Title: Identification of a pangolin niche for a 2019-nCoV-like coronavirus through an extensive meta-metagenomic search

Authors: Lamia Wahba[1]*, Nimit Jain[1,3]*, Andrew Z. Fire[1,2]*, Karen L. Artiles[1]*, Matthew J. McCoy[1]*, Dae-Eun Jeong[1]*

Affiliations:

[1]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA

[2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

[3]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

*Co-corresponding and equally contributing authors

Author for publication correspondence:

Andrew Z. Fire

afire@stanford.edu

Abstract:

In numerous instances, tracking the biological significance of a nucleic acid sequence can be augmented through the identification of environmental niches in which that sequence is present. Many metagenomic datasets are now available, with deep sequencing of samples from diverse biological niches. While any individual metagenomic dataset can be readily queried using web-based tools, meta-searches through all such datasets are less accessible. In this brief communication, we demonstrate such a meta-meta-genomic approach, examining close matches to the Wuhan coronavirus 2019-nCoV in all high-throughput sequencing datasets in the NCBI Sequence Read Archive accessible with the keyword "virome". In addition to the homology to bat coronaviruses observed in descriptions of the 2019-nCoV sequence [1,2], we note a strong homology to numerous sequence reads in a metavirome dataset generated from the lungs of deceased Pangolins reported by Liu et al. [3]. Our observations are relevant to discussions of the derivation of 2019-nCoV, and illustrate the utility of meta-metagenomic search tools in effective and rapid characterization of potentially significant nucleic acid sequences.

Introduction:

In the early years of nucleic acids sequencing, biological hypothesis generation and discovery were greatly aided by the aggregation of the majority of published DNA and RNA sequences into public sequence databases. 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. As of the early 2000s, sequencing methodologies and computational technologies advanced in tandem so that homology results from a novel sequence were generally available quickly and without substantial cost.

With the development of larger-scale, massively parallel sequencing methodologies, the time and resources to search all extant sequence data became untenable for most studies, although 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 [4,5]. As the diversity and sizes of such datasets expand, the utility of searching them with a novel sequence increases. Meta-metagenomic searches are currently under-utilized-- 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. 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 2019-nCoV coronavirus sequence through a set of metagenomic datasets with the tag "virome".

Materials and Methods:

Computing Hardware: A Linux workstation used for the bulk analysis of metagenomic datasets employs an 8-core i7 Intel Microprocessor, 128G of Random Access Memory, 12TB of conventional disk storage, and 1TB of SSD storage. Additional analysis of individual alignments were carried out with standard consumer-grade computers.

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 datasets 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, in the former case due to network issues and for unknown reasons in the latter case. Thus the workflow continually requests .sra files with prefetch until at least some type of file is obtained, and attempts to unpack into .fasta.gz format until at least one such file is obtained from each .sra file. Metagenomic datasets for analysis were chosen through a keyword search of the SRA descriptions for "virome". The "virome" keyword search will certainly not capture every metagenomic dataset with viral sequences, and likewise not capture every virus in the short sequence read archive. These datasets were downloaded between January 27 and January 31, 2020. With up to 16 threads running simultaneously, total download time (prefetch) was approximately 2 days, with a similar time required for conversion to gzipped fasta files. A total of 9014 sequence datasets were downloaded and converted to fasta.gz files, although a small fraction of these contained very little data (only a few reads or reads of at most a few base pairs). Most files contained large numbers of reads and the total dataset consists of 2.5TB of compressed sequence data, with total bases covered approaching 10TB.

Search Software: For rapid identification of close matches among large numbers of metagenomic reads, we used a simple dictionary based on the 2019-nCoV sequence (NCBI MN908947.3Wuhan-Hu-1) 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 a flu 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 datasets, but it rapidly finds any closely matching

sequence, with the downloading and conversion of the data, rather than the search, being rate limiting.

Alignment of reads with the human 2019-nCoV virus: Reads from the positive pangolin datasets were mapped to the 2019-nCov genome with BWA-MEM (version 0.7.12), using default settings for paired-end mode [6]. Alignments were visualized with the Integrated Genomics Viewer IGV tool [9].

