AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

OBSERVATION
Host-Microbe Biology

An Extensive Meta-Metagenomic Search Identifies SARS-CoV-2-Homologous Sequences in Pangolin Lung Viromes

Lamia Wahba,[a] Nimit Jain,[a,b,c] Andrew Z. Fire,[a,b] Massa J. Shoura,[a] Karen L. Artiles,[a] Matthew J. McCoy,[a] Dae-Eun Jeong[a]

[a]Department of Pathology, Stanford University School of Medicine, Stanford, California, USA
[b]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA
[c]Department of Bioengineering, Stanford University, Stanford, California, USA

All authors contributed equally to this work. Author order was chosen randomly.

ABSTRACT In numerous instances, tracking the biological significance of a nucleic acid sequence can be augmented through the identification of environmental niches in which the sequence of interest is present. Many metagenomic data sets are now available, with deep sequencing of samples from diverse biological niches. While any individual metagenomic data set can be readily queried using web-based tools, meta-searches through all such data sets are less accessible. In this brief communication, we demonstrate such a meta-metagenomic approach, examining close matches to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in all high-throughput sequencing data sets in the NCBI Sequence Read Archive accessible with the "virome" keyword. In addition to the homology to bat coronaviruses observed in descriptions of the SARS-CoV-2 sequence (F. Wu, S. Zhao, B. Yu, Y. M. Chen, et al., Nature 579:265–269, 2020, https://doi.org/10.1038/s41586-020-2008-3; P. Zhou, X. L. Yang, X. G. Wang, B. Hu, et al., Nature 579:270–273, 2020, https://doi.org/10.1038/s41586-020-2012-7), we note a strong homology to numerous sequence reads in metavirome data sets generated from the lungs of deceased pangolins reported by Liu et al. (P. Liu, W. Chen, and J. P. Chen, Viruses 11:979, 2019, https://doi.org/10.3390/v11110979). While analysis of these reads indicates the presence of a similar viral sequence in pangolin lung, the similarity is not sufficient to either confirm or rule out a role for pangolins as an intermediate host in the recent emergence of SARS-CoV-2. In addition to the implications for SARS-CoV-2 emergence, this study illustrates the utility and limitations of meta-metagenomic search tools in effective and rapid characterization of potentially significant nucleic acid sequences.

IMPORTANCE Meta-metagenomic searches allow for high-speed, low-cost identification of potentially significant biological niches for sequences of interest.

KEYWORDS COVID, SARS-nCoV-2, bioinformatics, coronavirus, metagenomics, pangolin

In the early years of nucleic acid sequencing, aggregation of the majority of published DNA and RNA sequences into public sequence databases greatly aided biological hypothesis generation and discovery. Search tools capable of interrogating the ever-expanding databases were facilitated by creative algorithm development and software engineering and by the ever-increasing capabilities of computer hardware and the Internet. In the early 2000s, sequencing methodologies and computational technologies advanced in tandem, enabling quick homology results from a novel sequence without substantial cost.

With the development of larger-scale sequencing methodologies, the time and resources to search all extant sequence data became untenable for most studies.

Citation Wahba L, Jain N, Fire AZ, Shoura MJ, Artiles KL, McCoy MJ, Jeong D-E. 2020. An extensive meta-metagenomic search identifies SARS-CoV-2-homologous sequences in pangolin lung viromes. mSphere 5:e00160-20. https://doi.org/10.1128/mSphere.00160-20.

Editor Michael J. Imperiale, University of Michigan—Ann Arbor

Address correspondence to Andrew Z. Fire, afire@stanford.edu.

An extensive meta-metagenomic search identifies SARS-CoV-2-homologous sequences in pangolin lung viromes. Fire lab, Stanford @Fire__Lab

Received 18 February 2020
Accepted 24 April 2020
Published 6 May 2020

However, creative approaches involving curated databases and feature searches ensured that many key features of novel sequences remained readily accessible. At the same time, the nascent field of metagenomics began, with numerous studies highlighting the power of survey sequencing of DNA and RNA from samples as diverse as the human gut and Antarctic soil (1, 2). As the diversity and size of such data sets expand, the utility of searching them with a novel sequence increases. Meta-metagenomic searches are currently underutilized. In principle, such searches would involve direct access to sequence data from a large set of metagenomic experiments on a terabyte scale, along with software able to search for similarity to a query sequence. We find that neither of these aspects of meta-metagenomic searches is infeasible with current data transfer and processing speeds. In this communication, we report the results of searching the recently described severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequence through a set of metagenomic data sets with the "virome" tag.

