COMMENTARY

CONTAMINATION OR VACCINE RESEARCH?

RNA Sequencing data of early COVID-19 patient samples show abnormal presence of vectorized H7N9 hemagglutinin segment

Steven C. Quay[#1], Monali C. Rahalkar[#2], Adrian Jones[#3] and Rahul Bahulikar[4]

[1] Atossa Therapeutics, Inc.

[2] Agharkar Research Institute, Pune, India

[3] Independent Bioinformatics Researcher, Melbourne, Australia

[4]BAIF Research Development, Foundation, Pune, India

#All the three authors share equal authorship

Corresponding authors: Steven@DrQuay.com; ORCID ID; monalirahalkar@aripune.org

Abstract

A re-analysis of the meta-transcriptome data (SRR11092059-63) generated at the Wuhan Institute of Virology (WIV) from bronchial-alveolar lavage fluid (BALF) specimens of five early SARS-CoV-2 patients (WIV02,04,05,06,07-2)[2] was done. The data of these five patients had been obtained by the WIV on two different NGS machines: MiSeq and MGISEQ-2000RS (HiSeq 3000 equivalent). The MGISEQ-2000RS gave 10X more data than the MiSeq machine (5.6-12 Gb) and therefore, it was possible to see more detail. Surprisingly, all the five samples analysed by MGISEQ-2000RS machine showed the presence of a sequence H7N9 'Hemagglutinin A (HA) segment 4' gene in a relatively high proportion, and in one case six-times the abundance of the SARS-CoV-2 sequences. The presence of non-SARS-CoV-2, including these influenza A genes, has been reported earlier, and this data was also used in our current study for comparison and analysis. The surprising finding was the HA segment 4 gene cloned in an expression vector, pVAX1, confirming previously identified vector sequences[3,4]. A WIV publication documented that DNA vaccines containing H7N9 HA genes were being developed and tested in mice in WIV at the same time as the outbreak (2019-2020). In addition, all five samples showed a relatively high proportion of *Spodoptera frugiperda* rhabdovirus (13-83% of SARS-CoV-2 reads). Additionally, the samples also showed the presence of other low-abundance, high homology (LAHH) sequences, mostly of viral origin and not expected to be associated with human BALF specimens. These LAHH sequences could be contaminants, and we identified these viruses as part of previously published research at the WIV, providing a genomic record of prior work. The ability to identify previously performed research in the meta-transcriptome raw data reads from a laboratory provides a new forensic tool. The presence of cloned H7N9 HA gene segment in the transcriptome data of the early five patients processed in the WIV should be treated as an important forensic clue and warrants a full investigation. The most important question considering the plausible hypothesis that the SARS-CoV-2 could have escaped due to a lab accident would be: what does the co-occurrence of vectorized H7N9 sequences with SARS-CoV2 sequences in the early COVID-19 patients suggests?

INTRODUCTION

In December 2019, a pneumonia outbreak started in Wuhan which has turned into the most devastating pandemic of this century. Since, we do not know how SARS-CoV-2 originated, it is of utmost importance to look at all data, especially data obtained in the earliest phase of the pandemic. We found that there were certain anomalies reported in the early patient's data analysed in the WIV (initial five patients-WIV02-07), in the transcriptome data sequenced on the MGISeq machine, which indicated that there was the presence of several non-SARS-CoV-2 sequences in the data. In his publication, Aboulkhair, 2020[1] reported influenza-related genes detected in considerable abundance in the same data set. In his study, public next-generation sequencing data from SARS-CoV-2 infected patients (SRR11092059-63) were analyzed by fastv using unique K-mers. His study reported the actual existence of genome sequences of other microbes in SARS-CoV-2 infected patients. Aboulkhair 2020[1] wrote that surprisingly, along with the SARS-CoV-2 virus genome, the Influenza type A segment 4 hemagglutinin (HA) gene was found. He also reported the genome sequence of the Nipah virus was detected in WIV05, WIV06, and WIV07 patients. Aboulkhair 2020[1] also reported the presence of other viruses such as human immunodeficiency virus (in one patient), rhabdovirus, human metapneumovirus, Human adenovirus, Human herpesvirus 1, coronavirus NL63, parvovirus, simian virus 40, and hepatitis virus genomes sequences in SARS-CoV-2 infected patients.
Later on, Quay 2021[3] and Zhang 2021[4], reported that there were vectors present in the same transcriptome data indicating that the H7N9 fragment could be attached to a vector.

The present study was initiated to validate the results obtained in earlier studies by Aboulkhair 2020[1], Quay 2021,[3] and Zhang 2021[4] and to correlate the overall findings from these studies. We also wanted to find out why the non-SARS-CoV-2 sequences were seen in these early patients, the nature of the sequences, and their probable origin.

