fortune Journals

Research Article

JOURNAL OF BIOINFORMATICS AND SYSTEMS BIOLOGY

ISSN: 2688-5107

Discovery of a Novel Merbecovirus DNA Clone Contaminating Agricultural Rice Sequencing Datasets from Wuhan, China

Adrian Jones[1], Daoyu Zhang[2], Steven E. Massey[3], Yuri Deigin[4*], Louis R. Nemzer[5], Steven C. Quay[6]

Abstract

HKU4-related coronaviruses belong to the same merbecovirus subgenus as Middle Eastern Respiratory Syndrome coronavirus (MERS-CoV), which causes severe respiratory illness in humans with a mortality rate of over 30%. The high genetic similarity between HKU4-related coronaviruses and MERS-CoV makes them an attractive subject of research for modeling potential zoonotic spillover scenarios. In this study, we identify a novel coronavirus contaminating agricultural rice RNA sequencing datasets from Wuhan, China. The datasets were deposited with NCBI by the Huazhong Agricultural University in early 2020. We were able to assemble the complete viral genome of a novel HKU4-related merbecovirus. The assembled genome is 98.38% identical to the closest known full genome sequence, *Tylonycteris pachypus* bat isolate BtTp-GX2012. Using *in silico* modeling, we show that the novel HKU4-related coronavirus spike protein likely binds to human dipeptidyl peptidase 4 (DPP4), the receptor used by MERS-CoV. We further show that the novel HKU4-related coronavirus genome has been inserted into a bacterial artificial chromosome in a format consistent with previously published coronavirus infectious clones. Additionally, we have found a near complete read coverage of the spike gene of the MERS-CoV reference strain, and identify the likely presence of a HKU4-related-MERS chimera in the datasets.

Keywords: HKU4-related CoV; MERS-CoV; Bacterial artificial chromosome; Reverse genetics system; Contamination.

Introduction

Coronaviruses (CoVs) are a large group of RNA viruses infecting a range of animals, including humans. The Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) epidemic in 2002-2003, the Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV) epidemic in 2012-2015 with sporadic outbreaks through 2022 [1] and the SARS-CoV-2 pandemic which began in late 2019 highlight the potential pathogenicity of coronaviruses to humans. MERS-CoV was first identified in 2012 [2], with 2,600 cases confirmed as of October 2022, and a case fatality rate of 36% [1]. The source of initial MERS-CoV human infection was traced to dromedary camels on the Arabian Peninsula [3,4]. Camels, however, are likely only intermediate hosts, with insectivorous bats the likely ancestral hosts of MERS-CoV [5].

MERS-CoV belongs to the merbecovirus (previously called lineage C) subgenus of the betacoronavirus genus of coronaviruses. Within merbecoviruses, in addition to the MERS-related CoV group, there are two

Affiliation:

[1]Independent bioinformatics researcher

[2]Independent genetics researcher

[3]Department of Biology, University of Puerto Rico - Rio Piedras, San Juan, PR 00931, USA;

[4]Youthereum Genetics Inc., Toronto, Ontario, Canada

[5]Department of Chemistry and Physics, Halmos College of Arts and Sciences, Nova Southeastern University, Ft. Lauderdale, FL, USA.

[6]Atossa Therapeutics, Inc., Seattle, WA 98104 USA

***Corresponding author:**
Yuri Deigin, Youthereum Genetics Inc., Toronto, Ontario, Canada.

Citation: Adrian Jones, Daoyu Zhang, Steven E. Massey, Yuri Deigin, Louis R. Nemzer, Steven C. Quay. Discovery of a Novel Merbecovirus DNA Clone Contaminating Agricultural Rice Sequencing Datasets from Wuhan, China. Journal of Bioinformatics and Systems Biology. 7 (2024): 139-156.

Received: April 05, 2024
Accepted: April 12, 2024
Published: July 29, 2024

other main phylogenetic groupings: HKU4-related CoVs and HKU5-related CoVs. HKU4-related CoVs were first identified in *Tylonycteris pachypus* bats in the Hong Kong Special Administrative Region [6], and have since been identified in *Tylonycteris* spp. across Southern China [7,8,9]. Notably, only two HKU4-related CoVs, Ty-BatCoV HKU4 SM3A, and recently MjHKU4r-CoV-1, have been documented as having been isolated, with the viruses replicating efficiently in human Caco-2 and Huh7 cells [10,11]. HKU5-related CoVs are hosted in *Pipistrellus abramus* bats, also found in Southern China [6].

The genomes of coronaviruses contain four genes coding for structural proteins: spike (S), envelope (E), membrane (M), and nucleocapsid (N). The coronavirus genome also contains six or more open reading frames (ORFs) including a large replicase gene (ORF1ab), 5' leader and untranslated region (UTR) and 3′ UTR and poly (A) tail [12]. The S protein is responsible for binding to a host cell receptor on the cell surface and facilitating host cell entry.

In coronaviruses, the trimeric S protein consists of S1 and S2 subunits in each monomer. In merbecoviruses and sarbecoviruses, the S1 subunit consists of an N-terminal domain and receptor-binding domain (RBD). The RBD of MERS-CoV binds to the human dipeptidyl peptidase 4 (hDPP4) receptor, which allows MERS-CoV to infect host cells [13]. Although the HKU4-CoV S protein RBD also binds to hDPP4, its binding affinity is less than that of MERS-CoV [14]. Notably HKU4-CoV has been demonstrated to infect hDPP4 transgenic mice [10]. HKU5-related CoVs however, lack the ability to bind to hDPP4 [14].

After the binding of the RBD to hDPP4, MERS-CoV relies on host cell protease cleavage to activate membrane fusion and gain cell entry [15,16]. Proteolytic cleavage occurs at two positions within the S protein, at both the boundary of the S1 and S2 subunits (S1/S2), and adjacent to the fusion peptide within the S2 subunit (S2′) [12,17,18]. Unlike MERS-CoV, HKU4-related CoVs do not possess furin cleavage sites at either the S1/S2 boundary or the S2' location, and cannot efficiently utilize endogenous human proteases for cell entry [14,19,20], although a furin cleavage site upstream of S1/S2 in pangolin-hosted MjHKU4r-CoV-1 has been proposed as facilitating human cell entry [11]. Notably [20] demonstrated that by introducing two mutations present in MERS-CoV into HKU4-CoV, including a mutation creating a furin cleavage site at the S1/S2 boundary, the HKU4-CoV S protein was able to gain the ability to enter human cells.

Sequence Read Archive (SRA) data submitted to the National Center for Biotechnology Information (NCBI) may contain contamination, either through library or sample contamination prior to sequencing, or index-hopping within multiplexed runs [21]. Indeed, [22] identified that in early 2020, over 2,000 Public Health England surveillance bacterial next-generation sequencing SRA datasets were likely contaminated with SARS-CoV-2 sequences. In some cases, the identification and characterization of cross-contaminating reads may provide useful biosecurity and or research insights [23,24,25,26].

NCBI BioProject PRJNA602160 contains 26 SRA datasets with *Oryza sativa* subsp. *japonica* (Japonica rice) DNA and RNA sequencing data (Supplementary Table S1). We identified several SRA datasets in this BioProject that contain HKU4-related and MERS-related CoV sequences. The datasets were generated by the Huazhong Agricultural University (HZAU), registered with NCBI on 2020-01-19 and published on 2020-02-09. BioProject PRJNA602160 utilized bisulfite sequencing and RNA sequencing to characterize DNA methylation patterns, and analysis of DNA glycosylases functionality in rice eggs, sperm cells, and zygotes. As a rice sequencing study, only rice genomic sequences, rice crop-hosted viruses, and bacterial sequences are expected to be present in the SRA datasets.

Using a bioinformatics workflow that included *de novo* assembly, as well as read and contig alignments, we identified a novel HKU4-related CoV genome sequence in four SRA datasets in rice sequencing BioProject PRJNA602160. We show that the novel HKU4-related CoV likely binds to hDPP4, potentially representing a human spillover risk. Additionally, we unexpectedly determined that the coronavirus genome is contained in a bacterial artificial chromosome (BAC) plasmid. This represents the first reverse genetics system documented for HKU4-related CoVs. In addition to the identification of a novel HKU4-related CoV in a BAC, we identified a near complete MERS-CoV spike sequence. We show the MERS-CoV spike was very likely substituted into the novel HKU4-related CoV backbone, representing a second clone in the datasets. Such research is indicative of enhanced potential pandemic pathogen (gain-of-function) research and we assess how this novel HKU4r-related clone may have contaminated agricultural rice sequencing datasets.

After our findings and analyses had been made, we were informed of another independent discovery of the HKU4-related CoV genome. This was achieved via large-scale viral alignment and assembly workflow conducted in 2020 [60]. Specifically, *de novo* assembled contigs in three SRA datasets in BioProject PRJNA602160 were found to match either BtCoV/133/2005 or Tylonycteris bat coronavirus HKU4 isolate CZ01 at between 97.6% to 98.3% identity. However, an in-depth characterization of the discovered HKU4-related CoV and attached vector sequences was not undertaken.

Materials and Methods

Datasets

Analysis of each SRA experiment in *Oryza sativa japonica* group sequencing BioProject accession PRJNA602160

Jones A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107086

registered on 2020-01-19 in the NCBI SRA database [27] was conducted using the NCBI SRA Taxonomy Analysis Tool (STAT), a k-mer-based taxonomic classification tool [22]. BtTp-BetaCoV/GX2012 (KJ473822.1) was identified as a genome with significant matches in four of the 26 SRA datasets in the BioProject: SRR10915167, SRR10915168, SRR10915173, and SRR10915174 (Supplementary Tables S1, S2).

SRA format datasets from NCBI were extracted as single fastq files using sratoolkit version 3.0.0 (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software).

Fastv [28] was run for each SRA against the Opengene viral genome kmer collection 'microbial.kc.fasta.gz' (https://github.com/OpenGene/UniqueKMER). Fastv extracts k-mers from reads and converts the result to keys. Keys are used to search against pre-built k-mer collections. Four datasets that contained HKU4-related CoV sequences at between 31% and 8% genome read coverage and three datasets that contained MERS-CoV sequences at between 8% and 5% genome coverage were identified (Supplementary Table S3).

Each SRA dataset in PRJNA602160 was filtered using fastp [29] with default settings, and filtered datasets were used for all subsequent analyses.

Alignment and Assembly

Read alignments were conducted using minimap2 version 2.24 [30] with the following parameters, unless indicated: "-MD -c -eqx -x sr --sam-hit-only --secondary=no -t 32".

Alignments of pooled reads from SRR10915167, SRR10915168, SRR10915173, and SRR10915174 to the MERS-CoV reference sequence HCoV-EMC/2012 (NC_019843.3) were made using bowtie2 version 2.4.2 [31] using the '--local' setting.

MEGAHIT v1.2.9 [32] with default settings was used for *de novo* assembly of each SRA dataset. To confirm the MEGAHIT assembly, we also undertook *de novo* assembly of SRR10915167-8 and SRR10915173-4 using coronaSPAdes v3.15.2 [33] using default settings, and SPAdes v3.15.2 [34] using the '--careful' parameter. Additionally, SRR10915167-8 and SRR10915173-4 were pooled and also assembled with coronaSPAdes and SPAdes with settings as per above.