Experimental procedures. (i) Computing hardware. A Linux workstation used for the bulk analysis of metagenomic data sets employs an 8-core i7 Intel microprocessor, 128 gigabyte (GB) of random access memory, 12 terabytes (TB) of conventional disk storage, and 1 TB of solid state drive (SSD) storage. Additional analyses of individual alignments were conducted with standard consumer-grade computers.

(ii) Sequence data. All sequence data for this analysis were downloaded from the National Center for Biotechnology Information (NCBI) website, with individual sequences downloaded through a web interface and metagenomic data sets downloaded from the NCBI Sequence Read Archive (SRA) using the SRA-tools package (version 2.9.1). The latter sequence data were downloaded as .sra files using the prefetch tool, with extraction to readable format (.fasta.gz) using the NCBI fastq-dump tool. Each of these manipulations can fail some fraction of the time. Obtaining the sequences can fail due to network issues, while extraction in readable format occasionally fails for unknown reasons. Thus, we developed a set of scripts implementing a workflow that continually requests .sra files with ncbi-prefetch until at least some type of file is obtained, followed by attempts to unpack into .fasta.gz format until one such file is obtained from each .sra file. Metagenomic data sets for analysis were chosen through a keyword search of the SRA descriptions for "virome" and downloaded between 27 January 2020 and 31 January 2020. We note that the "virome" keyword search will certainly not capture every metagenomic data set with viral sequences, and likewise not capture every virus in the short sequence read archive. Despite these clear limitations, the "virome" keyword search identified a broad and diverse set of experimental data sets for further analysis. With up to 16 threads running simultaneously, total download time (prefetch) was approximately 2 days. Similar time was required for conversion to gzipped fasta files. A total of 9,014 sequence data sets were downloaded and converted to fasta.gz files. Most files (as expected) contained large numbers of reads, while a small fraction contained very little data (only a few reads or reads of at most a few base pairs). The total data set consists of 2.5 TB of compressed sequence data corresponding to approximately 10^{13} bases.

(iii) Search software. For rapid identification of close matches among large numbers of metagenomic reads, we used a simple dictionary based on the SARS-CoV-2 sequence (our query sequence for SARS-CoV-2 was GenBank accession no. MN908947.3) and its reverse complement, querying every 8th k-mer along the individual reads for matches to the sequence. As a reference, and to benchmark the workflow further, we included several additional sequences in the query (vaccinia virus, an arbitrary segment of an influenza virus isolate, the full sequence of bacteriophage P4, and a number of putative polinton sequences from *Caenorhabditis briggsae*). The relatively small group of k-mers being queried ($<10^6$) allows a rapid search for homologs. This was implemented in a Python script run using the PyPy accelerated interpreter. We stress that this is by no means the most comprehensive or fastest search for large data sets. However, it is more than sufficient to rapidly find any closely

matching sequence (with the downloading and conversion of the data, rather than the search, being rate limiting). While the high degree of conservation between isolates of individual coronaviruses and between related coronaviruses (3, 4) facilitates this very simple approach to pattern matching for discovery of closely related viruses in this family, alternative approaches with more-sensitive matching algorithms for other gene or virus families would require only minimally expanded computing resources (5).

(iv) Alignment of reads to SARS-CoV-2. Reads from the pangolin hit data sets were adapter rimmed with cutadapt (version 1.18) (6), and mapped to the SARS-CoV-2 genome with BWA-MEM (version 0.7.12) (7) using default settings for paired-end mode. Alignments were visualized with the Integrated Genomics Viewer (IGV) tool (version 2.4.10) (8).

(v) Assessment of nucleotide similarity between SARS-CoV-2, pangolin metavirome reads, and closely related bat coronaviruses. All pangolin metavirome reads that aligned to the SARS-CoV-2 genome with BWA-MEM after adapter trimming with cutadapt were used for calculation. The bat coronavirus genomes were aligned to the SARS-CoV-2 genome in a multiple sequence alignment using the web interface for Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) (9) with default settings. We note that sequence insertions with respect to the SARS-CoV-2 genome in either the pangolin metavirome reads or the bat coronavirus genomes are not accounted for in the similarity traces shown in Fig. 1b.

(vi) Regional assessment of synonymous and nonsynonymous mutations. Although the incomplete nature of coverage in the pangolin metavirome data somewhat limits the application of measures such as normalized *dN*/*dS* (ratio of nonsynonymous to synonymous evolutionary changes) values, it remains possible to identify regions with the strongest matches of this inferred viral sequence with the human and bat homologs and to determine the distribution of synonymous and nonsynonymous variants in these regions. Details of this analysis are presented in Fig. S3 in the supplemental material.