RESULTS

Avian flu (H7N9 hemagglutinin segment 4) with vector sequences in early patients' data analysed by the WIV

The Wuhan Institute of Virology published a description of five early patients with COVID-19 that has been one of the most cited papers[2] from the early days of the pandemic. Previous reports have examined the raw SRA files from this WIV paper and have identified significant anomalous sequence data[1, 3,4]. In our analysis using fastv, we found that H7N9 hemagglutinin gene was detected in a high abundance and corresponded to a considerable proportion of the reads. In a patient (sample name WIV07-2), the H7N9 segment 4 reads were six times the abundance of SARS-CoV-2 reads (Supplementary Table 1, Table 1).

BioProject	Experiment	Run	Sample Name	SARS-CoV-2 (NC	H7N9 HA gene (N	H7N9 PB2 gene (	H7N9 M2 and M1	Nipah virus (NC	Human immunoc	Japanese encepl	Simian virus 40 (	Spodoptera frugiperda rhabdovirus (NC_025382.1)
PRJNA605983	SRX7730884	SRR11092059	WIV07-2	12921	76284	351	47	234	121	119	379	2284
PRJNA605983	SRX7730883	SRR11092060	WIV06-2	6400	257	4		25			12	2826
PRJNA605983	SRX7730882	SRR11092061	WIV05	3270	409	19		83			7	2719
PRJNA605983	SRX7730881	(SRR11092062)	WIV04-2	33469	1459	5					77	4508
PRJNA605983	SRX7730880	(SRR11092063)	WIV02-2	4043	485	5						2709

Supplementary Table 1 Fastv K-Mer Hit Summary

Run (NCBI SRR number)	Machine name	Sample name	SARS-CoV-2 K-mer hits	H7N9 HA segment 4* K-mer hits	Ratio of HA/CoV-2 (Percentage)
SRR11092059	MGISEQ-2000RS	WIV07-2	12921	76,284	600
SRR11092060	MGISEQ-2000RS	WIV06-2	6400	257	4
SRR11092061	MGISEQ-2000RS	WIV05-2	3270	409	12
SRR11092062	MGISEQ-2000RS	WIV04-2	33469	1459	4
SRR11092063	MGISEQ-2000RS	WIV02-2	4043	485	12

*Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene

Table 1: The read numbers with the corresponding K-mer hits of HA segment 4 of the H7N9 virus versus SARS-CoV-2 percentage (our study).

In our work, all five samples analysed by the MGISEQ2000RS machine showed a significant proportion of hemagglutinin (HA) from the same H7N9 virus Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 HA gene. The ratio of SARS-CoV-2: HA segment

was 4-600%, Table 1. Our results are consistent with Aboulkhair 2020, where similar k-mer values and proportions of HA/SARS-CoV-2 were detected. Surprisingly, the proportion of HA gene in patient id WIV-07-02, run SRR11092059, showed that the proportion of HA/SARS-CoV2 was about 5.9, almost 600% of the SARs-CoV-2 reads. Surprisingly, the SARS-CoV-2 proportion was much lesser than the HA in this particular patient (WIV07-2). In other cases, the HA genes were 4-12% of the SARS-CoV-2 reads. Aboulkhair 2020[1] had also found that the Influenza type A segment 4 hemagglutinin (HA) gene showed high coverage (100%, 100%, 98.57%, 95.71%, and 100%) and mean depth (5.58, 16.52, 4.8, 2.95, and 881.37) in WIV02, WIV04, WIV05, WIV06, and WIV07 patients, respectively.