The 38,583 nucleotide (nt) contig from the MEGAHIT *de novo* assembly of SRR10915173 'k141_13282' was trimmed to remove the first 149 nt. Reads from SRR10915167-8, and SRR10915173-4 were aligned to the resultant 38,434 nt trimmed contig 'k141_13282_del_149' using minimap2 (Supplementary Data S1-3). Additionally, pooled reads from the four SRA datasets above were aligned to the HKU4r-HZAU-2020 clone (i.e. contig 'k141_13282_del_149' containing the HKU4r-HZAU-2020 viral genome and attached vector sequences) using bowtie2 and bwa-mem2 [35] using default settings. Samclip (https://github.com/tseemann/samclip) with a maximum clip length allowed of 25 was used to clip minimap2 aligned reads for visualization and prior to genome alignment statistics calculations. Samclip removes reads from a sam file with soft or hard unaligned ends of read ends larger than the maximum length specified. This reduces the chances of misaligned reads. The choice of a maximum allowable length of 25 nt was arbitrarily chosen after testing values of 5, 25 and 30 nt.

MEGAHIT *de novo* assembled contigs and pooled reads from SRR10915167-8, and SRR10915173-4 were aligned to the MERS-CoV reference sequence (NC_019843.3) and the S gene of MERS-CoV using minimap2 (Supplementary Data S4-7).

Mapping statistics and coverages for all alignments were calculated using samtools version 1.15.1 [36] and bamdst version 1.0.9 [37].

Quality control

Reads in each of the four SRAs in BioProject PRJNA602160 containing merbecovirus sequences were aligned using minimap2 to the full HKU4r-HZAU-2020 clone (38,434 nt trimmed contig 'k141_13282_del_149'), as shown in Table 3. The four datasets were pooled and the combined reads again mapped to HKU4r-HZAU-2020 (Table 3).

The process was repeated using two additional short read aligners, bowtie2 and bwa-mem2, with similar results. Minimap2-aligned read depth distribution was fairly consistent across the full contig sequence for the SRA dataset with highest HKU4-related CoV read count, SRR10915173, except for three local regions: positions 1-71 with a read depth of 29-48; positions 1919-1973 with a read depth of 22-28 and located within the 5' UTR sequence 155 nt downstream from the 5' end of the HKU4r-HZAU-2020 viral genome; and between positions 32412 and 32521 immediately downstream of a bGH poly(A) signal and upstream of a CAP binding sequence, with a read depth between eight and 20 reads (Supplementary Figures S1 to S3). For the four pooled SRA datasets, slight differences in read depth were found between aligners betweens positions 32412 and 32521 of the HKU4r-HZAU-2020 clone with slightly higher read coverage over this section of the HKU4r-HZAU-2020 clone than SRA dataset SRR10915173 alone. Aligners minimap, bowtie2 and bwa-mem2 showed minimum read depths of 9, 10 and 14 reads respectively (Supplementary Figures S4 to S6). While we infer that this read depth is sufficient to indicate the likely presence of a single contiguous sequence, we cannot conclusively rule out misalignment of a second vector sequence trailing the bGH poly(A) signal.

Since the choice of assembler can affect the quality of betacoronavirus *de novo* assembly [38], two other *de novo* assemblers were tested and the results compared with the

MEGAHIT assembly. CoronaSPAdes [33] using default parameters and SPAdes [34] with the 'careful' setting, were used to assemble each of the four SRA datasets containing merbecovirus sequences. Using SPAdes careful assembly of SRR10915173, a 38,592 nt contig with 100% coverage and identity to the HKU4r-HZAU-2020 complete sequence was recovered (Supplementary Data S8). At the same time, *de novo* assembly of SRR10915173 using coronaSPAdes generated two contigs of lengths 32,683 nt and 5879 nt, which had a combined 100% coverage and 100% identity to the HKU4r-HZAU-2020 complete sequence (Supplementary Data S9).

To test the alignment to the HKU4r-HZAU-2020 viral genome sequence alone, each of the four SRAs with merbecovirus sequences, as well as a pooled set of these four datasets, were aligned to the HKU4r-HZAU-2020 viral genome sequence using minimap2, bowtie2 and bwa-mem2. In each of the four SRA datasets, at least one read crosses the 5' end of the genome, with soft-clipped overhangs upstream of the 5' end of the genome, all with a nearly identical sequence (Supplementary Figure S7). SRR10915173 has the most number of reads crossing the 5' end of the genome with 38x 150 nt reads, while SRR10915174 has four 150 nt reads covering this region.

Phylogenetic and recombination analyses

For all phylogenetic analyses, the following workflow was used. The HKU4r-HZAU-2020 genome and genome sub-regions were analyzed using NCBI blastn [39] to identify the 100 closest genomes. These were downloaded and aligned using MAFFT v7.490 [40] using the 'auto' parameter with default settings. A PhyML [41] maximum likelihood tree was generated with smart model selection (SMS) [42] using default settings. Tree branches distant from the query genome were pruned. A maximum-likelihood tree using the pruned genome set was then generated using PhyML with SMS using default settings. We repeated the phylogenetic analysis using the maximum likelihood method in MEGA11 [43] with best fit model selection based on lowest Bayesian information criterion. 1,000 bootstrap replicates were calculated for the full genome, S, and RdRp gene analysis. The respective topologies were consistent with those generated using PhyML (Supplementary Figures S8 to S10).

RDP4 [44] was used with the following algorithms to test recombination regions in the HKU4r-HZAU-2020 genome: RDP method [45], BOOTSCAN [46], MAXCHI [47], CHIMAERA [48], 3SEQ [49], GENECONV [50], LARD [51], and SISCAN [52]. The HKU4r-HZAU-2020 genome was tested against 13 closely related HKU4-related CoV genomes identified using blast against the nt database: BtTp-BetaCoV/GX2012; HKU4 isolate CZ07; HKU4 isolate CZ01; BtCoV/133/2005; HKU4 isolate SM3A; HKU4-4; HKU4-related isolate GZ131656; HKU4 isolate SZ140324; HKU4-2; HKU4-3; HKU4; Bat CoV isolate JPDB144 and Tr CoV isolate 162275. See Supplementary Table S4 for GenBank accession numbers.

Simplot analysis

The 12 HKU4-related CoV genomes with highest identity to the HKU4r-HZAU-2020 genome identified using blastn were aligned with the HKU4r-HZAU-2020 genome using MAFFT with the 'auto' parameter setting. Simplot++ [53] using default parameters was used to review sequence identity. The five closest genomes to the HKU4r-HZAU-2020 genome (BtCoV/133/2005; BtTp-BetaCoV/GX2012; HKU4 isolate CZ01; HKU4 isolate CZ07 and BtCoV HKU4 SM3A) were then shortlisted and re-plotted in Simplot++ using a 300 nt window, 30 nt step and Jukes-Cantor model.

Modeling the Receptor Binding Domain

The structure of the RBD for the novel HKU4-related CoV was modeled using the SWISS-MODEL web server [54] and aligned to PDB id: 4QZV using PyMOL Version 2.4 [55] (Supplementary Data S10). Contacts were identified using default distance settings. The binding free energy of the complex was calculated using the PRODIGY web server [56] and compared against that of the canonical complex PDB id: 4QZV.

Sequence annotation and restriction site mapping

Gene and vector sequence annotation was conducted using ApE version 3.1.4 [57] to identify ORFs and SnapGene version 6.2 [58] for feature annotation. Restriction enzyme site annotation was conducted in SnapGene version 6.2 (Supplementary Data S11-13).

Contamination and negative control

Reads in all SRAs in BioProject PRJNA602160 were aligned to the entire NCBI mitochondrial database downloaded on 2022-05-05, using minimap2 (Supplementary Table S5). Ribosomal RNA (rRNA) matching reads in the four SRAs containing HKU4r-HZAU-2020 sequences, SRR10915167-8 and SRR10915173-4 were identified using Metaxa2 version 2.2.3 [59] with default parameters. The reads were *de novo* assembled using MEGAHIT, and blastn was then used on the assembled contigs to identify closest identity in the NCBI nt database (Supplementary Table S6, Supplementary Data S14).

De novo assembled contigs were aligned using minimap2 to a concatenated virus database consisting of the NCBI viral database downloaded on 2021-11-28 and all CoVs on NCBI downloaded on 2020-03-30 by [60] (https://github.com/ababaian/serratus/wiki/Working-Data-Dir). Contigs from SRR10915173 were aligned using minimap2 to the NCBI nt database downloaded on 2021-11-27. A set of viruses with greater than 5% coverage after fastv read analysis and contig alignments to NCBI databases was then generated. Each SRA

dataset in PRJNA602160 was aligned using minimap2 to this selected set of viruses.

The same workflow was used for analysis of SRA datasets in BioProject PRJNA602115, used as a negative control.

Results

Identification of a novel merbecovirus

Of the 26 SRA datasets in BioProject PRJNA602160, Tylonycteris sp. bat CoV HKU4 sequence matches were identified in four datasets using the NCBI SRA Taxonomy Analysis Tool (STAT) [22], as tabulated in Supplementary Table S2. Tylonycteris bat CoV isolate BtTp-GX2012 was found to be the HKU4 strain with the highest percentage of matching reads in SRA datasets SRR10915168 and SRR10915173-4. To confirm the presence of HKU4-related CoV sequences in the four SRA datasets we undertook genome alignment analysis as described in the Methods section.

To extract a complete genome sequence, *de novo* assembly of each of the four SRA datasets in BioProject PRJNA602160 containing HKU4-related CoV sequences was conducted using MEGAHIT [32]. The resulting contigs were aligned to a combined set of the complete NCBI viral database and all coronaviruses on NCBI using minimap2 [30]. Of the 40 contigs matching HKU4-related CoVs or MERS-CoV, contig k141_13282 from assembly of SRA dataset SRR10915173 was found to have 100% coverage of the BtTp-BetaCoV/GX2012 genome. This contig was queried against the NCBI nucleotide (nt) database using NCBI BLAST (blastn) [39], and the highest identity match was found to be Tylonycteris bat coronavirus isolate BtTp-GX2012 (Supplementary Table S7). The four bat coronavirus genomes with highest identity to contig k141_13282 were then input as queries against contig k141_13282 as a subject using blastn. The overall identity of the three closest known genomes to the novel HKU4-related CoV was between 97.97% and 98.38% (Table 1). We named the novel HKU4-related CoV 'HKU4r-HZAU-2020' to reflect its phylogenetic affiliation, institutional source, and date of sequencing HKU4r-HZAU-2020 was assigned NCBI assession number OK560913.

Table 1: Blastn nucleotide sequence identity of the four most closely related sequences on GenBank to HKU4r-HZAU-2020. See Supplementary Table S4 for Genbank accession numbers.

Description	Max Score	Coverage	% identity
BtTp-BetaCoV/GX2012	53149	100%	98.38%
T. BtCov HKU4 isolate CZ07	52711	99%	98.14%
T. BtCov HKU4 isolate CZ01	52436	99%	97.97%
BtCoV/133/2005	48710	99%	95.67%

A BLAST search of HKU4r-HZAU-2020 against all HKU4-related sequences was conducted against the nt database on NCBI. Critical components of HKU4r-HZAU-2020, including the spike glycoprotein and the membrane glycoprotein, bear substantial differences when compared with other known sequences of HKU4 CoVs (Table 2, Supplementary Table S8).

Table 2: Nucleotide percentage identity of HKU4r-HZAU-2020 to the four most closely related sequences on GenBank for S and M gene sequences. See Supplementary Table S4 for Genbank accession numbers.

S gene	% identity	M gene	% identity
T. BtCov HKU4 isolate CZ07	97.59	BtTp-BetaCoV/GX2012	98.33
BtTp-BetaCoV/GX2012	96.62	HKU4 GZ131656	97.88
T. BtCov HKU4 isolate CZ01	96.53	BtCoV/133/2005	97.73
HKU4 GZ160421	96.23	HKU4-4	96.36

Prior to our analyses, which were conducted in 2021 and 2023, other researchers had already identified a HKU4-related CoV in three *de novo* assembled contigs in BioProject PRJNA602160 in 2020 [60]. Specifically, a 333 nt contig with a 97.9% identity to BtCoV/133/2005 was assembled from SRR10915168; a 961 nt contig with a 97.6% identity to BtCoV/133/2005 was assembled from SRR10915174; and a 38,489 nt contig with a 98.3% identity to T. BtCoV HKU4 isolate CZ01 was assembled from SRR10915173 [60]. The contig containing the complete viral genome is available at https://serratus.io/explorer/rdrp?run=SRR10915173.