(vii) Accessibility of software. Scripts used for the observations described in this communication are available at https://github.com/firelabsoftware/Metasearch2020.

Findings. To identify biological niches that might harbor viruses closely related to SARS-CoV-2, we searched through publicly available metavirome data sets. We were most interested in viruses with highly similar sequences, as these would likely be most useful in forming hypotheses about the origin and pathology of the recent human virus. We thus set a threshold requiring matching of a perfect 32-nucleotide segment with a granularity of 8 nucleotides in the search (i.e., interrogating the complete database of k-mers from the virus with k-mers starting at nucleotide 1, 9, 17, 25, 33 of each read from the metagenomic data for a perfect match). This would catch any perfect match of 39 nucleotides or greater (regardless of phasing relative to the 8-base search granularity), with some homologies as short as 32 nucleotides captured depending on the precise phasing of the read.

All metagenomic data sets with the keyword "virome" in NCBI SRA as of January 2020 were selected for analysis in a process that required approximately 2 days each for downloading and conversion to readable file formats and 1 day for searching by k-mer match on a desktop workstation computer (i7 8-core). Together the data sets included information from 9,014 NCBI Short Read Archive entries with (in total) 6.2×10^{10} individual reads and $8.4 \times$ 1,012 bp. Despite the relatively large mass of data, the 32-nucleotide k-mer match remains a stringent measure, with spurious matches to the ~30-kb SARS-CoV-2 genome expected at only 1 in 3×10^{14}. Positive matches among the metagenomic data sets analyzed were relatively rare, with the vast majority of data sets (8,994/9,014 or 99.8%) showing no matched 32-mers to SARS-CoV-2. Of the data sets with matched k-mers, one was from a synthetic mixture of viral sequences that included a feline alphacoronavirus (10), while the remaining were all from vertebrate animal sources. The latter matches were from five studies: two bat-focused studies (11,