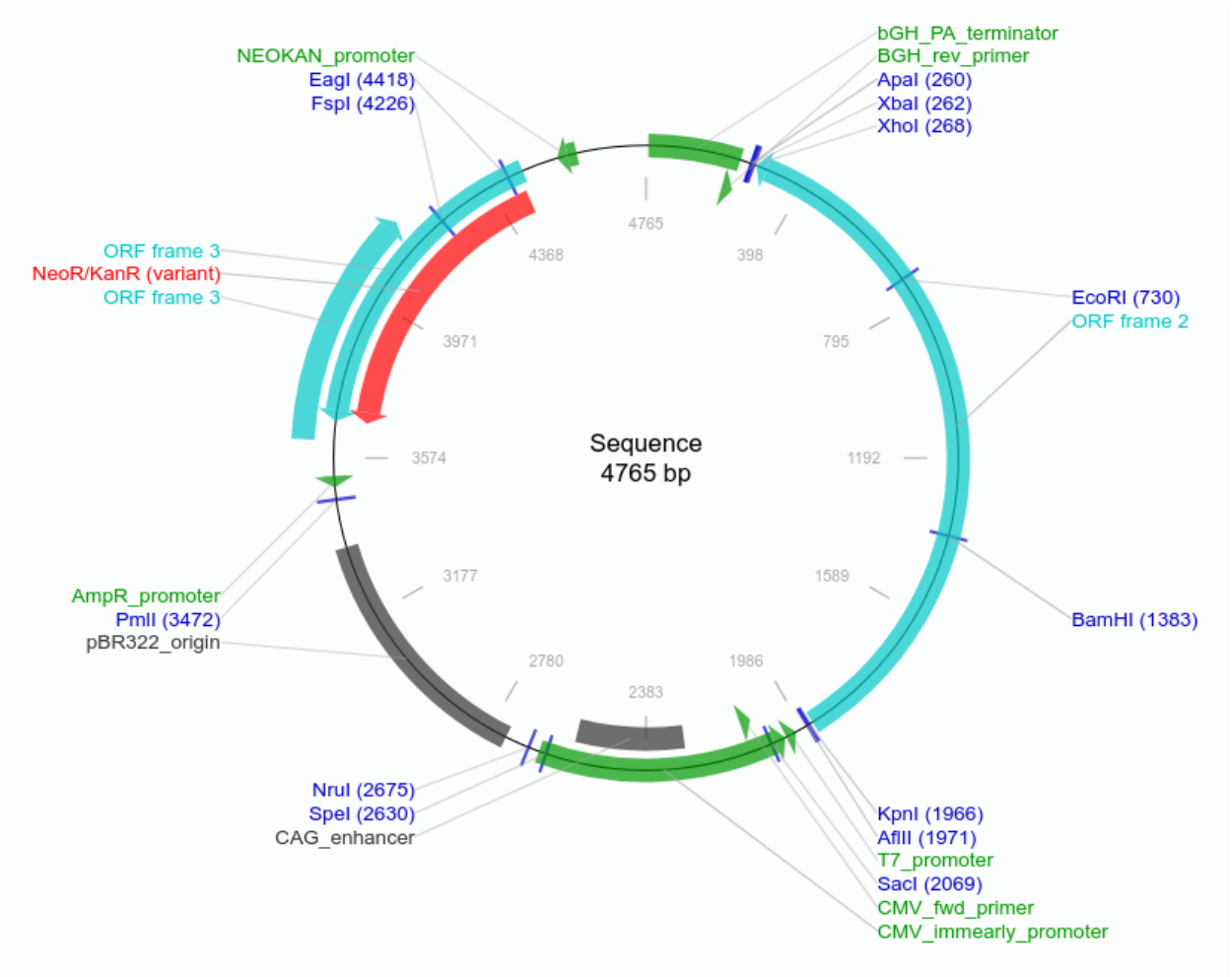

Fig. 1. pVAX1 plasmid containing Influenza A H7N9 HA gene from de novo assembly of SRR11092059 and plotted in AddGene.

A further surprising fact was that pVAX1, a DNA vaccine expression vector was found containing the cloned HA gene. A complete pVAX1 plasmid sequence (Fig. 1) (Supplementary Sequence file) was identified from WIV07-2 SRA dataset SRR11092059 as a single contig

(Suppl. Fig. 10) and was also identified in WIV04-2 (SRR11092062) and WIV02-2 (SRR11092063).

The ORF 2 region (274-1956bp) was analysed using NCBI BLASTN for alignment to the nt database and 100% homology was found to Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds KF021597.1 and Influenza A virus (A/chicken/Guangdong/G2/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds KJ395948.1 and synthetic constructs (Supp. Fig. 11).

The Hemagglutinin gene (HA) has a sequence coverage that is significantly above the coverage of the other known contaminants.

To determine if the presence of the influenza HA gene in these specimens was most consistent with being a low abundance, high homology contaminant from the laboratory or with being different than simple equipment or reagent contamination, the percent coverage of the HA gene was compared to the coverage for a true-positive specimen virus, SARS-CoV-2, on the one hand, and to the sequence coverage for the above contaminants on the other hand.

In order to compute the p-values for whether the viruses in question were either above or below coverage for the known contaminants, we examined the distribution of coverages for all of the known contaminants across the five subjects. These data were evaluated using the Box-Cox family of power transformations in order to ascertain the optimal transformation to normality. This turned out to be $X^{-0.1}$. All coverage values for these contaminants were transformed and the mean and standard deviation computed (1.3216 ± 0.1518). From these values, it is possible to compute a z-value for each of the transformed coverage values for the viruses of interest. These z-values were converted into probabilities using the standard normal distribution. For any coverage above the mean (approximately 6.2% untransformed), the probability was computed from the upper tail of the distribution (representing the probability that the coverage for the given virus was higher than for the contaminants), and those below the mean were competed from the lower tails (representing the probability that the coverage for the given virus was lower than for the contaminants.

The p-values for the coverage of SARS-CoV-2 and the H7N9 HA gene in all five specimens is significantly higher than the mean coverage for the known contaminants (Table 2). On the other hand, the additional influenza-related sequences are not statistically different from the

LAHH contamination sequences. The SARS-CoV-2 results are highlighted in yellow, the H7N9 HA genes in green, and all other influenza sequences in brown.