Characterization of HKU4r-HZAU-2020 as a cDNA clone

The length of contig k141_13282 at 38,583 nt was significantly longer than HKU4-related or MERS-related CoV genomes, which are 30 kb to 33 kb in length. Reads in each of the four SRAs containing HKU4r-HZAU-2020 sequences were aligned to contig k141_13282, and the regions covering the 5' and 3' ends of the HKU4r-HZAU-2020 genome were analyzed using Addgene sequence analyser [61], Integrative Genomics Viewer (IGV) [62], and NCBI blastn. A human cytomegalovirus (CMV) immediate-early promoter sequence was identified immediately upstream of the 5' end of the HKU4r-HZAU-2020 genome (Supplementary Figure S11). Trailing the poly(A) tail at the 3' end of the genome, a hepatitis delta virus (HDV) ribozyme and bovine growth hormone (bGH) polyadenylation signal was found (Supplementary Figure S12). With a read depth of 40-60 over 200 nt regions centered on both the 3' and 5' junction locations, we infer that synthetic sequences were attached to the genome, with good confidence.

The first 316 nt of the 6343 nt sequence attached to the poly(A) tail at the 3' end of the HKU4r-HZAU-2020 genome exhibited the highest blastn maximum score match to vector pBAC-Beaudette-FU (92.11% identity, 97% coverage), which is a BAC plasmid containing infectious bronchitis virus [63]. The trailing 6027 nt was identified using blastn as having a 100% identity match to the BAC cloning vector pDEV-CHa [64]. Overall, however, the highest blast maximum score match we identified was to Sequence 11 from Patent WO2006236448 "Attenuated SARS and use as a vaccine" [65] with a 6329/6368 nt (99.39%) match (Supplementary Table S9). The matching section of Sequence 11 from Patent WO2006236448 forms part of a pBAC-SARS-CoV vector backbone [66,67,68]. The first 627 nt of the HKU4r-HZAU-2020 sequence trailing the poly(A) tail of the genome has a similar HDV ribozyme and bGH polyadenylation signal layout to the pBAC-SARS-CoV Sequence 11 trailing a SARS poly(A) tail (Supplementary Figure S13).

A 1912 nt section of contig k141_13282 was found upstream of the 5' end of the HKU4r-HZAU-2020 genome. Anomalous read depth and coverage was found over the first 149 nt of the sequence, and this region was trimmed from the contig. The resultant 1763 nt sequence was analyzed using blastn, with highest max score match to BAC constructs Gallid alphaherpesvirus 1 strain A489 and BAC cloning vector pDEV-CHa. The sequence was compared with pBAC-SARS-CoV Sequence 11 using blastn and Addgene sequence analyser (Supplementary Figure S14), with two regions exhibiting a 100% identity. A 956 nt sequence includes a lambda cos and a loxP site, which are commonly used for phage package and cre cleavage, respectively [69]. A 587 nt sequence includes a M13 forward primer and CMV promoter. The complete contig k141_13282 with 149 nt at the 5' end trimmed ('k141_13282_del_149') was annotated using SnapGene [58] (Figure 1). We refer to the recovered full length pBAC and genome sequence in contig k141_13282_del_149 as 'HKU4r-HZAU-2020 clone' and refer to the viral genome sequence as the HKU4r-HZAU-2020 genome or HKU4r-HZAU-2020 CoV.

The presence of T7 and CMV promoters before the 5'-end of the HKU4r-HZAU-2020 genome and a Hepatitis D virus ribozyme, followed by a bGH polyA signal after the 3'-end of the genome, indicates that the novel HKU4-related CoV obtained from SRR10915173 is probably infectious or intended to be infectious. This is evidenced by its format, which could generate full-length infectious RNA when expressed in mammalian cells.

The number of reads mapping to the full HKU4r-HZAU-2020 clone sequence for each SRA dataset and the four datasets pooled together is shown in Table 3.

Recombination and Simplot analysis

A recombination analysis using RDP4 [44], utilizing eight different methods, indicated only a single small 316 nt potential recombination fragment located between positions 26027 and 26343 in the HKU4r-HZAU-2020 genome (Figure 2), which was detected by four methods (Supplementary Table S11). Thus, there is no clear evidence that the HKU4r-HZAU-2020 genome CoV identified here could have been the result of simple recombination from known HKU4 genomes.

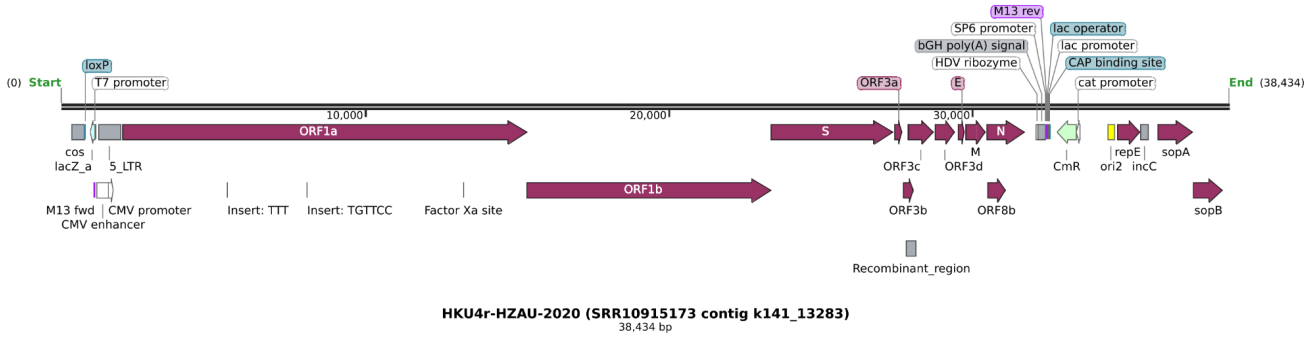

Figure 1: Annotation of the largest assembled HKU4-related contig sequence obtained from rice dataset SRR10915173. A 149 nt region at the 5' end of the sequence with anomalously high coverage was removed prior to annotation.

Table 3: Number of reads, coverage and average read depth for each of the four SRA datasets (*and pooled datasets) containing merbecovirus sequences in BioProject PRJNA602160 mapping to the full HKU4r-HZAU-2020 sequence (trimmed contig 'k141_13282_del_149'). Detailed statistics can be found in Supplementary Table S10.

SRA	Reference	Average depth	Coverage%	Reads mapped
4 pooled SRAs*	HKU4r-HZAU-2020	65.6	100	19952
SRR10915173	HKU4r-HZAU-2020	59.68	100	18309
SRR10915174	HKU4r-HZAU-2020	3.6	93.68	1004
SRR10915168	HKU4r-HZAU-2020	1.91	77.41	518
SRR10915167	HKU4r-HZAU-2020	0.4	32.97	121

Jones A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107086

Simplot analysis shows that three sequences: BtTp-BetaCoV/GX2012, HKU4 isolate CZ01 and HKU4 isolate CZ07 have high identity across the ORF1ab coding region, with CZ07 having the highest identity match in the S1 region of the spike gene (Figure 3). The ORF3a/3b gene sections of the genome exhibit low identity to the HKU4-related CoV genomes analyzed, in a region overlapping with and slightly larger than the potential recombination region discussed above.

Phylogenetic analysis

Maximum likelihood phylogenetic trees were generated for the full-length genome, spike protein coding sequence, RdRp gene, and partial RdRp gene. The HKU4r-HZAU-2020 genome forms a basal sister relationship to BtTp Beta-CoV/GX2012, HKU4 isolate CZ07 and HKU4 isolate CZ01 for the full genome (Figure 4).

For the spike gene, the HKU4r-HZAU-2020 genome forms a sister group to the HKU4 isolates CZ07/CZ01 and BtTp Beta-CoV/GX2012 clade (Figure 5). The sister clade to the HKU4r-HZAU-2020 genome, HKU4 isolates CZ07/CZ01 and BtTp Beta-CoV/GX2012 clade consists of three *Tylonycteris pachypus* HKU4 isolates, GZ1912, GZ1862 and GZ1832. These were collected in Guizhou province, China on 2015-09-09 and the sequences were submitted to Genbank on 2020-11-04 by [10].

Phylogenetic analysis of the S protein shows a similar grouping of the HKU4r-HZAU-2020 genome with the S proteins of HKU4 isolates CZ07/CZ01 and BtTp Beta-CoV/GX2012. One of only two documented HKU4 cell culture isolates, HKU4 SM3A, groups in a separate clade with BtCoV/133/2005 and HKU4, in both S gene and S protein phylogenetic trees (Figure 5; Supplementary Figure S15).

Receptor binding domain *in silico* analysis

The RBD of HKU4-related CoVs are known to bind to the human dipeptidyl peptidase 4 (hDPP4) receptor [19]. We identified the closest match to the HKU4r-HZAU-2020 RBD

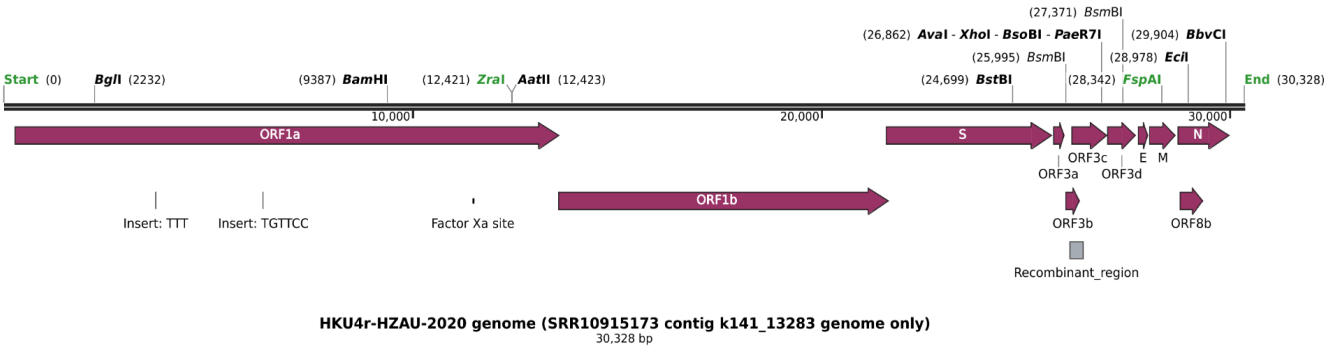

Figure 2: Structural and nonstructural protein and potential recombinant region locations and restriction mapping of the HKU4r-HZAU-2020 genome.

