R	MG693168.1	47	84.6

807