SRR Identifier	Identity of the reads	K-mer	Coverage	Read Depth	p-values
SRR11092064					
	Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1	1404	72.56%	2.21451	0.028
SRR11092063					
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene	391	100%	5.58571	0.017
	Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1	4965	87.54%	7.83123	0.021
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 1 polymerase PB2 (PB2) gene	5	4.50%	0.045045	0.393
SRR11092062					
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene	1157	100%	16.5286	0.017
	Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1	68528	100%	108.088	0.017
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 1 polymerase PB2 (PB2) gene	3	2.70%	0.027027	0.228
SRR11092061					
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene	336	98.57%	4.8	0.017
	Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1	2622	89.27%	4.13565	0.02
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 1 polymerase PB2 (PB2) gene	14	9.91%	0.126126	0.657
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 2 polymerase PB1 (PB1) and PB1-F2 protein (PB1-F2) genes	8	7.34%	0.0733945	0.562
SRR11092060					
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene	207	95.71%	2.95714	0.0182
	Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1	5085	95.43%	8.0205	0.0185
	Influenza A virus na gene for neuraminidase, genomic RNA, strain A/Hong Kong/1073/99(H9N2)	4	3.39%	0.0677966	0.296
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 1 polymerase PB2 (PB2) gene	4	1.80%	0.036036	0.128
SRR11092059					
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene	61696	100%	881.371	0.017
	Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1	10468	99.37%	16.511	0.017
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 1 polymerase PB2 (PB2) gene	276	36.94%	2.48649	0.077
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 8 nuclear export protein (NEP) and nonstructural protein 1 (NS1) genes	17	15.63%	0.53125	0.219
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 2 polymerase PB1 (PB1) and PB1-F2 protein (PB1-F2) genes	18	9.17%	0.165138	0.367
	Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 1	8	8.89%	0.0888889	
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 7 matrix protein 2 (M2) and matrix protein 1 (M1) genes	40	8.70%	0.869565	0.384
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 5 nucleocapsid protein (NP) gene	5	6.98%	0.116279	0.456
	Influenza A virus (A/New York/392/2004(H3N2)) segment 7	4	6.67%	0.133333	
	Influenza A virus (A/Hong Kong/1073/99(H9N2)) segment 7	3	5.41%	0.0810811	
	Influenza A virus (A/Korea/426/1968(H2N2)) segment 7	6	4.17%	0.25	
	Influenza A virus (A/goose/Guangdong/1/1996(H5N1)) segment 7	5	4.17%	0.104167	
	Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 3 polymerase PA (PA) and PA-X protein (PA-X) genes	6	3.41%	0.0681818	0.298
	Influenza A virus (A/California/07/2009(H1N1)) segment 4 hemagglutinin (HA) gene	3	2.86%	0.0428571	
	Influenza A virus ha gene for Hemagglutinin, genomic RNA, strain A/Hong Kong/1073/99(H9N2)	6	2.60%	0.0779221	
	Influenza A virus pb2 gene for polymerase Pb2, genomic RNA, strain A/Hong Kong/1073/99(H9N2)	10	2.33%	0.116279	
	Influenza A virus pb1 gene for polymerase Pb1, genomic RNA, strain A/Hong Kong/1073/99(H9N2)	2	2.04%	0.0204082	
	Influenza A virus na gene for neuraminidase, genomic RNA, strain A/Hong Kong/1073/99(H9N2)	4	1.69%	0.0677966	
	Influenza A virus (A/California/07/2009(H1N1)) segment 3 polymerase PA (PA) gene	2	1.33%	0.0266667	

Table 2: The p-values for the coverage of SARS-CoV-2 and the H7N9 HA gene in all five specimens is significantly higher than the mean coverage for the known contaminants (k-mer data table was used from[1] for the analysis.

High proportion of insect cell line viruses, e.g., *Spodoptera frugiperda* rhabdovirus (13-83% of SARs-CoV-2)

In our analysis, we re-confirmed the results which were obtained by Aboulkhair 2020, that the data set showed a high abundance of *Spodoptera frugiperda* rhabdovirus, a virus associated with insect cell lines and the reads were 13-83% of the SARS-CoV-2 reads. These observations are in congruence to the ones reported by Aboulkhair, (2020). He had also reported similar proportions of *Spodoptera frugiperda* related K-mers. Additionally**,** Aboulkhair, (2020) had

also reported the presence of *Autographa californica* nucleopolyhedrovirus reads also in high proportion. The *Spodoptera frugiperda* rhabdovirus could come from the Sf9 cell line is broadly used for manufacturing baculovirus-expressed viral vaccines[5]. Similarly, *Autographa californica* nucleopolyhedrovirus (AcMNPV), the type member of the virus family Baculoviridae, infects pest insects and has been the subject of many studies for its development as a biopesticide. It is also the virus upon which most of the commercial baculovirus protein expression systems are based.[5]

Run (NCBI SRR number)	Platform for NGS	Sample name	SARS-CoV-2 K-mer hits	*Spodoptera frugiperda* rhabdovirus K-mer hits	Ratio of SF/CoV-2 (%)
SRR11092059	MGISEQ-2000RS	WIV07-2	12921	2284	18
SRR11092060	MGISEQ-2000RS	WIV06-2	6400	2826	44
SRR11092061	MGISEQ-2000RS	WIV05-2	3270	2719	83
SRR11092062	MGISEQ-2000RS	WIV04-2	33469	4508	13
SRR11092063	MGISEQ-2000RS	WIV02-2	4043	2709	67

Table 3. The read numbers with the corresponding K-mer hits of *Spodoptera frugiperda* rhabdovirus versus SARS-CoV-2 percentage.