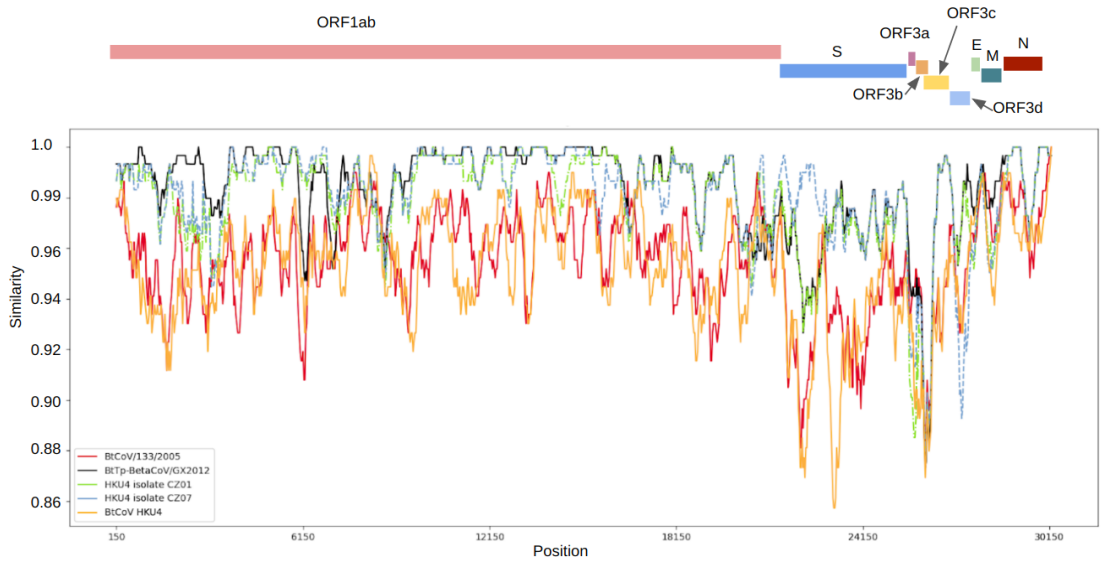

Figure 3: Similarity plot of the HKU4r-HZAU-2020 genome plotted against selected HKU4-related CoV genomes. HKU4 isolate CZ07 in light blue dashed pattern, HKU4 isolate CZ01 in green dash-dot pattern. Parameters: Window: 300 nt, step: 30 nt, model: Jukes-Cantor. Generated using Simplot++.

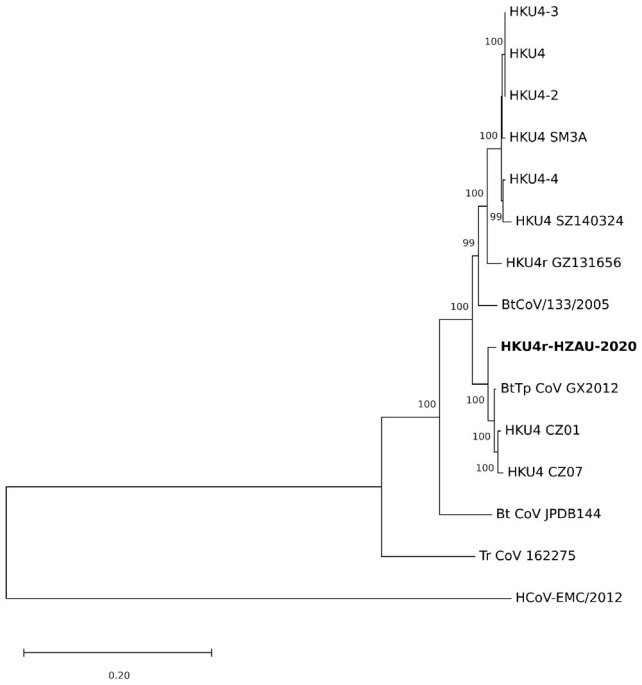

Figure 4: Maximum likelihood tree generated for selected HKU4-related CoV and MERS-CoV full genomes using a GTR+R model in PhyML using smart model selection [42]. Tree rooted on midpoint. The distance scale represents nucleotide substitutions per site. Branch support values of 70% or higher are shown at nodes. See Supplementary Table S4 for Genbank accession numbers.

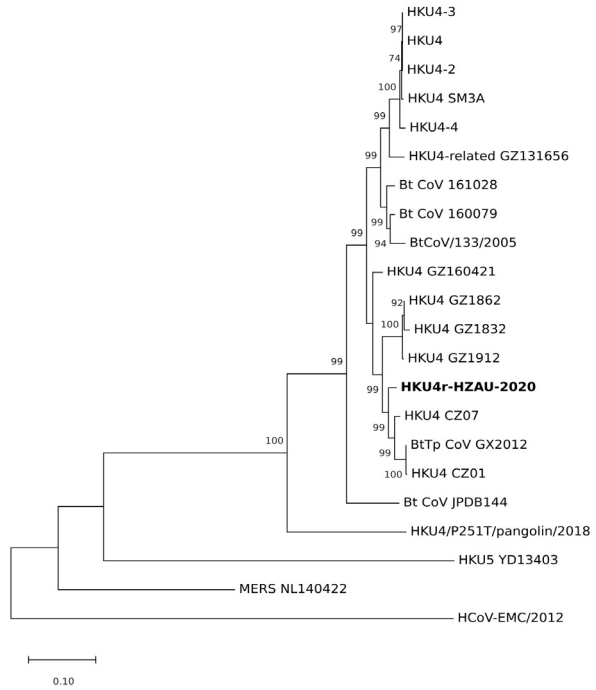

Figure 5: Maximum likelihood tree generated for HKU4-related CoV, MERS CoV and HKU5 CoV S genes using a GTR+G model in PhyML using smart model selection [42]. Tree rooted on midpoint. The distance scale represents nucleotide substitutions per site. Branch support values of 70% or higher are shown at nodes. See Supplementary Table S4 for Genbank accession numbers.

protein sequence on the NCBI nr database as PDB structure 4QZV:B [19]. To ascertain if the HKU4r-HZAU-2020 RBD was also likely to bind to hDPP4, we generated a structural model of the HKU4r-HZAU-2020 RBD using SWISS-MODEL and visualized its docking to human DPP4 using PyMOL [55] (Figure 7). High structural homology between the RBD from the HKU4r-HZAU-2020 clone and the RBD of PDB structure 4QZV was observed.

We then undertook molecular docking modeling in PRODIGY [56], which showed comparable binding energy of the two RBD molecules to hDPP4 (Table 4), indicating that the HKU4r-HZAU-2020 CoV obtained from SRR10915173 may be capable of binding to the hDPP4 receptor.

We compared five HKU4-related and three MERS-CoV sequences to HKU4r-HZAU-2020 in the key region of the RBD where interaction with hDPP4 occurs (Figure 8). Six of the thirteen key hDPP4 binding residues have an exact match to MERS-CoV residues, while eleven of the thirteen key residues match those found in PDB:4QZV_B.

MERS-CoV spike gene

From *de novo* assembly of the four pooled SRA datasets containing the novel HKU4-related CoV, we recovered two contigs of lengths 2,486 nt and 1,649 nt with 99.68% and 99.62% identity to the MERS-CoV reference strain HCoV-EMC/2012 (NC_019843.3), respectively (Supplementary

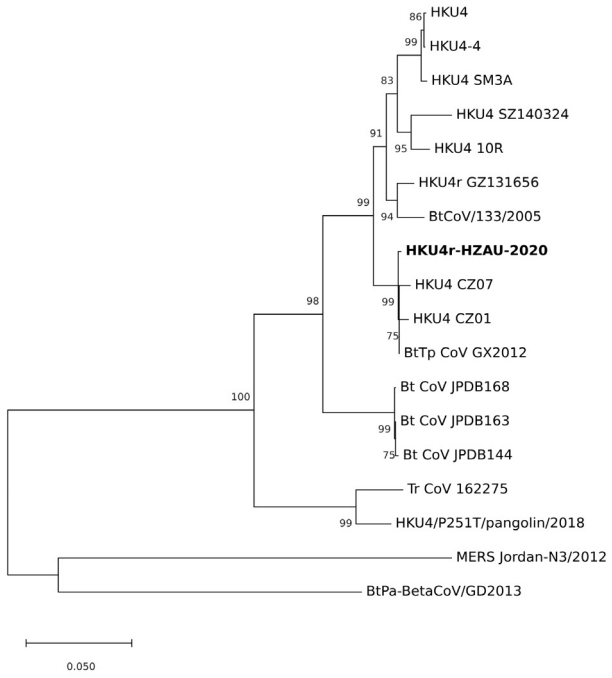

Figure 6: Maximum likelihood tree generated for selected HKU4-related CoVs, MERS-CoV and HKU5-related CoV RdRp genes using a GTR+G model in PhyML using smart model selection [42]. Tree rooted on midpoint. The distance scale represents nucleotide substitutions per site. Branch support values of 70% or higher are shown at nodes. See Supplementary Table S4 for Genbank accession numbers.

Jones A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107086

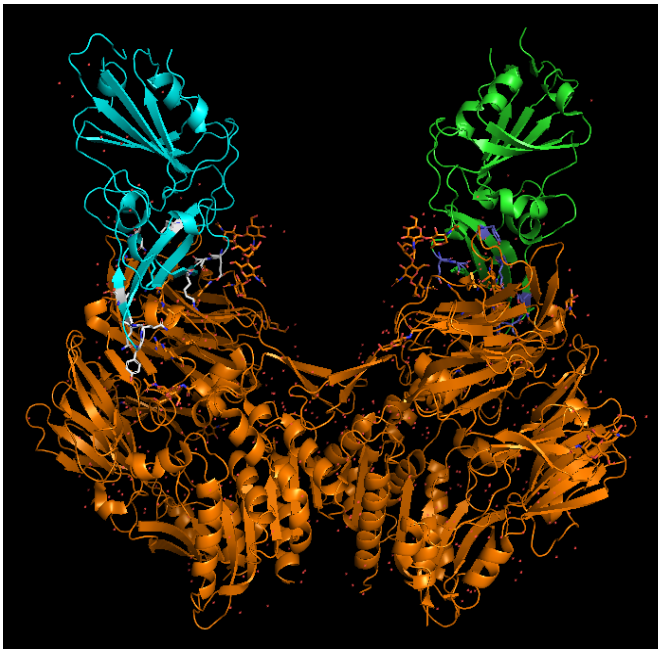

Figure 7: Alignment of the modeled RBD of the HKU4r-HZAU-2020 clone (light blue) to the HKU4 RBD in PDB (PDB ID 4QZV) (green) is consistent with HKU4r-HZAU-2020 RBD binding to hDPP4 (orange). Contacts between the RBD structures and hDPP4 were identified using PyMOL [55] (which were found to be 2.6 to 3.5 Angstroms in distance) and are indicated by blue sticks between the molecules.

Table 4: PRODIGY [56] binding energy and predicted binding affinity of HKU4r-HZAU-2020 spike protein RBD and hDPP4 as compared to the RBD of HKU4 (PDB ID 4QZV)

RBD protein	Binding free energy (kcal/mol)	Predicted Kd (M) at 25℃
HKU4r-HZAU-2020	-10.4	2.40E-08
4QZV:B	-10.5	2.00E-08

Table S13). Coverage of the MERS-CoV S gene sequence by reads is 99.09%, however, the average read depth is low at 6.13 reads. A consensus C21695T single nucleotide variation (SNV) (read depth of 12) with respect to HCoV-EMC/2012 was found (Figure 9). The only unmapped region of the MERS-CoV S gene is a 33 nt section located between positions 23,908 and 23,940 which is within the S2 subunit, upstream of the fusion peptide. The two 150 nt reads directly upstream of the missing section exhibit 100% identity to the MERS-CoV reference sequence, while the one 150 nt read directly downstream of the gap exhibits a SNV, not evident in other reads covering this position. No evidence of vector sequences in reads on either side of the gap are evident.

Only four reads cover the 3' most section of the spike sequence. Notably, all four reads have a 100% identity match to HKU4r-HZAU-2020 directly downstream of the MERS spike sequence (Supplementary Figure S17, Supplementary Data S5, S7). A 14 nt sequence at the 3' end of the MERS S gene is identical to the equivalent position in the HKU4r-HZAU-2020 S gene, with the exception of C25514T (MERS-CoV reference) in HKU4r-HZAU-2020 (Supplementary Figures S17, S19); as such, the exact position of the join of the MERS-CoV S gene with the HKU4r-HZAU-2020 backbone is uncertain.