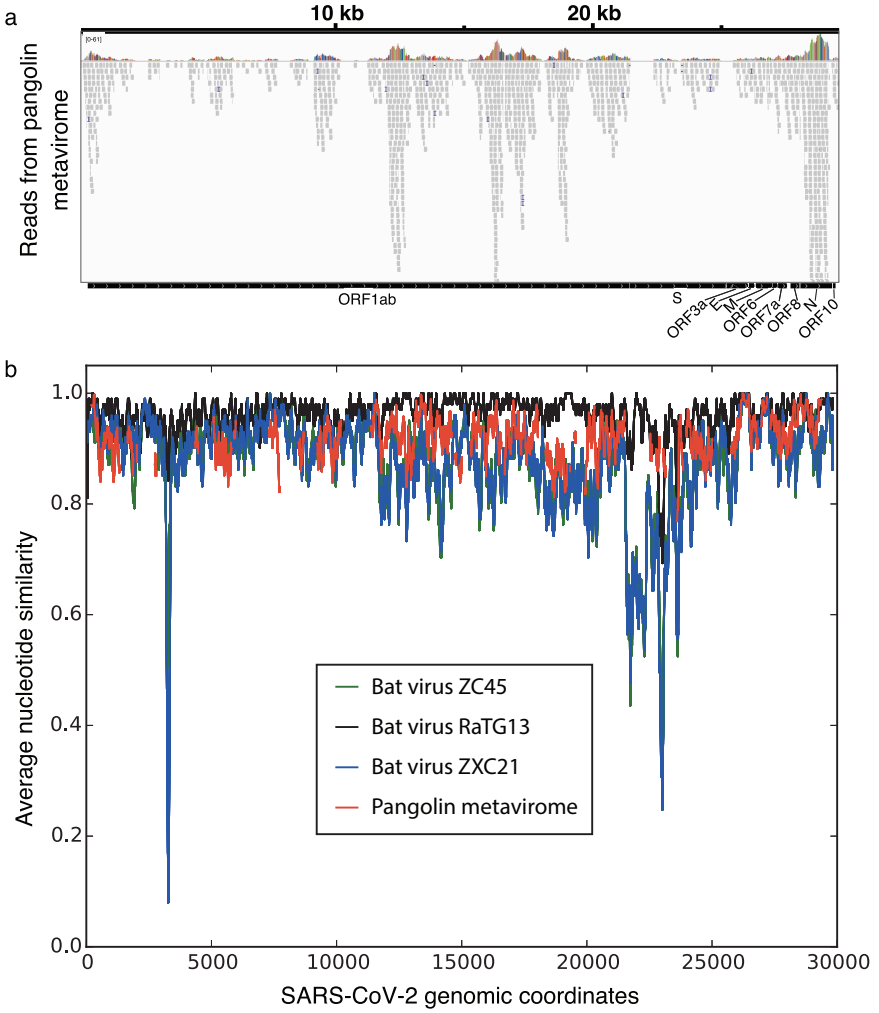

FIG 1 (a) Integrated Genomics Viewer (IGV) snapshot of alignment. Reads from the pangolin lung virome samples (SRA accession no. SRR10168377, SRR10168378, and SRR10168376) were mapped to a SARS-CoV-2 reference sequence (GenBank accession no. MN908947.3). The total numbers of aligned reads from the three samples were 1,107, 313, and 32 reads, respectively. Figure S1 in the supplemental material shows an enlarged view for these alignments within the spike RBD region. (b) Quantification of nucleotide-level similarity between the SARS-CoV-2 genome and pangolin lung metavirome reads aligning to the SARS-CoV-2 genome. Average similarity was calculated in 101-nucleotide windows along the SARS-CoV-2 genome and is only shown for those windows where each nucleotide in the window had coverage of ≥2. Average nucleotide similarity calculated (in 101-nucleotide windows) between the SARS-CoV-2 genome and reference genomes of three relevant bat coronaviruses (bat-SL-CoVZC45 [accession no. MG772933.1], bat-SL-CoVZXC21, [accession no. MG772934.1], and RaTG13 [accession no. MN996532.1]) is also shown. Note that the pangolin metavirome similarity trace is not directly comparable to the bat coronavirus similarity traces, because the former uses read data for calculation, whereas the latter uses reference genomes.

12), one bird-focused study (13), one study focused on small animals and rodents (14), and a study of pangolins (15) (Table 1).

The abundance and homology of viruses within a metagenomic sample are of considerable interest in interpreting possible characteristics of infection and relevance to the query virus. From the quick k-mer search, an initial indicator could be inferred from the number of matching reads and k-mer match counts for those reads (Table 1; see also Table S1 in the supplemental material). For the SARS-CoV-2 matches among the available metagenomic data sets, the strongest and most abundant matches in these analyses came from the pangolin lung metaviromes. The matches were observed throughout the SARS-CoV-2 query sequence, and many of the matching reads showed numerous matching 32-mer sequences. The vast majority of matches were in two lung

TABLE 1 Metagenomic data sets with k = 32-mer matches to GenBank accession no. MN908947.3 (SARS-CoV-2)[a]

Description	File	TotalReads	TotalBases	HitReads	HitKmers
Pangolin Lung	SRR10168376_1	18067615	2710142250	5	9
Pangolin Lung	SRR10168376_2	18067615	2710142250	6	21
Pangolin Lung	SRR10168377_1	16414925	2462238750	308	955
Pangolin Lung	SRR10168377_2	16414925	2462238750	285	904
Pangolin Lung	SRR10168378_1	19045923	2856888450	91	352
Pangolin Lung	SRR10168378_2	19045923	2856888450	96	337
Pangolin Lung	SRR10168392_1	39738679	5960801850	2	4
Pangolin Lung	SRR10168392_2	39738679	5960801850	4	6
"Mock virome"	SRR3458564_2	10957589	1654595939	1	1
Bat Feces	SRR5040897_1	4020145	607041895	5	5
Bat Feces	SRR5040897_2	4020145	607041895	4	4
Bat Feces	SRR5040918_1	4804340	725455340	4	4
Bat Feces	SRR5040918_2	4804340	725455340	3	3
Virome analysis of Rodents and other small animals	SRR5343975_1	19775296	1582023680	1	1
Virome analysis of Rodents and other small animals	SRR5343977_1	20227246	1618179680	1	2
Bat	SRR5351751_2	363186	32644600	1	1
Bat	SRR5351752_1	305206	27328640	24	57
Bat	SRR5351752_2	305206	27608440	20	55
Bat	SRR5351758_2	301889	27514580	1	3
Bat	SRR5351760_1	788212	70634120	407	962
Bat	SRR5351760_2	788212	71244040	378	1006
Virome analysis of Rodents and other small animals	SRR5365807_1	21646498	1731719840	2	5
Virome analysis of Rodents and other small animals	SRR5365809_1	22799703	1823976240	1	1
Virome analysis of Rodents and other small animals	SRR5431767_1	10213803	1031594103	2	2
Virome analysis of Rodents and other small animals	SRR5447167_1	13849913	1398841213	2	2
Virome analysis of Rodents and other small animals	SRR5447174_1	14063346	1420397946	1	1
Virome analysis of Rodents and other small animals	SRR5447175_1	7510256	758535856	1	1
Avocet (RNA Virome in Wild Birds)	SRR7239364_1	22336985	2233698500	2	2