Large number of LAHH (low abundance high homology sequences) affiliated to other viruses

The paper by Abouelkhair[1] reports on the large-scale LAHH contamination of the five specimens sequenced on the MGI-SEQ2000RS at the Wuhan Institute of Virology. We confirmed this by re-analysis and could find most of these, namely, Nipah, *Brevundimonas* phage vB_BsubS-Delta, Saccharomyces 20S RNA narnavirus, etc (Supplementary Table.xlsx).

A "contamination signature" of three sequences is present in similar proportions in all WIV specimens, suggesting, at least in part, a shared contamination process or step.

Table 4 was prepared with the three most abundant LAHH sequences, excluding the SARS-CoV-2 and Influenza A H7N9 HA gene. If one calculates the mean and standard deviation of the coverage data for the three sequences identified one obtains values of Autographa; 0.191 ± 0.037; *Spodoptera*: 0.618 ± 0.081; Bamboo mosaic virus: 0.085 ± 0.019. Since all values are within two standard deviations of the mean, one can conclude that there is a common reagent, sample processing step, machine part, etc. that can account for some of the contamination.

On the other hand, each specimen has one or more sequences identified that are unique to that specimen suggesting that, in addition to the commonality of the above contamination, there are also specimen-specific contaminations identified.

SRR Identifier	Identity of the reads	K-mer	Coverage	Read Depth
SRR11092063				
	Autographa californica nucleopolyhedrovirus	4002	18.82%	0.506454
	Spodoptera frugiperda rhabdovirus isolate Sf	2227	57.36%	4.82035
	Bamboo mosaic virus satellite RNA	17	9.76%	0.207317
SRR11092062				
	Autographa californica nucleopolyhedrovirus	6725	25.30%	0.85105
	Spodoptera frugiperda rhabdovirus isolate Sf	3757	72.73%	8.13203
	Bamboo mosaic virus satellite RNA	20	7.32%	0.243902
SRR11092061				
	Autographa californica nucleopolyhedrovirus	3388	15.87%	0.428752
	Spodoptera frugiperda rhabdovirus isolate Sf	2239	57.58%	4.84632
	Bamboo mosaic virus satellite RNA	25	10.98%	0.304878
SRR11092060				
	Autographa californica nucleopolyhedrovirus	3833	19.25%	0.485067
	Spodoptera frugiperda rhabdovirus isolate Sf	2351	67.75%	5.08874
	NC_003497.1 Bamboo mosaic virus satellite RNA	12	6.10%	0.146341
SRR11092059				
	Autographa californica nucleopolyhedrovirus	2991	16.50%	0.378512
	Spodoptera frugiperda rhabdovirus isolate Sf	1877	53.46%	4.06277
	Bamboo mosaic virus satellite RNA	15	8.54%	0.182927

Table 4: Percentage of Sf rhabdovirus, Autographa, and Bamboo mosaic virus in all the samples (k-mer data from Aboulkhair 2020[1] was used for the calculations).

Unique nature of WIV07-2 sample: Six-times HA gene in comparison to SARS-CoV-2 reads

The WIV07-2 showed six-times the number of HA reads compared to SARS-CoV-2 (Supplementary Table 1, Table 1). Exactly the same observations were made by Aboulkhair 2020. Also, this sample showed a large number of other sequences related to influenza and also HIV (our data and data from[1]). According to Aboulkhair 2020, the influenza genes were 8%, SARS-CoV-2 2%, and rhabdovirus 0.6%. Our data also indicated six-times the sequence reads for HA compared to SARS-CoV-2. As all the samples were run in the same MGISEQ-2000 machine using the same protocol, the detection of vectorized hemagglutinin gene in higher proportion compared to the actual pathogen, SARS-CoV-2, at least in one case, raises a red flag. The results indicate that this could not be mere low-level contamination, as the k-mer reads in WIV07-2 were 76,000 and the SARS-CoV-2 is ~14,000. Additionally a complete pVAX1 plasmid sequence was identified from WIV07-2 SRA dataset SRR11092059 as a

single contig (Supp. Fig. 10), The ORF 2 region (274-1956bp) was analysed using NCBI BLASTN for alignment to the nt database and 100% homology was found to Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds KF021597.1 and Influenza A virus (A/chicken/Guangdong/G2/2013(H7N9)) segment 4 hemagglutinin (HA) gene, complete cds KJ395948.1 and synthetic constructs (Supp. Fig. 11). This particular sample (WIV07-2) also showed a large number of bacteria such as *Acinetobacter*, *Brevundimonas*, *Enterobacter*, etc. which could be an indication of secondary bacterial infections. Since the machine used was the same in all five cases, the high number of HA reads in WIV07-2 is highly strange and significant.