At the 5' end of the MERS-CoV S gene, three 150 nt reads with a 100% identity match to HCoV-EMC/2012 at their 3' ends (44/44 nt, 78/78 nt and 125/125 nt) (Supplementary Figure S18), exhibit a 100% identity match to HKU4r-HZAU-20 at their 5' ends (106/106nt, 72/72nt and 25/25 nt, respectively) (Supplementary Data S5, S6). Two other reads aligning across the 5' end of the MERS-CoV S gene exhibit only a single SNV relative to HKU4r-HZAU-20 in a 25 nt section at their 5' ends match, while at their 3' ends a 36/36 nt match to the reverse complement of a 36 nt sequence at the 5' end of the HCoV-EMC/2012 S gene is found (Supplementary Figure S18). We infer the 36 nt sequences at their 5' ends of these two reads likely represent ligation or reverse transcription step artifacts.

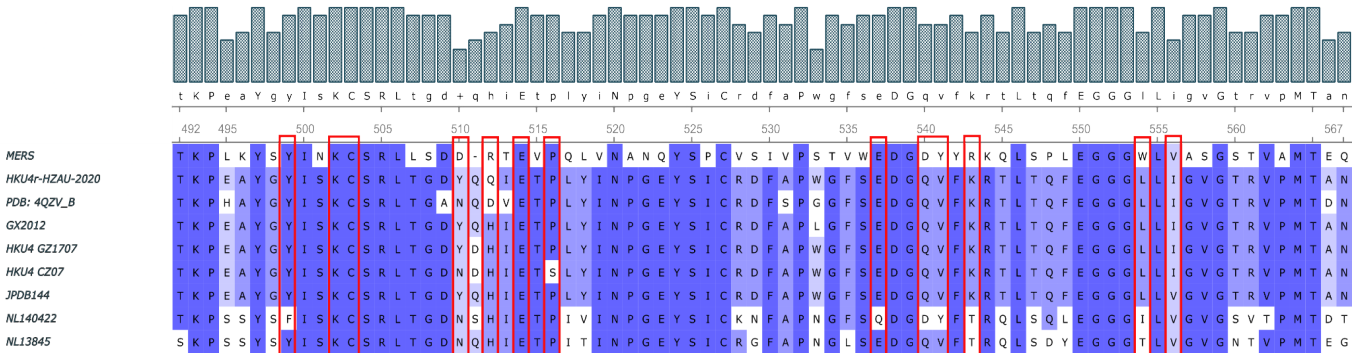

Figure 8: RBD subsequence alignment. Alignment of the external subdomain of the receptor binding domain for selected merbecoviruses. Key residue positions found to have direct interaction with the human DPP4 receptor [71] are marked with red bounding boxes. Plotted using UGENE [72]. For protein accession numbers see Supplementary Table S12.

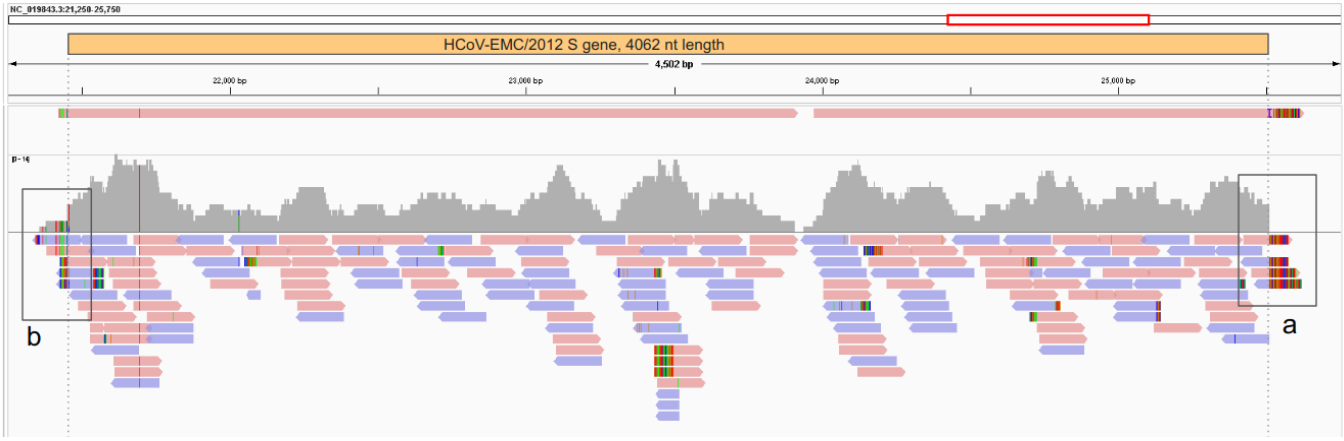

Figure 9: MERS-CoV genome alignment. In the top track pooled datasets SRR10915167, SRR10915168, SRR10915173 and SRR10915174 were *de novo* assembled using MEGAHIT and contigs aligned to the MERS-CoV reference genome HCoV-EMC/2012 (NC_019843.3). Middle track shows pooled read depth, scale 0-14. Bottom track shows pooled read alignments using minimap2. Boxes drawn around 5' end (b) and 3' end (a) of S gene and shown in Supplementary Figures S17 and S18.

As both the 3' and 5' ends of the MERS-CoV S gene sequence are attached to HKU4r-HAZU-2020 sequences and as the 33 nt gap in coverage discussed above occurs in a low read depth region we infer it likely that the gap is sequencing-related rather than a result of a deletion or a non-functional MERS-CoV S protein.

We infer that Golden Gate cloning was likely used for assembly of the HKU4r-HZAU-2020+S(MERS) chimera, since we did not identify any 1-8 cutter enzyme sequences in the MERS-CoV spike gene or HKU4r-HZAU-2020 in positions that could have been utilized for *in vitro* ligation. Additionally, we did not recover any synthetic vector sequences attached to the ends of the MERS-CoV spike sequences and no MERS-CoV sequences outside of the spike gene were found.

Restriction enzymes and synthetic engineering detection

In an attempt to find an obvious signature of genetic engineering in the HKU4r-HZAU-2020 genome, we performed a restriction enzyme mapping with SnapGene using the set of all type II and type IIS restriction endonucleases with unique sites. Since BsmBI and BsaI are often used for constructing coronavirus reverse genetic systems [73,74], we analyzed the number of these sites, as well as the number of unique restriction sites in the genome (Figure 2). For comparison, we also obtained and performed similar restriction site mapping of two related coronaviruses: BtTp-BetaCoV/GX2012 (Supplementary Figure S20) and HKU4 CZ07 (Supplementary Figure S21).

Several restriction sites were found to be conserved, including unique BglI, AatII, XhoI, EciI, and BvbCI sites between two closely related HKU4-related CoV sequences and the HKU4r-HZAU-2020 genome. However BsaI, BsmBI, and BamHI restriction sites were not conserved across these genomes. In the HKU4r-HZAU-2020 genome, there are only two BsmBI sites and a complete lack of BsaI sites. This appears anomalous, as all of the 14 most closely related HKU4-related CoVs, and MERS HCoV-EMC/2022 contain a minimum of five combined BsaI and BsmBI sites (Supplementary Figure S22, Supplementary Table S14).

HKU4r-HZAU-2020 host

To determine if some of the merbecovirus sequences may have originated from a bat sample, we performed a blastn search using the available nucleotide sequences of the bat *Tylonycteris pachypus*, the reservoir host of HKU4, against SRR10915173. We did not obtain any sequence that matched any nucleotide sequence from this species.

In addition, we aligned each SRA dataset in BioProject PRJNA602160 against all mitochondrial genomes present in Genbank. We used a minimum genome coverage of 10% to infer presence (Figure 10; Supplementary Table S5). Only *Danio rerio* (zebrafish), *Homo sapiens*, and three rice species *Oryza minuta*, *Oryza rufipogon*, and *Oryza sativa* had greater than 20% coverage, with several other plant, yeast, and parasites having between 10 and 20% mitochondrial genome coverage. For the four SRA datasets containing HKU4r-HZAU-2020 sequences, *Homo sapiens* mitochondrial genome coverage varied between 37.1% and 38.4%. However, no bat mitochondrial genome coverage of greater than 0.8% was detected in any of the 26 SRA datasets in BioProject PRJNA602160.

Homo sapiens mitochondrion ribosomal RNA matching contigs were found in all four datasets, comprising between 12.5% and 25% of *de novo* assembled contigs assembled from rRNA matching reads (Supplementary Table S6).

Jones A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107086

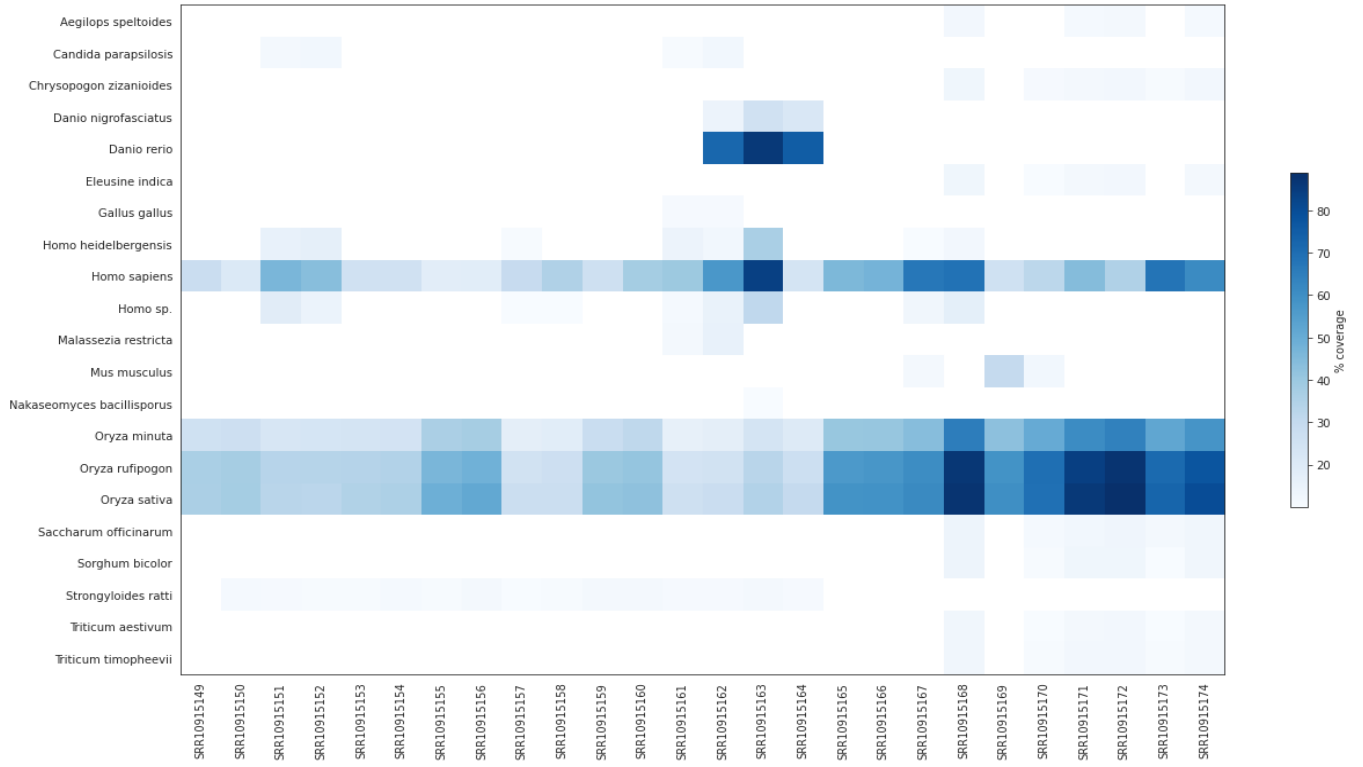

Figure 10: Mitochondrial genome coverage for each SRA dataset in BioProject PRJNA602160 aligned to all mitochondrial genomes on Genbank. A minimum genome coverage of 10% was used to infer presence. See Supplementary Table S1 for SRA dataset descriptions.