[a]Details of the search are described in the legend to Table S1 in the supplemental material.

samples—lung07 and lung 08—with small numbers of matches in two additional lung data sets, lung02 and lung09 (15). No matches were detected for seven additional lung data sets, and no matches were seen in eight spleen samples and a lymph node sample (15). Further analysis of coverage and homology through alignment of the metagenomic data sets revealed an extensive, if incomplete, coverage of the SARS-CoV-2 genome (Fig. 1a and Fig. S1A to C). Percent nucleotide similarity can be calculated for pangolin metavirome reads aligning to SARS-CoV-2 (Fig. 1b), and these segmental homologies consistently showed strong matches, approaching (but still overall weaker than) the similarity of the closest known bat coronavirus (RaTG13). A provisional comparison of synonymous differences at the nucleotide level between the pangolin reads, bat coronavirus RaTG13, and SARS-CoV-2 was also feasible for genes where pangolin sequences were available and readily aligned. Many synonymous (generally codon third base) changes were visible in such comparisons (Fig. S2 and S3). Comparisons of RaTG13 to SARS-CoV-2 revealed synonymous changes at 10% of conserved amino acid residues, while comparisons of the aggregate (but incomplete) pangolin reads indicated synonymous changes at 23% of conserved amino acid residues. Within the receptor binding domain (RBD) of the spike protein (16), these values are 26% and 34%, respectively.

The potential structural implications of protein sequence divergence in the RBD region of the spike protein were explored through combined sequence-directed structural alignment. The bat coronavirus RaTG13 RBD is markedly divergent relative to the SARS-CoV-2 RBD, with several amino acid differences located at the ACE-2 receptor-RBD interface (Fig. S4, with coordinate information shown in Text S1 in the supplemental material) (17). Thus, changes in these amino acid sequences, as previously described (17) and shown for comparison in Fig. S4 could be expected to influence interactions with the human ACE-2 receptor. In the case of the pangolin sequences, amino acid changes relative to the SARS-CoV-2 RBD seem to be, for the most part, located outside of the ACE-2 interface, with the exception of two residues (417 and 498) at the interface (17, 18). Overall, the SARS-CoV-2 and inferred pangolin virus amino acid sequences differ at seven positions in the RBD (residues 346, 372, 402, 417, 498, 519, and 529) (Fig. S4).

Conclusions. Meta-metagenomic searching can provide unique opportunities to understand the distribution of nucleic acid sequences in diverse environmental niches. As metagenomic data sets proliferate and as both the need and capability to identify pathogenic agents through sequencing increase, meta-metagenomic searching may prove extremely useful in tracing the origins and spreading of causative agents. In the example we present in this paper, such a search identifies a number of niches with sequences matching the genome of the SARS-CoV-2 virus. These analyses raise a number of relevant points for the origin of SARS-CoV-2. Before describing the details of these points, however, it is important to stress that while environmental, clinical, and animal-based sequencing is valuable in understanding how viruses traverse the animal ecosphere, static sequence distributions cannot be used to construct the full transmission history of a virus among different biological niches. So even if the closest relative of a virus-causing disease in species X were to be found in species Y, we cannot define the source of the outbreak or the direction(s) of transmission. As some viruses may move more than once between hosts, the sequence of a genome at any time may reflect a history of selection and drift in several different host species. This point is also accentuated in the microcosm of our searches for this work. When we originally obtained the SARS-CoV-2 sequence from the posted work of Wu et al. (3), we recapitulated their result that bat-SL-CoVZC45 was the closest related sequence in NCBI's nonredundant (nr/nt) database. In our screen of metavirome data sets, we observed several pangolin metavirome sequences, which were not in the NCBI nr/nt database at the time, that are more closely related to SARS-CoV-2 than bat-SL-CoVZC45. An assumption that the closest relative of a sequence identifies the origin would at that point have transferred the extant model to zoonosis from pangolin instead of bat. To complicate such a model, an additional study from Zhou et al. (4) described a previously unpublished coronavirus sequence, designated RaTG13 with much stronger homology to SARS-CoV-2 than either bat-SL-CoVZC45 or the pangolin reads from Liu et al. (15). While this observation certainly shifts the discussion (legitimately) toward a possible bat-borne intermediate in the chain leading to SARS-CoV-2, it remains difficult to determine if any of these are true intermediates in the chain of infectivity.