DISCUSSION

The present paper concerns the earliest COVID-19 patient samples sequenced at the Wuhan Institute of Virology. These are the samples that were used for RNA sequencing/ transcriptome used to investigate the pathogen in the samples of the cluster pneumonia cases, which occurred in Wuhan in December 2019. The sequencing was done in the WIV using two NGS platforms: MGISeq2000RS and Miseq (Zhou 2020, NCBI data).

The MGISeq2000 machine used by the WIV is a highly sensitive machine that would give a large amount of data from the patients' samples. This machine is equivalent to HiSeq3000 platform. The same samples were also analysed by MiSeq machine (Supplementary Table. xlsx). However, the MiSeq dataset size is 10 to 20 times lower (~0.5 Gb to 1.2 Gb), in contrast to 5-12 Gb data obtained in the MGISeq2000 sequencing. Hence, the MGISeq data seems to be of utmost importance in this case.

The ability to identify previously performed research in the meta-transcriptome raw data reads from a laboratory provides a new forensic tool[1,4]. Previous work by Zhang et al.[8] documented the unexpected discovery of multiple coronaviruses and a BSL-3 pathogen in agricultural cotton and rice sequencing datasets, including a novel HKU5-related Merbecovirus in a cotton dataset sequenced by the Huazhong Agricultural University in 2017. These two studies demonstrate that meta-transcriptome data sets can be forensically probed for unreported research activities.

It was found by Aboulkhair 2020[1] that there was the presence of non-SARS-CoV-2 sequences in these samples, including that of influenza virus. We confirmed this observation and found that the majority of influenza genes were of H7N9 HA segment 4, which was cloned in a human expression vector pVAX1, often used for the creation of DNA vaccines. Similar observations were made earlier[1] where among viral communities, influenza type A (8%), rather than SARS-CoV-2 (2%), was found to be dominant in the WIV07-2 patient. In the same paper by Abouelkheir 2020 SARS-CoV-2, rhabdovirus, and influenza type A dominated the sequence data from patients WIV06-2 and WIV04-2.

Avian influenza A(H7N9) is a subtype of influenza viruses that have been detected in birds in the past. After no reported human cases of highly pathogenic avian influenza (HPAI) H7N9 for over a year, a fatal case occurred in late March 2019[7]. The currently available epidemiological and virological evidence suggests that this virus has not acquired the ability of sustained transmission among humans; thus the likelihood of human-to-human transmission of the A(H7N9) virus is low (WHO). Therefore, the possibility that the H7N9 HA being from a natural infection is very low. Moreover, only HA segment 4 cloned in a vector were found, which indicates the laboratory origin of this material.

Two hypotheses regarding the H7N9 HA segment in pVAX1 are offered: Either it was contamination during the library preparation or in the sequencer or the vectorized HA was present in the patients' samples. The second possibility would raise a serious alarm, as to why this type of artificial construct was present in patients' samples. There are no records of any vaccine administered specifically for H7N9 in Wuhan at that time. We know of no such vaccine trials being conducted in humans in China. There are published references showing that H7N9 sequences and constructs using plasmid vectors were being studied at the time in WIV, including a group working in the WIV preparing DNA vaccines against avian influenza using mice.[6] The researchers in WIV were experimenting with DNA vaccination using consensuses H7 which elicited a broad antibody response against H7 lineage viruses. For these experiments, they were using four H7 HA protein sequences generated and named CH7-22, CH7-24, CH7-26, and CH7-28.

The highly intriguing and simultaneous study by a group in WIV[6] involved H7N9, and was used to infect six- to 8-week-old female BALB/c mice (10 mice per group) and were immunized twice with a 3-week interval. In the study, DNA vaccines encoding consensus H7

proteins elicited broadly reactive antibody responses in mice. Six to 8-week-old female BALB/c mice (10 mice per group) were immunized twice with a 3-week interval. Each vaccination consisted of 30 μg of pCH7-22, pCH7-24, pCH7-26, pCH7-28, H7N9, and pH7N7 dissolved in 30 μL Tris-EDTA buffer. These were injected into mice. They had also found that Eurasian lineage H7N9 and pH7N7 vaccination of mice had less efficacy against American lineage H7N3 viral infection compared with consensus pCH7-22 or pCH7-24 proteins.[6] The HA genes of A/Shanghai/02/2013 (H7N9) and A/Phalacrocorax carbo/Hubei/HH179/2013(H7N7) were amplified and cloned into pVAX1 to generate H7N9 and pH7N7.[6]

Since the pVAX1 is capable of expression in human systems and is used as a DNA vaccine, there is a possibility that it can pass from human to human through respiratory droplets. Given the possible replication, proliferation, and expression ability of this cloned HA gene in human cells, it is a possibility that this could be actually present in the patient's BALF and not just explained by contamination in the sequencing facility. Aboulkhair, (2020) also hypothesised that the patients had co-infection with the influenza A, rather than the flu sequences being mere contamination. However, as the influenza segment was an H7N9 HA segment 4 in a pVAX1 vector, this could not be a natural source of infection but indicates a laboratory source of this vectorized segment.