Assessment of nucleotide similarity between 2019-nCoV, Pangolin metavirome reads and closely related bat coronaviruses: All Pangolin metavirome reads that aligned to the 2019-nCov genome with BWA-MEM after adapter trimming with cutadapt were used for calculation. The bat coronavirus genomes were aligned to the 2019-nCov genome in a multiple sequence alignment using the web-interface for Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) with default settings. We note that sequence insertions with respect to the 2019-nCoV genome in either the Pangolin metavirome reads or the bat coronavirus genomes are not accounted for in the similarity traces shown in Figure 1b.

Results:

To identify biological niches that might harbor viruses closely related to 2019-nCoV, we searched through publicly available metaviromic datasets. 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, with some homologies as short as 32 nucleotides captured depending on the precise phasing of the read.

All metagenomic datasets 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 one day for searching by k-mer match on a desktop workstation computer (i7 8-core). Together the datasets included information from 9014 NCBI Short Read Archive entries with (in total) $6.2*10^{10}$ individual reads and $8.4*10^{12}$ base pairs. Despite the relatively large mass of data, the 32-nucleotide k-mer match remains a stringent measure, with spurious matches to the ~30 kb 2019-nCoV genome expected at only 1 in $3*10^{14}$. Positive matches among the metagenomic datasets analyzed were relatively rare, with the vast majority of datasets (8994/9014 or 99.8%) showing no matched 32-mers to 2019-nCoV. Of the datasets with matched k-mers, one was from an apparent synthetic mixture of viral sequences, while the remaining were all from vertebrate animal sources. The matches were from five

studies: two bat-focused studies, one bird-focused study, one small-animal-and-rodent focused study, and a study of pangolins [3] [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 can be inferred from the number of matching reads and k-mer match counts for those reads [Table 1, Supplementary Table 1]. For the 2019-nCoV matches amongst the available metagenomic datasets, the strongest and most abundant matches in these analyses come from the pangolin lung metaviromes. The matches are observed throughout the 2019-nCoV query sequence and many of the matching reads show numerous matching 32-mer sequences. The vast majority of matches were in two lung samples with small numbers of matches in two additional lung datasets. No matches were detected for five additional lung datasets, and no matches were seen in eight spleen samples and a lymph node sample. Further analysis of coverage and homology through alignment of the entire metagenomic datasets revealed an extensive, if incomplete, coverage of the 2019-nCoV genome [Figure 1a]. Percent nucleotide similarity can be calculated for pangolin metavirome reads aligning to 2019-nCoV [Figure 1b], and these segmental homologies consistently show strong matches, approaching (but still overall weaker than) the similarity of the closest known bat coronavirus (RaTG13).

Discussion:

Meta-meta-genomic searching can provide unique opportunities to understand the distribution of nucleic acid sequences in diverse environmental niches. As metagenomic datasets proliferate and as both the need and capability to identify pathogenic agents through sequencing grows, 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 recent 2019-nCoV virus, in so doing raising a number of relevant points for the origin of 2019-nCoV. 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 a virus' full transmission history amongst different biological niches. So even were the closest relative of a virus causing disease in species X 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 2019-nCoV sequence from the posted work of Wu et al., we recapitulated their result that bat-SL-CoVZC45 was the closest related sequence in NCBI's non-redundant (nr/nt) database. Several pangolin metavirome sequences—which are not currently in the (nr/nt) database— are more closely

related to 2019-nCoV 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. [2] described a previously unpublished Coronavirus sequence, designated RaTG13 with much stronger homology to 2019-nCoV than either bat-SL-CoVZC45 or the pangolin reads from Liu et al. While this observation certainly shifts the discussion (legitimately) toward a possible bat-borne intermediate in the chain leading to 2019-nCoV, it remains difficult to determine if any of these are true intermediates in the chain of infectivity.

The match of 2019-nCoV to the pangolin coronavirus sequences also enables a link to substantial context on the pangolin samples from Liu et al., with information on the source of the rescued animals (from smuggling activity), the nature of their deaths despite rescue efforts, and the accompanying pathology. That work describes analyses indicating several coronaviruses present in two of the pangolin lungs as well as other viral species. The pangolins appear to have died from lung-related illness, which may have involved the 2019-nCoV closely-related virus. Notably, however, two of the deceased pangolin lungs had much lower coronavirus signals, while five showed no signal, with sequencing depths in the various lungs roughly comparable. Although it remains possible that the coronavirus was the primary cause of death for these animals, it is also possible (as noted by Liu et al. [3]) 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 2019-nCoV and Pangolin coronavirus sequences in a particular genomic subregion was also noted and discussed in an online forum ("Virological.org") with some extremely valuable sharing of analyses and insights. Matthew Wong and colleagues bring up the homology to the Pangolin metagenomic dataset in this thread and appears to have encountered it through a more targeted search than ours. Several models are discussed in this forum, in parallel to discussions in the virology community, as to the source of 2019-nCoV. The availability of numerous paths (both targeted and agnostic) toward identification of natural niches for pathogenic sequences 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.