Virus sequence contamination

Fastp-filtered reads in each SRA dataset in BioProject PRJNA602160 were aligned to a set of viruses identified as present in the datasets as described in Methods. We identified six viral genomes and a synthetic sequence, each with greater than 10% coverage in one or more SRA datasets in BioProject PRJNA602160 (Figure 11; Supplementary Figure S23). As discussed above, HKU4r-HZAU-2020 CoV and HCoV-EMC/2012 were only identified in four SRA datasets. These were the only RNA-sequencing datasets in PRJNA602160 using cDNA selection and a unicellular zygote source (Supplementary Table S1).

Human endogenous retrovirus H HERV-H/env62 sequences were identified in all SRA datasets. Murine hosted viruses were found in six SRA datasets but not obviously correlated with the presence of HKU4r-HZAU-2020.

The bacterial content in the four RNA-sequencing datasets containing HKU4r-HZAU-2020 sequences identified using NCBI STAT [22] analysis averaged 58.14%, with a range of 40.98% to 83.41% (Supplementary Table S2), whereas the average bacterial content for the six RNA-sequencing datasets in BioProject PRJNA602160 that did not contain HKU4r-HZAU-2020 sequences averaged 1.65%, and ranged from 0.14% to 4.03%. We calculated a very strong Spearman correlation coefficient of 0.81 between bacterial percentage and percentage of reads mapping to the HKU4r-HZAU-2020 viral genome in RNA-sequencing datasets in BioProject PRJNA602160 (Supplementary Code).

Negative control

We identified that RNA-Sequencing BioProject PRJNA602115, a zebrafish miR-462-731 regulation study [75] was submitted by the HZAU to NCBI on the same day as BioProject PRJNA602160. Using it as a control dataset, we ran the same workflow we used for identifying and aligning HKU4r-HAZU-2020 on the three SRA datasets in BioProject PRJNA602115. Only a small number of viruses were identified using fastv, all with low coverage (Supplementary Table S15). We used minimap2 to align the three SRA datasets to the same set of HKU4-related CoVs, MERS-CoV, and miscellaneous viruses used for PRJNA602160 contamination identification, as well as the three viruses identified using fastv with more than 3% coverage (Supplementary Table S16). Only Sugarcane mosaic virus (NC_003398.1) and the Harvey murine sarcoma virus p21 v-has protein gene (NC_038668.1) were found to have more than 10% coverage. No trace of HKU4-related CoVs or MERS-CoV was found. Additionally, we aligned the three SRA datasets to all mitochondrial sequences on NCBI (Supplementary Table S17). As expected, the *Danio rerio* mitochondrial gene was mapped with near complete coverage in all datasets. We additionally identified *Ochotona curzoniae* and *Rattus norvegicus* genomic sequence contamination in the datasets.

Jones A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107086

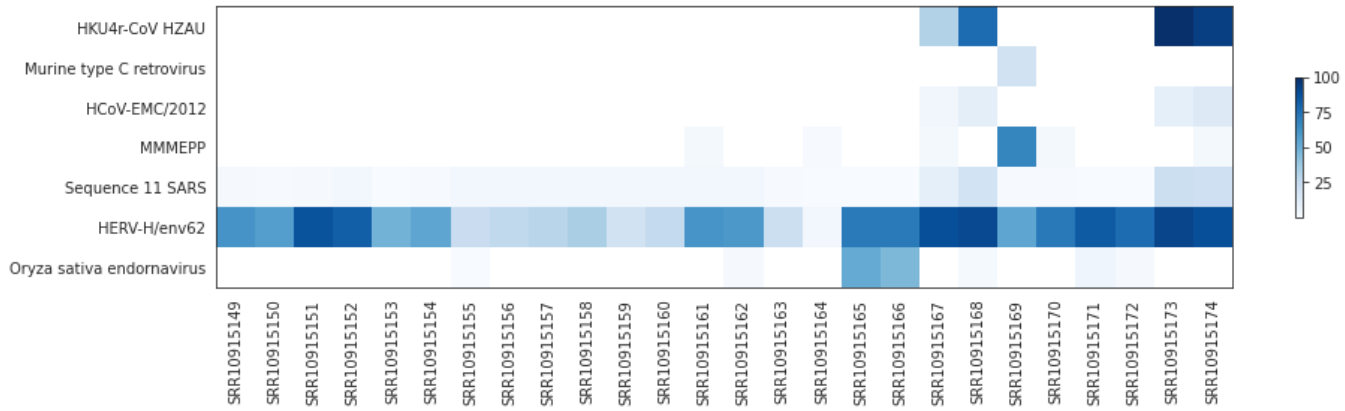

Figure 11: Percentage genome coverage for six viral genomes and one SARS BAC for each SRA dataset in PRJNA602160. A minimum cutoff of 10% genome coverage in any SRA was applied.

Discussion

SARS-related and MERS-related coronaviruses (CoVs) with zoonosis potential have been a subject of intense research since the SARS-CoV and MERS-CoV spillovers to humans in 2002 and 2012, respectively. In addition to MERS-related CoVs, the two other groups belonging to the merbecovirus subgenus of betacoronaviruses are HKU4-related and HKU5-related CoVs. Here we have identified a full-length HKU4-related CoV cDNA clone contaminating an *Oryza sativa japonica* agricultural sequencing BioProject. The average read depth of 66 for the clone in pooled datasets is good, and we have high confidence in both the presence of the clone as well as its sequence.

Our assembly and analysis approach differs from [60] who independently identified the presence of a HKU4-related CoV genome in BioProject PRJNA602160. The researchers used a large scale data mining approach using RdRp viral classification and de novo assembly of identified coronaviruses using coronaSPAdes [60]. The authors proposed that the most parsimonious explanation was that the HKU4-related CoV found in the rice sequencing dataset was due to contamination by feces or fertilizer sourced from a mammalian host. While the novel CoV was identified as an HKU4-related CoV, an in-depth characterization of the CoV and attached vector sequences was not undertaken.

HKU4-related CoVs are only known to be hosted by Tylonycteris bats (Vespertilionidae family, Vespertilioninae subfamily), a rare insectivorous genus only known to live in bamboo forests, and in China found in the Guangdong, Guangxi, Yunnan, Guizhou, Hong Kong, and Macau regions of southern China [76]. HKU4r-HZAU-2020, the novel HKU4-related CoV identified here, phylogenetically groups with the Tylonycteris pachypus isolate BtTp-BetaCoV/GX2012 sampled in the Guangxi province and the Tylonycteris bat HKU4 isolates CZ07 and CZ01 over the full genome, S and RdRp genes. HKU4 isolates CZ07 and CZ01 were sampled in southern China in 2012 by researchers affiliated with the Wuhan Institute of Virology (WIV) (Supplementary Table S4). BtTp-BetaCoV/GX2012, which exhibits highest identity over the full genome, was sampled by [8] in the Guangxi province in 2012 (Supplementary Table S4). Tylonycteris bat coronavirus HKU4 isolate 152762, with closest partial RdRp sequence grouping with HKU4r-HZAU-2020, was sampled between 2010 and 2015 by [70] as part of research co-authored by WIV-affiliated researchers (Supplementary Table S4). Consequently, we infer that the HKU4r-HZAU-2020 genome is either a *Tylonycteris* sp. hosted CoV sampled by researchers from southern China, or is a consensus genome generated from *Tylonycteris* spp. hosted CoVs.

Prior to the identification of the HKU4r-HZAU-2020 DNA clone, no reverse genetics systems were previously published for any HKU4-related CoVs. It is interesting to note that a 5-year grant awarded to EcoHealth Alliance by the National Institute of Health involved chimeric MERS-related CoV construction at the WIV [77]. Grant award year 5 (2018-06-01 – 2019-05-31) included generating full-length infectious clones of MERS-CoV with the RBD replaced by those from various HKU4-related CoVs and assessing the infectivity of novel chimeras in human cells. Two full MERS-related CoV genomes from Guangdong province, BtCoV/li/GD/2013–845 and BtCoV/li/GD/2014–422, were published by [71] and discussed in [77], while notably 24 HKU4-related CoV sequences were not published.

We find that instead of using a published MERS-CoV backbone and inserting novel HKU4-related CoV fragments [77], a completely undocumented HKU4-related CoV clone was apparently being researched in Wuhan. The discovery of a near complete MERS-CoV spike sequence in the same BioProject as HKU4r-HZAU-2020 is concerning. Although we identified only a low number of reads mapping across the 5’ and 3’ ends of the MERS spike gene sequence, these reads clearly indicate that a MERS-CoV S gene sequence had been inserted into the HKU4r-HZAU-2020 backbone. The upstream section of reads mapping across the 5’ end of the

MERS S gene had a 100% match part of the ORF1b gene of HKU4r-HZAU-2020 genome directly upstream of its spike gene. Similarly, the downstream section of reads mapping across the 3' end of the MERS-CoV S gene had a 100% match to a non-coding region of the HKU4r-HZAU-2020 genome directly downstream of its spike gene. The probability that such a configuration is artefactual we believe is extremely remote. As such, we infer it highly probable that a second HKU4-related CoV clone 'HKU4r-HZAU-2020+S(MERS)', with a replacement of the spike sequence by a MERS-CoV spike is present in the BioProject, albeit at lower abundance than HKU4r-HZAU-2020.

The lack of BsaI sites and the low combined count of BsaI and BsmBI sites are both anomalous in HKU4r-HZAU-2020 relative to the 14 other HKU4-related CoVs. We infer it is likely that "No See'm" cloning [78] or Golden Gate assembly [79] was used to assemble the HKU4r-HZAU-2020 genome fragments, whereby fragments can be ligated in one step for a seamless assembly [80,81].

This differs from the approach proposed by [82] for assembly of SARS-CoV-2 whereby BsmBI and BsaI TypeIIS restriction enzymes were oriented to allow directed assembly leaving restriction enzyme recognition sequences in the final assembled genome. This also differs from the approach used for the assembly of the WIV1 DNA clone using Type II BglI restriction endonucleases [83].

Merbecoviruses were actively researched in Wuhan, predominantly by the WIV with documented plasmid construction, genetic engineering and viral entry studies [20,71,77,84,85,86]. Indeed we note that the synthetic vector backbone of pBAC-SARS-CoV, which best matches the synthetic sequences present at the 3' end of HKU4r-HZAU-2020 and shows 100% sequence matches of 956 nt and 587 nt lengths at the 5' end, was used previously by researchers at the WIV [87]. However, merbecovirus research has also been conducted by HZAU researchers [88,89] and Wuhan University [90,91]. As the WIV outsourced CoV sequencing to Wuhan sequencing centers [92], the originating laboratory for the HKU4r-HZAU-2020 clone is uncertain. The BioProject for the *Oriza sativa japonica* sequencing datasets was registered in GenBank on 2020-01-19, but to this date the novel HKU4-related CoV has not been published by the HZAU, WIV, or Wuhan University.

Cross-contamination of samples is a well-documented challenge in biological sequencing laboratories, as noted by several studies including those by [93] and [94]. In particular, [95] have highlighted several key stages in the sequencing process where contamination can occur, including sample collection and shipping, where "passenger" viruses can cross-contaminate, as well as during virus purification, nucleic acid isolation, amplification, and library preparation, where there is a risk of reagent contamination and amplification bias. Additionally, contamination may also occur due to human error, such as improper use of protective equipment or failure to follow established protocols, or due to laboratory infrastructure issues, such as poor ventilation systems or inadequate physical separation of samples. Contamination can also occur during sequencing itself, through machine contamination and index hopping.