The match of SARS-CoV-2 to the pangolin coronavirus sequences also enables a link to substantial context on the pangolin samples from Liu et al. (15), with information on the source of the rescued animals (from smuggling activity), the nature of their deaths despite rescue efforts, the potential presence of other viruses in the same whole-lung tissue, and the accompanying gross pathology. The pangolins appear to have died from lung-related illness, which may have involved a SARS-CoV-2-homologous virus. Notably, however, two of the deceased pangolin lungs had much lower SARS-CoV-2 signals, while seven showed no signal, with sequencing depths in the various lungs roughly comparable. Although it remains possible that the SARS-CoV-2-like coronavirus was the primary cause of death for these animals, it is also possible (as noted by Liu et al. [15]) that the virus was simply present in the tissue, with mortality due to another virus, a combination of infectious agents, or other exposures.

During the course of this work, the homology between SARS-CoV-2 and pangolin coronavirus sequences in a particular genomic subregion was also noted and discussed in an online forum ("Virological.org") with some extremely valuable analyses and insights. Matthew Wong and colleagues bring up the homology to the pangolin metagenomic data sets in this thread and appear to have encountered it through a more targeted search than ours (this study has since been posted online on bioRxiv [19]). As noted by Wong et al. (19), the spike region includes a segment of ~200 nucleotides encompassing the RBD where the inferred divergence between RaTG13 and SARS-CoV-2 dramatically increases. This region is of interest, as it is a key determinant of viral host range and under heavy selection (20). The observed spike region divergence indeed includes a substantial set of nonsynonymous differences (Fig. S2 and S3). Notably, while reads from the pangolin lung data sets mapped to this region do not show a similar increase in variation relative to SARS-CoV-2, we also did not

observe a significant drop in variation between SARS-CoV-2 and pangolin sequences in this region (Fig. S2 and S3). Instead, variation in the region is comparable to numerous other conserved regions of the spike and to the viral genome as a whole. While Wong et al. (19) and others (21–28) raised the model that recombination occurred in the RBD region in the derivation of SARS-CoV-2, the lack of a singular dip in the landscape of pangolin-SARS-CoV-2 variation in the region would seem counterintuitive were SARS-CoV-2 a result of a localized recombination between a close relative of RaTG13 and a close relative of the putative pangolin coronaviruses under consideration. Thus alternative models for the observed sequence variation seem important to consider and indeed parsimonious, including that of selection acting on the RaTG13 sequences in bats or another intermediate host resulting in a rapid variation of the amino acids at the highly critical virus-receptor interface. Overall, definitive conclusions regarding the origins of SARS-CoV-2 or other coronaviruses will remain difficult with limited sequencing data and without knowledge of evolutionary trajectories in different lineages (29, 30).

A number of literature contributions now discuss the potential role for bats, pangolins, and other possible progenitor/intermediate species in derivation of SARS-CoV-2 from different approaches and perspectives, with a diversity of approaches and interpretations in understanding the origin of the virus. In particular, there has been extensive discussion and debate about the possible pangolin origin of SARS-CoV-2 (19, 21–28, 31–41). These studies provide useful insights into the evolution of SARS-CoV-2 but have limitations and uncertainty in drawing conclusions regarding the viral origin, as most studies were mainly performed through sequence-based comparison and simulation. Thus, better understanding of the current pandemic requires additional information on investigational, experimental, and epidemiological levels that may resolve questions of origin and of preventing the reemergence of SARS-CoV-2 and other pathogens. Nevertheless, the availability of numerous paths (both targeted and agnostic) toward identification of natural niches for pathogenic sequences, including our meta-metagenomic search, will remain useful to the scientific community and to public health, as will vigorous sharing of ideas, data, and discussion of potential origins and modes of spread for epidemic pathogens.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, TXT file, 2.2 MB.

FIG S1, PDF file, 0.2 MB.

FIG S2, PDF file, 0.1 MB.

FIG S3, PDF file, 2.1 MB.

FIG S4, PDF file, 0.4 MB.

TABLE S1, TXT file, 1.3 MB.