The low abundance anomalous sequence reads found in the RNA seq data correspond to previous research activities at the WIV

The sequences of viruses found in low or high abundance were compared to the research done in WIV. To verify that these sequences were legitimate contaminants within the WIV a table was created using an internet search engine to provide references for the listed term in the first column together with "Wuhan Institute of Virology."

The hyperlinks in the second column are representative papers documenting publications with the identified sequences were in fact being studied.

LAHH-related Genomes	**Publications Related to Sequences Identified**
Autographa californica nucleopolyhedrovirus	Autographa Californica Multiple Nucleopolyhedrovirus Enters Host Cells
Avian leukosis virus - RSA	Avian leukosis virus

Bamboo mosaic virus satellite RNA	Bamboo mosaic virus
Cactus virus X	China National Center for Bioinformation
Coliphage phi-X174	**Illumina PhiX control library**
Hepatitis delta virus	Hepatitis delta virus (HDV) ribozyme construct
Human immunodeficiency virus 1	Human T-lymphoytopic virus
Human T-lymphotropic virus 1	Human T-Cell Leukemia Virus Type 1 Particle Morphology
Human T-lymphotropic virus 1	Human T-cell Leukemia Virus Type 1
Influenza A virus (A/Shanghai/02/2013(H7N9)) segment 4 hemagglutinin (HA) gene	Vaccination with Consensus H7 Elicits Broadly Reactive and Protective Antibodies against Eurasian and North American Lineage H7 Viruses
Japanese encephalitis virus, genome	Replication-defective JEV vaccine candidate
Moloney murine leukemia virus	Four novel bat hepadnaviruses and a hepevirus in China
Nipah virus	Assay for Rapid and Specific Detection of All Known Nipah virus Strains
Ralstonia_5_7_47FAA_uid39415	A bacteriophage cocktail for biocontrol of potato bacterial wilt
Simian virus 40	Virus-Based Nanoparticles of Simian Virus 40
Spodoptera frugiperda rhabdovirus isolate Sf	*Spodoptera frugiperda* Nucleopolyhedrovirus from China
Tobacco mosaic virus	TMV was obtained from Wuhan Institute of Virology
Woodchuck hepatitis virus	Gain of function human hepatitis B gene inserted into woodchuck virus

The above analysis documents that all twenty viral sequences found in these specimens were under study at the WIV during or before December 2019.

OPEN QUESTIONS

The presence of high reads of H7N9 HA fragment in the RNA sample of the BALF RNA of the five patients is highly surprising and generates two main questions:

Was the H7N9 in an expression vector contamination in the sequencing facility or a leak of a vaccine pH7N9?

Considering the first possibility, how can WIV explain the data of WIV07-2, the number of reads contributed by H7N9 is six times that of Sars-CoV-2? Also, the pVAX1 with the cloned hemagglutinin segment is clearly seen as a single contig in the WIV07-2 sample. Also, the coverage for the HA segment was almost complete in all the cases.
Considering that the HA reads were an actual part of the human BALF samples, how did the H7N9 enter the humans when there are no records of trials of any oral or nasal vaccines being used at that time in Wuhan? Were these people infected with some lab material, or a lab worker was infected with a mouse with this infectious material? It is well known that WIV was working with SARS-like coronaviruses and used humanized mice, i.e., hACE-2 expressing mice. So, could cross-contamination of the mice with H7N9 and SARS-CoV-2 be a possibility, looking at the co-occurrence of these sequences?

The highly abundant *Spodoptera frugiperda* rhabdovirus and Autographa viruses which are usually present in insect cell lines used for vaccine research, raises further questions. The presence of other numerous contaminant viruses, including Nipah, which were detected in lower proportions, raises the critical question of the kind of hygiene and safety level in the sequencing laboratory of WIV.

In conclusion, the RNA sequencing data of the earliest patient samples analyzed in WIV on the MGISeq2000 machine raise several critical questions. The HA segment found in the five patients was not of the natural origin as it was cloned in a vector and showed affiliation to only segment 4. The most important question would be: considering the plausible hypothesis that SARS-CoV-2 could have escaped due to a lab accident, what does the co-occurrence of vectorized H7N9 sequences with SARS-CoV2 sequences in the early COVID-19 patients, suggests?