HKU4r-HZAU-2020 clone is manifestly a full-length cDNA clone engineered as a BAC, and as such would be unrelated to RNA contamination. We found that all RNA-sequencing datasets in rice BioProject PRJNA602160 without merbecovirus sequences had a maximum of 4% bacteria, while the minimum quantity of bacteria contaminating the four SRA datasets containing merbecovirus reads was 41%. Indeed we found a very high Spearman correlation coefficient between bacterial content and the number of reads mapping to the HKU4r-HZAU-2020 genome in RNA-sequencing datasets in the rice BioProject (Supplementary code). This finding is strongly indicative of upstream contamination prior to sequencing. Furthermore, the RNA-sequencing of samples in BioProject PRJNA602160 involved a specialized library format and protocol. As such, we infer it unlikely that contamination with plasmid sequences was due to index hopping from a multiplexed run with another library prepared using the same format and pooled in the same run as samples in BioProject PRJNA602160. We propose that the most likely scenario of contamination of SRR10915167-8 and SRR10915173-4 is HKU4r-HZAU-2020 plasmids contaminating the samples prior to sequencing. Similarly, we infer that contamination by MERS spike gene sequences likely occurred via HKU4r-HZAU-2020+S(MERS) plasmids also contaminating the samples prior to sequencing.

The potential for bat-hosted coronaviruses to spill over to humans and spark epidemics has been widely documented and is a subject of extensive research [10,73,96,97]. Our finding that HKU4r-HZAU-2020 likely binds to hDPP4 adds to previously documented research that three HKU4-related CoVs are able to bind to hDPP4. HKU4 strain B04f from the Guangdong province, HKU4 strain SM3A, and pangolin hosted MjHKU4r-CoV-1 are all able to bind hDPP4 [10,11,14,19,98]. However, HKU4-related-CoVs including HKU4r-HZAU-2020 do not possess a human proprotein convertase cleavage motif at the S1/S2 boundary. As such, we infer that HKU4r-HZAU-2020 may be incapable of mediating sustained human cell entry without the evolution of, or introduction of a furin cleavage motif [20].

A topic which receives far less attention than zoonotic spillover is the emerging threat of accidental human infection during laboratory research of wild-type viruses, or research involving enhanced potential pandemic pathogens [99,100,101]. Although the exact nature of the research related to the HKU4r-HZAU-2020 clone is unknown, there are two experimental purposes in generating infectious

clones of coronaviruses: to study a wildtype virus for which no 'live' isolate exists but for which a genome sequence is available, or to manipulate the genome of a coronavirus in order to study changes in infectivity, tropism (including host or tissue specificity), or pathogenicity.

To clarify the circumstances under which HKU4r-HZAU-2020 and HKU4r-HZAU-2020+S(MERS) sequences could have come to contaminate BioProject PRJNA602160, we contacted its corresponding author via email but did not receive a response.

Conclusion

In this work, we have reported the unexpected discovery of a HKU4-related coronavirus (CoV) clone in agricultural rice sequencing data from the Huazhong Agricultural University in Wuhan. The novel HKU4r-HZAU-2020 genome exhibits a 98.38% identity to the closest published HKU4-related CoV, BtTp-BetaCoV/GX2012. The regions of most significant difference to the most closely related HKU4-related CoVs are in the spike, ORF3a and ORF3b coding regions. We modeled the interface of the RBD of the novel HKU4-related CoV to human dipeptidyl peptidase 4 (hDPP4) and found that HKU4r-HZAU-2020 likely binds to human cells. This represents the third known HKU4-related CoV with hDPP4 binding potential.

The HKU4r-HZAU-2020 genome was found to have been inserted into a bacterial artificial chromosome. T7 promoter and cytomegalovirus promoter sequences were identified upstream of the genome, likely to facilitate in vitro transcription of full-length RNA copies of the genome. Downstream of the 3' end of the poly(A) tail, a hepatitis delta virus ribozyme and bovine growth hormone termination and polyadenylation sequence were found, which would allow truncation of the 3' end of the transcribed viral genome. This constitutes the first known reverse genetics system used for HKU4-related CoV research.

We further identified the presence of a near complete MERS-CoV spike sequence which had been inserted into the HKU4r-HZAU-2020 backbone forming a second clone in the rice RNA-sequencing datasets. As the MERS-CoV RBD binds more efficiently to hDPP4 than known HKU4r-CoVs, and as the MERS-CoV S protein has the demonstrated capability of utilizing human cell proteases for mediating cell entry, the HKU4r-HZAU-2020+S(MERS) chimera appears to constitute enhanced potential pandemic pathogen (gain-of-function) research.

We infer the most likely source of contamination of the rice RNA-sequencing datasets was by bacterial plasmids upstream of the sequencing step.

Finally, this work serves as a cautionary story to show that the absence of a documented related reverse genetic system cannot be relied upon as evidence to disprove laboratory involvement of a novel viral pathogen.

Acknowledgements

We thank Francisco A. de Ribera, Valentin Bruttel, Adrian Gibbs and Jonathan Latham, as well as an anonymous reviewer, for review and helpful comments.

Data Availability

Supplementary information and data are available on Zenodo:

doi: 10.5281/zenodo.7633113

link: https://zenodo.org/records/8351689

Supplementary Data

https://www.fortunejournals.com/suppli/suppliJBSB_10555.zip

References

1. World Health Organization. Middle East respiratory syndrome coronavirus (MERS-CoV) (2022).
2. Bermingham A, Chand MA, Brown CS, et al. Severe Respiratory Illness Caused by a Novel Coronavirus, in a Patient Transferred to the United Kingdom from the Middle East. Eurosurveillance 17 (2012).
3. Alraddadi BM, Watson JT, Almarashi A, et al. Risk Factors for Primary Middle East Respiratory Syndrome Coronavirus Illness in Humans, Saudi Arabia, 2014. Emerg Infect Dis 22 (2016): 49–55.
4. El-Kafrawy SA, Corman VM, Tolah AM, et al. Enzootic Patterns of Middle East Respiratory Syndrome Coronavirus in Imported African and Local Arabian Dromedary Camels: A Prospective Genomic Study. Lancet Planet Heal 3 (2019): e521–e528.
5. Lau SKP, Fan RYY, Luk HKH, et al. Replication of MERS and SARS Coronaviruses in Bat Cells Offers Insights to Their Ancestral Origins. Emerg Microbes Infect 7 (2018): 1–11.
6. Woo PCY, Lau SKP, Li KSM, et al. Molecular Diversity of Coronaviruses in Bats. Virol 351 (2006): 180–187.
7. TangXC, Zhang JX, Zhang SY, et al. Prevalence and Genetic Diversity of Coronaviruses in Bats from China. J Virol 80 (2006): 7481–7490.
8. Wu Z, Yang L, Ren X, et al. Deciphering the Bat Virome Catalog to Better Understand the Ecological Diversity of Bat Viruses and the Bat Origin of Emerging Infectious Diseases. ISME J 10 (2016): 609–620.
9. Li B, Si H-R, Zhu Y, et al. Discovery of Bat Coronaviruses through Surveillance and Probe Capture-Based Next-Generation Sequencing. mSphere 5 (2020): 1–10.
10. Lau SKP, Fan RYY, Zhu L, et al. Isolation of MERS-

Related Coronavirus from Lesser Bamboo Bats That Uses DPP4 and Infects Human-DPP4-Transgenic Mice. Nat Commun 12 (2021): 216.

11. Chen J, Yang X, Si H, et al. A Bat MERS-like Coronavirus Circulates in Pangolins and Utilizes Human DPP4 and Host Proteases for Cell Entry. Cell 186 (2023): 850-863.
12. Fehr AR, Perlman S. Coronaviruses: An Overview of Their Replication and Pathogenesis. In Coronaviruses: Methods and Protocols; Springer New York (2015): 1–23.
13. Raj VS, Mou H, Smits SL, et al. Dipeptidyl Peptidase 4 Is a Functional Receptor for the Emerging Human Coronavirus-EMC. Nat 495 (2013): 251–254.
14. Yang Y, Du L, Liu C, et al. Receptor Usage and Cell Entry of Bat Coronavirus HKU4 Provide Insight into Bat-to-Human Transmission of MERS Coronavirus. Proc Natl Acad Sci 111 (2014): 12516–12521.
15. Millet JK, Whittaker GR. Host Cell Entry of Middle East Respiratory Syndrome Coronavirus after Two-Step, Furin-Mediated Activation of the Spike Protein. Proc Natl Acad Sci 111 (2014): 15214–15219.
16. Kleine-Weber H, Elzayat MT, Hoffmann M, et al. Functional Analysis of Potential Cleavage Sites in the MERS-Coronavirus Spike Protein. Sci Rep 8 (2018): 16597.
17. Yamada Y, Liu DX. Proteolytic Activation of the Spike Protein at a Novel RRRR/S Motif Is Implicated in Furin-Dependent Entry, Syncytium Formation, and Infectivity of Coronavirus Infectious Bronchitis Virus in Cultured Cells. J Virol 83 (2009): 8744–8758.
18. Belouzard S, Chu VC, Whittaker GR. Activation of the SARS Coronavirus Spike Protein via Sequential Proteolytic Cleavage at Two Distinct Sites. Proc Natl Acad Sci 106 (2009): 5871–5876.
19. Wang Q, Qi J, Yuan Y, et al. Bat Origins of MERS-CoV Supported by Bat Coronavirus HKU4 Usage of Human Receptor CD26. Cell Host Microbe 16 (2014): 328–337.
20. Yang Y, Liu C, Du L, et al. Two Mutations Were Critical for Bat-to-Human Transmission of Middle East Respiratory Syndrome Coronavirus. J Virol 89 (2015): 9119–9123.
21. Ballenghien M, Faivre N, Galtier N. Patterns of Cross-Contamination in a Multispecies Population Genomic Project: Detection, Quantification, Impact, and Solutions. BMC Biol 15 (2017): 25.
22. Katz KS, Shutov O, Lapoint R, et al. STAT: A Fast, Scalable, MinHash-Based k-Mer Tool to Assess Sequence Read Archive next-Generation Sequence Submissions. Genome Biol 22 (2021): 270.
23. Lewis G, Jordan JL, Relman DA, et al. The Biosecurity Benefits of Genetic Engineering Attribution. Nat Commun 11 (2020): 6294.
24. Csabai I, Papp k,Visontai D, et al. Unique SARS-CoV-2 Variant Found in Public Sequence Data of Antarctic Soil Samples Collected in 2018-2019 (2021).
25. Quay SC, Zhang D, Jones A, et al. Nipah Virus Vector Sequences in COVID-19 Patient Samples Sequenced by the Wuhan Institute of Virology (2021).
26. Jones A, Massey SE, Zhang D, et al. Forensic Analysis of Novel SARS2r-CoV Identified in Game Animal Datasets in China Shows Evolutionary Relationship to Pangolin GX CoV Clade and Apparent Genetic Experimentation. Appl Microbiol 2 (2022): 882–904.
27. Agarwala R, Barrett T, Beck J, et al. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res 46 (2018): D8–D13.
28. Chen S, He C, Li Y, et al. A Computational Toolset for Rapid Identification of SARS-CoV-2, Other Viruses and Microorganisms from Sequencing Data. Brief Bioinform 22 (2021): 924–935.
29. Chen S, Zhou Y, Chen Y, et al. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. Bioinformatics 34 (2018): i884–i890.
30. Li H. Minimap2: Pairwise Alignment for Nucleotide Sequences. Bioinformatics 34 (2018): 3094–3100.
31. Langmead B, Salzberg SL. Fast Gapped-Read Alignment with Bowtie 2. Nat Methods 9 (2012): 357–359.
32. Li D, Liu C.-M, Luo R, et al MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph. Bioinformatics 31 (2015): 1674–1676.
33. Meleshko D, Hajirasouliha I, Korobeynikov A. CoronaSPAdes: From Biosynthetic Gene Clusters to RNA Viral Assemblies. Bioinformatics 38 (2021): 1–8.
34. Prjibelski A, Antipov D, Meleshko D, et al. Using SPAdes De Novo Assembler. Curr. Protoc. Bioinforma 70 (2020): 1–29.
35. Vasimuddin M, Misra S, Li H, et al. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS); IEEE (2019): 314–324.
36. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map Format and SAMtools. Bioinformatics 25 (2009): 2078–2079.
37. Bamdst (2021). https://github.com/shiquan/bamdst.