Declarations:

Competing interests: The authors declare that they have no competing interests; **Funding:** This study was supported by the following programs, grants and fellowships: Human Frontier Science Program (HFSP) to DEJ, Stanford Genomics Training Program (5T32HG000044-22; PI: M. Snyder) to MJM, and R35GM130366 to AZF. **Authors' contributions:** All authors contributed equally, and author order was selected randomly.

References:

1. 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. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Feb 3. doi: 10.1038/s41586-020-2008-3 PMID: 32015508

2. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020. Epub 2020/02/06. doi: 10.1038/s41586-020-2012-7. PubMed PMID: 32015507.

3. Liu P, Chen W, Chen JP. Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica). Viruses. 2019;11(11). Epub 2019/10/28. doi: 10.3390/v11110979. PubMed PMID: 31652964; PubMed Central PMCID: PMCPMC6893680.

4. Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annu Rev Virol. 2016;3(1):237-61. Epub 2016/09/01. doi: 10.1146/annurev-virology-110615-042301. PubMed PMID: 27578435; PubMed Central PMCID: PMCPMC5457962.

5. Bry L, Falk PG, Midtvedt T, Gordon JI. A model of host-microbial interactions in an open mammalian ecosystem. Science. 1996;273(5280):1380-3. Epub 1996/09/06. doi: 10.1126/science.273.5280.1380. PubMed PMID: 8703071.

6. Bowman JS. Identification of Microbial Dark Matter in Antarctic Environments. Front Microbiol. 2018;9:3165. Epub 2019/01/09. doi: 10.3389/fmicb.2018.03165. PubMed PMID: 30619224; PubMed Central PMCID: PMCPMC6305705.

7. Li H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics. 2012;28(14):1838-44. Epub 2012/05/10. doi: 10.1093/bioinformatics/bts280. PubMed PMID: 22569178; PubMed Central PMCID: PMCPMC3389770.

8. McCarthy TW, Chou HC, Brendel VP. SRAssembler: Selective Recursive local Assembly of homologous genomic regions. BMC Bioinformatics. 2019;20(1):371. Epub 2019/07/04. doi: 10.1186/s12859-019-2949-4. PubMed PMID: 31266441; PubMed Central PMCID: PMCPMC6604332.

9. Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).

Figure Legends:

Figure 1. (a) Integrated Genomics Viewer (IGV) snapshot of alignment. Reads from the pangolin lung virome samples (SRR10168377, SRR10168378 and SRR10168376) were mapped to a 2019-nCoV reference sequence (Genbank accession: MN908947.3). The total number of aligned reads from the three samples was 1107, 313 and 32 reads respectively. **(b) Quantification of nucleotide-level similarity between the 2019-nCoV genome and pangolin lung metavirome reads aligning to the 2019-nCoV genome.** Average similarity was calculated in 101-nucleotide windows along the 2019-nCoV genome, and is only shown for those windows where each nucleotide in the window had coverage ≥ 2. Average nucleotide similarity calculated (in 101-nucleotide windows) between the 2019-nCoV genome and reference genomes of three relevant bat coronaviruses (bat-SL-CoVZC45: accession MG772933.1, bat-SL-CoVZXC21: accession MG772934.1, RaTG13: accession 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 use reference genomes.

Figure 1.