ACKNOWLEDGMENTS

This study was supported by the following programs, grants, and fellowships: Human Frontier Science Program (HFSP) to D.-E.J., Arnold O. Beckman Award to M.J.S., Stanford Genomics Training Program (5T32HG000044-22; principal investigator [PI], M. Snyder) to M.J.M., and R35GM130366 to A.Z.F.

We declare that we have no competing interests.

REFERENCES

1. Bry L, Falk PG, Midtvedt T, Gordon JI. 1996. A model of host-microbial interactions in an open mammalian ecosystem. Science 273:1380–1383. https://doi.org/10.1126/science.273.5280.1380.
2. Bowman JS. 2018. Identification of microbial dark matter in Antarctic environments. Front Microbiol 9:3165. https://doi.org/10.3389/fmicb.2018.03165.
3. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265–269. https://doi.org/10.1038/s41586-020-2202-3.
4. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7.
5. Edgar RC. 2020. URMAP, an ultra-fast read mapper. bioRxiv https://doi.org/10.1101/2020.01.12.903351.
6. Marcel M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17:10–12.
7. Li H, Durbin R. 2010. Fast and accurate long-read alignment with

Burrows-Wheeler transform. Bioinformatics 26:589–595. https://doi.org/10.1093/bioinformatics/btp698.
8. Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. https://doi.org/10.1093/bib/bbs017.
9. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636–W641. https://doi.org/10.1093/nar/gkz268.
10. Conceicao-Neto N, Zeller M, Lefrere H, De Bruyn P, Beller L, Deboutte W, Yinda CK, Lavigne R, Maes P, Van Ranst M, Heylen E, Matthijnssens J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci Rep 5:16532. https://doi.org/10.1038/srep16532.
11. Hu D, Zhu C, Wang Y, Ai L, Yang L, Ye F, Ding C, Chen J, He B, Zhu J, Qian H, Xu W, Feng Y, Tan W, Wang C. 2017. Virome analysis for identification of novel mammalian viruses in bats from Southeast China. Sci Rep 7:10917. https://doi.org/10.1038/s41598-017-11384-w.
12. Yinda CK, Zeller M, Conceicao-Neto N, Maes P, Deboutte W, Beller L, Heylen E, Ghogomu SM, Van Ranst M, Matthijnssens J. 2016. Novel highly divergent reassortant bat rotaviruses in Cameroon, without evidence of zoonosis. Sci Rep 6:34209. https://doi.org/10.1038/srep34209.
13. Wille M, Eden JS, Shi M, Klaassen M, Hurt AC, Holmes EC. 2018. Virus-virus interactions and host ecology are associated with RNA virome structure in wild birds. Mol Ecol 27:5263–5278. https://doi.org/10.1111/mec.14918.
14. Wu Z, Lu L, Du J, Yang L, Ren X, Liu B, Jiang J, Yang J, Dong J, Sun L, Zhu Y, Li Y, Zheng D, Zhang C, Su H, Zheng Y, Zhou H, Zhu G, Li H, Chmura A, Yang F, Daszak P, Wang J, Liu Q, Jin Q. 2018. Comparative analysis of rodent and small mammal viromes to better understand the wildlife origin of emerging infectious diseases. Microbiome 6:178. https://doi.org/10.1186/s40168-018-0554-9.
15. Liu P, Chen W, Chen JP. 2019. Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (Manis javanica). Viruses 11:979. https://doi.org/10.3390/v11110979.
16. Li F, Li W, Farzan M, Harrison SC. 2005. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 309:1864–1868. https://doi.org/10.1126/science.1116480.
17. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, Wang X. 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature https://doi.org/10.1038/s41586-020-2180-5.
18. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367:1444–1448. https://doi.org/10.1126/science.abb2762.
19. Wong MC, Javornik Cregeen SJ, Ajami NJ, Petrosino JF. 2020. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. bioRxiv https://doi.org/10.1101/2020.02.07.939207.
20. Li F. 2016. Structure, function, and evolution of coronavirus spike proteins. Annu Rev Virol 3:237–261. https://doi.org/10.1146/annurev-virology-110615-042301.
21. Lam TT, Shum MH, Zhu HC, Tong YG, Ni XB, Liao YS, Wei W, Cheung WY, Li WJ, Li LF, Leung GM, Holmes EC, Hu YL, Guan Y. 2020. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature https://doi.org/10.1038/s41586-020-2169-0.
22. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, Li N, Guo Y, Li X, Shen X, Zhang Z, Shu F, Huang W, Li Y, Zhang Z, Chen R-A, Wu Y-J, Peng S-M, Huang M, Xie W-J, Cai Q-H, Hou F-H, Liu Y, Chen W, Xiao L, Shen Y. 2020. Isolation and characterization of 2019-nCoV-like coronavirus from Malayan pangolins. bioRxiv https://doi.org/10.1101/2020.02.17.951335.
23. Zhang T, Wu Q, Zhang Z. 2020. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol 30:1346–1351.e2. https://doi.org/10.1016/j.cub.2020.03.022.
24. Wu A, Niu P, Wang L, Zhou H, Zhao X, Wang W, Wang J, Ji C, Ding X, Wang X, Lu R, Gold S, Aliyari S, Zhang S, Vikram E, Zou A, Lenh E, Chen J, Ye F, Han N, Peng Y, Guo H, Wu G, Jiang T, Tan W, Cheng G. 2020. Mutations, recombination and insertion in the evolution of 2019-nCoV. bioRxiv https://doi.org/10.1101/2020.02.29.971101.
25. Huang J-M, Jan SS, Wei X, Wan Y, Ouyang S. 2020. Evidence of the recombinant origin and ongoing mutations in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). bioRxiv https://doi.org/10.1101/2020.03.16.993816.
26. Li X, Giorgi EE, Marichann MH, Foley B, Xiao C, Kong X-P, Chen Y, Korber B, Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. bioRxiv https://doi.org/10.1101/2020.03.20.000885.
27. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry B, Castoe T, Rambaut A, Robertson DL. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. bioRxiv https://doi.org/10.1101/2020.03.30.015008.
28. Tagliamonte MS, Abid N, Chillemi G, Salemi M, Mavian C. 2020. Re-insights into origin and adaptation of SARS-CoV-2. bioRxiv https://doi.org/10.1101/2020.03.30.015685.
29. Magiorkinis G, Magiorkinis E, Paraskevis D, Vandamme AM, Van Ranst M, Moulton V, Hatzakis A. 2004. Phylogenetic analysis of the full-length SARS-CoV sequences: evidence for phylogenetic discordance in three genomic regions. J Med Virol 74:369–372. https://doi.org/10.1002/jmv.20187.
30. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect Genet Evol 79:104212. https://doi.org/10.1016/j.meegid.2020.104212.
31. Liu P, Jiang J-Z, Wan X-F, Hua Y, Wang X, Hou F, Chen J, Zou J, Chen J. 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (2019-nCoV)? bioRxiv https://doi.org/10.1101/2020.02.18.954628.
32. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. 2020. Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. J Proteome Res 19:1351–1360. https://doi.org/10.1021/acs.jproteome.0c00129.
33. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, Chaillon A. 2020. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol 92:602–611. https://doi.org/10.1002/jmv.25731.
34. Luan J, Lu Y, Jin X, Zhang L. 2020. Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection. Biochem Biophys Res Commun https://doi.org/10.1016/j.bbrc.2020.03.047.
35. Gu H, Chu D, Peiris M, Poon L. 2020. Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. bioRxiv https://doi.org/10.1101/2020.02.15.950568.
36. Wu Y. 2020. Strong evolutionary convergence of receptor-binding protein spike between COVID-19 and SARS-related coronaviruses. bioRxiv https://doi.org/10.1101/2020.03.04.975995.
37. Fang B, Liu L, Yu X, Li X, Ye G, Xu J, Zhang L, Zhan F, Liu G, Pan T, Shu Y, Jiang Y. 2020. Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2). bioRxiv https://doi.org/10.1101/2020.03.04.976662.
38. Armijos-Jaramillo V, Yeager J, Muslin C, Perez-Castillo Y. 2020. SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. bioRxiv https://doi.org/10.1101/2020.03.21.001933.
39. Ou J, Zhou Z, Zhang J, Lan W, Zhao S, Wu J, Seto D, Zhang G, Zhang Q. 2020. RBD mutations from circulating SARS-CoV-2 strains enhance the structural stability and human ACE2 affinity of the spike protein. bioRxiv https://doi.org/10.1101/2020.03.15.991844.
40. Sun K, Gu L, Ma L, Duan Y. 2020. Atlas of ACE2 gene expression in mammals reveals novel insights in transmission of SARS-Cov-2. bioRxiv https://doi.org/10.1101/2020.03.30.015644.
41. Chiara M, Horner DS, Gissi C, Pesole G. 2020. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-CoV-2. bioRxiv https://doi.org/10.1101/2020.03.30.016790.