38. Islam R, Raju RS, Tasnim N, et al. Choice of Assemblers Has a Critical Impact on de Novo Assembly of SARS-CoV-2 Genome and Characterizing Variants. Brief. Bioinform 22 (2021): 1–11.

39. Johnson M, Zaretskaya I, Raytselis Y, et al. NCBI BLAST: A Better Web Interface. Nucleic Acids Res 36 (2018): W5–W9.

40. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30 (2013), 772–780.

41. Guindon S, Dufayard J.-F, Lefort V, et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 59 (2010): 307–321.

42. Lefort V, Longueville J.-E, Gascuel O. SMS: Smart Model Selection in PhyML. Mol Biol Evol 34 (2017): 2422–2424.

43. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38 (2021): 3022–3027.

44. Martin DP, Murrell B, Golden M, et al B. RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes. Virus Evol 1 (2015): 1–5.

45. Martin D, Rybicki E. RDP: Detection of Recombination amongst Aligned Sequences. Bioinformatics 16 (2000): 562–563.

46. Salminen MO, Carr JK, Burke DS, et al. Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning. AIDS Res. Hum. Retroviruses 11 (1995): 1423–1425.

47. Smith J. Analyzing the Mosaic Structure of Genes. J Mol Evol 34 (1992): 126–129.

48. Posada D, Crandall K. A. Evaluation of Methods for Detecting Recombination from DNA Sequences: Computer Simulations. Proc Natl Acad Sci 98 (2001): 13757–13762.

49. Boni MF, Posada D, Feldman MW. An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets. Genetics 176 (2007): 1035–1047.

50. Padidam M, Sawyer S, Fauquet CM. Possible Emergence of New Geminiviruses by Frequent Recombination. Virol 265 (1999): 218–225.

51. Holmes EC, Worobey M, Rambaut A. Phylogenetic Evidence for Recombination in Dengue Virus. Mol Biol Evol 16 (1999): 405–409.

52. Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-Scanning: A Monte Carlo Procedure for Assessing Signals in Recombinant Sequences. Bioinformatics 16 (2000): 573–582.

53. Samson S, Lord É, Makarenkov V, et al. A Python Application for Representing Sequence Similarity and Detecting Recombination. Bioinformatics 38 (2022): 3118–3120.

54. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. Nucleic Acids Res 46 (2018): W296–W303.

55. PyMOL. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC (2022). https://pymol.org/2/

56. Xue LC, Rodrigues JP, Kastritis PL, et al. PRODIGY: A Web Server for Predicting the Binding Affinity of Protein–Protein Complexes. Bioinformatics 32 (2016): 3676–3678.

57. Davis MW, Jorgensen EM, ApE A. Plasmid Editor: A Freely Available DNA Manipulation and Visualization Program. Front Bioinforma 2 (2022): 1–15.

58. SnapGene® software (from Dotmatics; available at snapgene.com) (2021).

59. Bengtsson-Palme J, Hartmann M, Eriksson KM, et al. METAXA2: Improved Identification and Taxonomic Classification of Small and Large Subunit RRNA in Metagenomic Data. Mol Ecol Resour 15 (2015):

60. Edgar RC, Taylor J, Lin V, et al. Petabase-Scale Sequence Alignment Catalyses Viral Discovery. Nat 602 (2022): 142–147.

61. Addgene. Addgene: Analyze Sequence (2022). https://www.addgene.org/analyze-sequence/

62. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. Brief Bioinform 14 (2013): 178–192.

63. Inayoshi Y, Oguro S, Tanahashi E, et al. Bacterial Artificial Chromosome-Based Reverse Genetics System for Cloning and Manipulation of the Full-Length Genome of Infectious Bronchitis Virus. Curr Res Microb Sci 3 (2022): 100155.

64. Chen L, Yu B, Hua J, et al. Construction of a Full-Length Infectious Bacterial Artificial Chromosome Clone of Duck Enteritis Virus Vaccine Strain. Virol J 10 (2013): 328.

65. Enjuanes L, DeDiego ML, Álvarez E, et al. Attenuated SARS and use as a vaccine (2006).

66. Almazán F, DeDiego ML, Galán C, et al. Construction of a Severe Acute Respiratory Syndrome Coronavirus Infectious CDNA Clone and a Replicon To Study

Coronavirus RNA Synthesis. J Virol 80 (2006): 10900–10906.

67. DeDiego ML, Álvarez E, Almazán F, et al. A Severe Acute Respiratory Syndrome Coronavirus That Lacks the E Gene Is Attenuated In Vitro and In Vivo. J Virol 81 (2007): 1701–1713.

68. Almazán F, Galán C, Enjuanes, L. Engineering Infectious CDNAs of Coronavirus as Bacterial Artificial Chromosomes. In Coronaviruses: Methods and Protocols 454 (2008): 275–291.

69. Wang W, Peng X, Jin Y, et al. Reverse Genetics Systems for SARS-CoV-2. J Med Virol 94 (2022): 3017–3031.

70. Latinne A, Hu B, Olival KJ, et al. Origin and Cross-Species Transmission of Bat Coronaviruses in China. Nat Commun 11 (2020): 4235.

71. Luo C-M, Wang N, Yang X-L, et al. Discovery of Novel Bat Coronaviruses in South China That Use the Same Receptor as Middle East Respiratory Syndrome Coronavirus. J Virol 92 (2018): 1–15.

72. Okonechnikov K, Golosova O, Fursov M, et al. Unipro UGENE: A Unified Bioinformatics Toolkit. Bioinformatics 28 (2012): 1166–1167.

73. Hu B, Zeng L-P, Yang X-L, et al. Discovery of a Rich Gene Pool of Bat SARS-Related Coronaviruses Provides New Insights into the Origin of SARS Coronavirus. PLOS Pathog 13 (2017): e1006698.

74. Xie X, Muruato A, Lokugamage KG, et al. An Infectious CDNA Clone of SARS-CoV-2. Cell Host Microbe 27 (2020): 841-848.

75. Huang Y, Huang C-X, Wang W-F, et al. Zebrafish MiR-462-731 Is Required for Digestive Organ Development. Comp. Biochem. Physiol. Part D Genomics Proteomics 34 (2020): 100679.

76. Fan Y, Zhao K, Shi Z-L, et al. Bat Coronaviruses in China. Viruses 11 (2019): 210.

77. Daszak P. Understanding the Risk of Bat Coronavirus Emergence. NIH Grant: 5R01AI110964-05 (2021).

78. Yount B, Denison MR, Weiss SR, et al. Systematic Assembly of a Full-Length Infectious CDNA of Mouse Hepatitis Virus Strain A59. J Virol 76 (2002): 11065–11078.

79. Engler C, Kandzia R, Marillonnet S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. PLoS One 3 (2008): e3647.

80. Engler C, Marillonnet S. Golden Gate Cloning. In Methods in Molecular Biology; Humana Press Inc 1116 (2014): 119–131.

81. Hou YJ, Okuda K, Edwards CE, et al. SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. Cell 182 (2020): 429-446.

82. Bruttel V, Washburne A, VanDongen A. Endonuclease Fingerprint Indicates a Synthetic Origin of SARS-CoV-2. bioRxiv (2022).

83. Zeng L-P, Gao Y-T, Ge X-Y, et al. Bat Severe Acute Respiratory Syndrome-Like Coronavirus WIV1 Encodes an Extra Accessory Protein, ORFX, Involved in Modulation of the Host Immune Response. J Virol 90 (2016): 6573–6582.

84. Sun Y, Zhang H, Shi J, et al. Identification of a Novel Inhibitor against Middle East Respiratory Syndrome Coronavirus. Viruses 9 (2017): 255.

85. Xia S, Lan Q, Pu J, et al. Potent MERS-CoV Fusion Inhibitory Peptides Identified from HR2 Domain in Spike Protein of Bat Coronavirus HKU4. Viruses 11 (2019): 56.

86. Zhang W, Zheng X-S, Agwanda B, et al. Serological Evidence of MERS-CoV and HKU8-Related CoV Co-Infection in Kenyan Camels. Emerg. Microbes Infect 8 (2019): 1528–1534.

87. Wang J-M, Wang L-F, Shi Z-L. Construction of a Non-Infectious SARS Coronavirus Replicon for Application in Drug Screening and Analysis of Viral Protein Function. Biochem Biophys Res Commun 374 (2008): 138–142.

88. Chen Z, Bao L, Chen C, et al. Human Neutralizing Monoclonal Antibody Inhibition of Middle East Respiratory Syndrome Coronavirus Replication in the Common Marmoset. J Infect Dis 215 (2017): 1807–1815.

89. Yuan Y, Cao D, Zhang Y, et al. Cryo-EM Structures of MERS-CoV and SARS-CoV Spike Glycoproteins Reveal the Dynamic Receptor Binding Domains. Nat Commun 8 (2017).

90. Wu A, Wang Y, Zeng C, et al. Prediction and Biochemical Analysis of Putative Cleavage Sites of the 3C-like Protease of Middle East Respiratory Syndrome Coronavirus. Virus Res 208 (2015): 56–65.

91. Han X, Qi J, Song H, et al. Structure of the S1 Subunit C-Terminal Domain from Bat-Derived Coronavirus HKU5 Spike Protein. Virol 507 (2017): 101–109.

92. Cohen, J. Wuhan Coronavirus Hunter Shi Zhengli Speaks Out. Sci 369 (2020): 487–488.

93. Lusk RW. Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data. PLoS One 9 (2014): e110808.

94. Selitsky SR, Marron D, Hollern D, et al. Virus Expression Detection Reveals RNA-Sequencing Contamination in TCGA. BMC Genomics 21 (2020): 79.

95. Cantalupo PG, Pipas JM. Detecting Viral Sequences in NGS Data. Curr. Opin. Virol 39 (2019): 41–48.

96. Yu P, Hu B, Shi Z-L, et al. Geographical Structure of Bat SARS-Related Coronaviruses. Infect. Genet Evol 69 (2019): 224–229.

97. Epstein JH, Anthony SJ, Islam A, et al. Nipah Virus Dynamics in Bats and Implications for Spillover to Humans. Proc Natl Acad Sci 117 (2020): 29190–29201.

98. Lu G, Wang Q, Gao GF. Bat-to-Human: Spike Features Determining 'Host Jump' of Coronaviruses SARS-CoV, MERS-CoV, and Beyond. Trends Microbiol 23 (2015): 468–478.

99. Butler D. Engineered Bat Virus Stirs Debate over Risky Research. Nat (2015): 1–2.

100. Klotz L. Human error in high-biocontainment labs: a likely pandemic threat. Bulletin of the Atomic Scientists (2019).

101. Shinomiya N, Minari J, Yoshizawa G, et al. Reconsidering the Need for Gain-of-Function Research on Enhanced Potential Pandemic Pathogens in the Post-COVID-19 Era. Front Bioeng Biotechnol 10 (2022): 1–14.