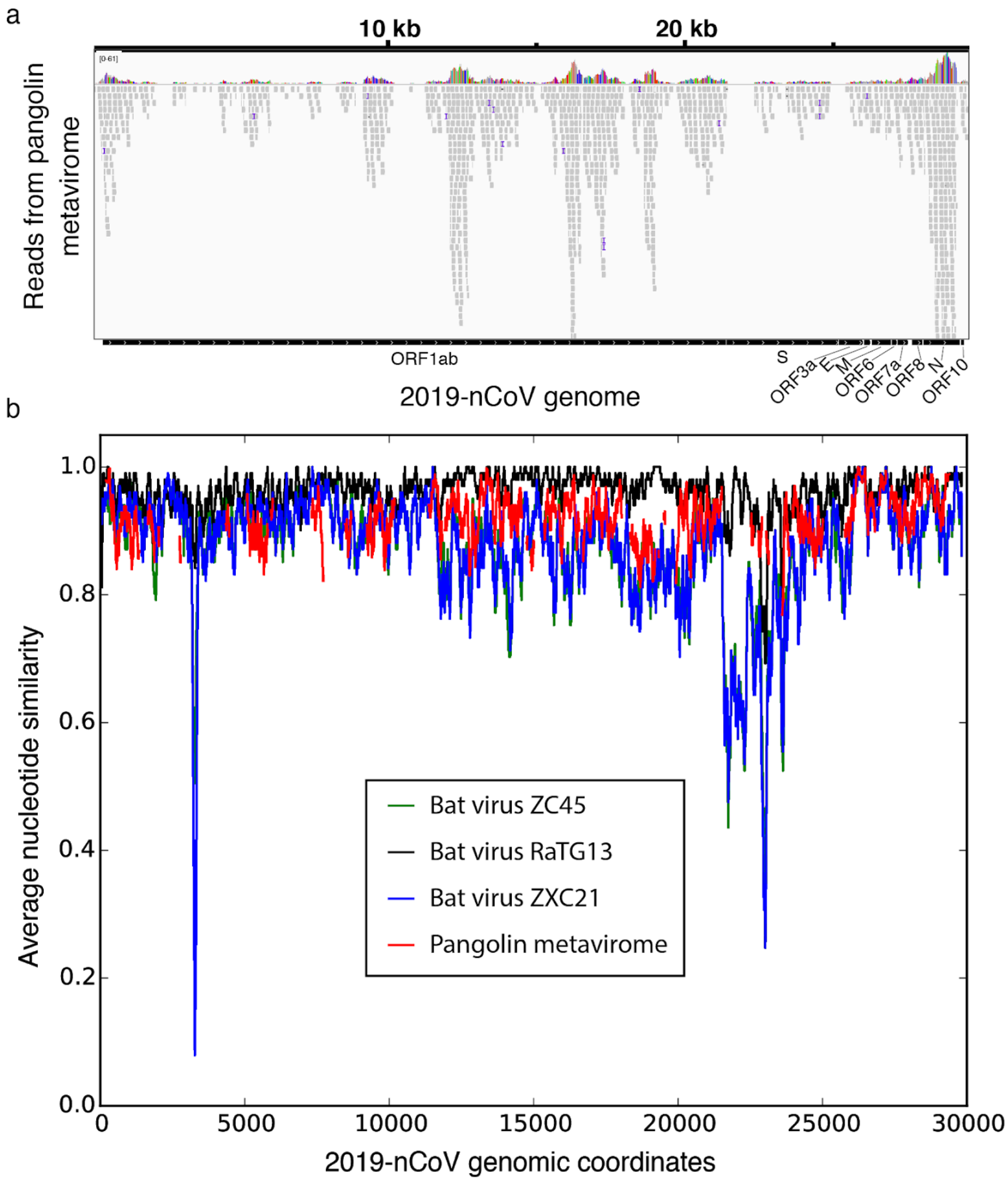
a
10 kb
20 kb
Reads from pangolin metavirome
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10
2019-nCoV genome
b
Average nucleotide similarity
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
30000
2019-nCoV genomic coordinates
Bat virus ZC45
Bat virus RaTG13
Bat virus ZXC21
Pangolin metavirome

Table 1

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 a	SRR5343975_1	19775296	1582023680	1	1
Virome analysis of Rodents a	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 a	SRR5365807_1	21646498	1731719840	2	5
Virome analysis of Rodents a	SRR5365809_1	22799703	1823976240	1	1
Virome analysis of Rodents a	SRR5431767_1	10213803	1031594103	2	2
Virome analysis of Rodents a	SRR5447167_1	13849913	1398841213	2	2
Virome analysis of Rodents a	SRR5447174_1	14063346	1420397946	1	1
Virome analysis of Rodents a	SRR5447175_1	7510256	758535856	1	1
Avocet (RNA Virome in Wild E	SRR7239364_1	22336985	2233698500	2	2

Metagenomic datasets with k=32-mer matches to MN908947.3 2019-nCoV
Details of search and are described in legend to Table S1