References

[1] Aboulkhair, M. (2020). Non-SARS-CoV-2 genome sequences identified in clinical samples from COVID-19 infected patients: Evidence for co-infections. *Peerj* DOI 10.7717/peerj.10246.

[2] Zhou, P., Yang, XL., Wang, XG. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7

[3] Quay SC. (2021) A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory derived https://zenodo.org/record/4642956#.YNM7KehKjOg

[4] Zhang D. (2021) Vector sequences in early WIV SRA sequencing data of SARS-CoV-2 inform on a potential large-scale security breach at the beginning of the COVID-19 pandemic https://zenodo.org/record/4486195#.YNM6uuhKjOg

[5] Ma H, Nandakumar S, Bae EH, Chin PJ, Khan AS. The *Spodoptera frugiperda* Sf9 cell line is a heterogeneous population of rhabdovirus-infected and virus-negative cells: Isolation and characterization of cell clones containing rhabdovirus X-gene variants and virus-negative cell clones. Virology. 2019 Oct;536: 125-133. doi: 10.1016/j.virol.2019.08.001. Epub 2019 Aug 2. PMID: 31494355.

[6] Fadlallah GM, Ma F, Zhang Z, et al. Vaccination with Consensus H7 Elicits Broadly Reactive and Protective Antibodies against Eurasian and North American Lineage H7 Viruses. *Vaccines (Basel)*. 2020;8(1):143. Published 2020 Mar 23. doi:10.3390/vaccines8010143

[7] Yu, D, Xiang G, Zhu W, et al. The re-emergence of highly pathogenic avian influenza H7N9 viruses in humans in mainland China, 2019. Eur Surveill. 2019 May 23; 24(21): 1900273. doi: 10.2807/1560-7917.ES.2019.24.21.1900273

[8] Zhang D, Jones A, Deigin Y, Sirotkin K, Sousa A. Unexpected novel Merbecovirus discoveries in agricultural sequencing datasets from Wuhan, China. https://arxiv.org/abs/2104.01533

Additional Data is available as Supplemental Analyses (Fig 1-12) and a sequence file of pVAX1 with the HA gene. An excel sheet with the fast-v analysis and blasts (Suppl_Table_SARS-CoV2-Early Patients)

Methods

SRA datasets from PRJNA605983 (Zhou et al. 2020) were analysed using fastv version 0.9.0 (Chen S. 2020) against the Opengene vial genome kmer collection 'microbial.kc.fasta.gz' (https://github.com/OpenGene/UniqueKMER). A summary table of read hits against key viruses can be found in Supp. Table 1. Viruses and microbes with fastv K-mer hits in four or more SRA datasets in PRJNA605983 were analysed using a correlation matrix using pearson and spearman correlation coefficients using Seaborn version 0.11.1 (Waskom, 2021). Box plots of read abundance were created with Pandas version 1.15 (The pandas development team, 2020).

Each SRA dataset in PRJNA605983 was de novo assembled using MEGAHIT v1.2.9 (Li et al. 2015) using default parameters. Addgene sequence analyzer (https://www.addgene.org/analyze-sequence/) was used for contig analysis.

Raw reads from SRR11092059 were aligned to NC_026427.1 (H7N9 M gene) using BWA MEM version 0.7.17 (Li, 2013) with default parameters. Reads from SRR11092059 were separately aligned to an identified pVax1 plasmid (de novo assembled contig k141_31) using bowtie2 version 2.4.2 (Langmead et al. 2012) to assess coverage and depth. GATK version 4.1.9 (Van der Auwera & O'Connor, 2020) was used for sorting and duplicate marking. Samtools version 1.11 (Li et al. 2009) was used for indexing for viewing in IGV. IGV version 2.8.13 (Thorvaldsdóttir et al. 2013) was used for read alignment analysis and plotting.

References for the Method Section:

Chen S, He C, Li Y, Li Z, Melançon CE. A computational toolset for rapid identification of SARS-CoV-2, other viruses and microorganisms from sequencing data. Brief Bioinform. 2020;00(August):1-12. doi: https://doi.org/10.1101/2020.05.12.092163

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-359. doi:10.1038/nmeth.1923

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

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. arXiv:1303.3997v1 [q-bio.GN]

Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-1676. doi:10.1093/bioinformatics/btv033.

The pandas development team. pandas-dev/pandas: Pandas. Feb. 2020. Zenodo. doi: 10.5281/zenodo.3509134

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178-192. doi:10.1093/bib/bbs017

Van der Auwera GA & O'Connor BD. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media.

Waskom M. Seaborn: Statistical Data Visualization. J Open Source Softw. 2021;6(60):3021. doi:10.21105/joss.03021

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

Acknowledgments:

We thank Prof. Virginie Courtier, H. Lawrence Remmel, and Dr. M. Aboulkhair for the important discussions during the writing of this pre-print.