medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Epidemiological and Genomic Analysis of SARS-CoV-2 in Ten Patients from a Mid-sized City outside of Hubei, China

Jinkun Chen[1§], Evann E. Hilt[2§], Huan Wu[3], Zhuojing Jiang[1], QinChao Zhang[1], JiLing Wang[1], Yifang Wang[3], Fan Li[4], Ziqin Li[5], Jialiang Tang[1*], Shangxin Yang[2,5*]

[1]Shaoxing Center for Disease Control and Prevention, Shaoxing, Zhejiang, China; [2]Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, CA, USA; [3]IngeniGen XunMinKang Biotechnology Inc., Shaoxing, Zhejiang, China; [4]Three Coin Analytics, Inc. Pleasanton, CA, USA; [5]Zhejiang-Californina International Nanosystems Institute, Zhejiang University, Hangzhou, Zhejiang, China

§Authors contributed equally.

*Corresponding authors: Jialiang Tang: 992488904@qq.com; Shangxin Yang, PhD, D(ABMM): shangxinyang@mednet.ucla.edu

RUNNING TITLE: Genomic Analysis of SARS-CoV-2 by Metagenomic Sequencing

KEYWORDS: SARS-CoV-2, Metagenomic Sequencing, Mutation Rate, S Genotype, COVID-19, 2019-nCoV

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ABSTRACT

A novel coronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the ongoing COVID-19 pandemic. In this study, we performed a comprehensive epidemiological and genomic analysis of SARS-CoV-2 genomes from ten patients in Shaoxing, a mid-sized city outside of the epicenter Hubei province, China, during the early stage of the outbreak (late January to early February, 2020). We obtained viral genomes with > 99% coverage and a mean depth of 296X demonstrating that viral genomic analysis is feasible via metagenomics sequencing directly on nasopharyngeal samples with SARS-CoV-2 Real-time PCR C_t values less than 28. We found that a cluster of 4 patients with travel history to Hubei shared the exact same virus with patients from Wuhan, Taiwan, Belgium and Australia, highlighting how quickly this virus spread to the globe. The virus from another cluster of two family members living together without travel history but with a sick contact of a confirmed case from another city outside of Hubei accumulated significantly more mutations (9 SNPs vs average 4 SNPs), suggesting a complex and dynamic nature of this outbreak. We also found 70% patients in this study had the S genotype, consistent with an early study showing a higher prevalence of S genotype out of Hubei than that inside Hubei. We calculated an average mutation rate of 1.37×10^{-3} nucleotide substitution per site per year, which is similar to that of other coronaviruses. Our findings add to the growing knowledge of the epidemiological and genomic characteristics of SARS-CoV-2 that are important for guiding outbreak containment and vaccine development. The moderate mutation rate of this virus also lends hope that development of an effective, long-lasting vaccine may be possible.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Coronaviruses (CoVs) are a large family of single-stranded RNA viruses that can be isolated from a variety of animals including camels, rats, birds and bats (1). These coronaviruses can cause a range of disease states in animals including respiratory, enteric, hepatic and neurological disease (2). Before late 2019, there were six known CoVs capable of infecting humans (Hu-CoVs). The first four Hu-CoVs that cause mild disease are HKU1, NL63, OC43 and 229E and are known to circulate in the human population (3). The other two Hu-CoVs, known as severe acute respiratory syndrome-CoV (SARS-CoV) and middle east respiratory syndrome-CoV (MERS-CoV), caused two previous epidemics in 2003 (4) and 2012 (5) respectively. Both SARS-CoV and MERS-CoV were the results of recent spillover events from animals. These two epidemics highlighted how easy it is for recombination events in CoVs to occur and cause outbreaks in humans.

In December 2019, another spillover event occurred and a seventh Hu-CoV appeared known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), previously named 2019-nCoV (6). SARS-CoV-2 has been spreading rapidly across the world since it was first reported in Wuhan, Hubei province, China (6, 7). The advances and accessibility of sequencing technologies have allowed researchers all over the world to quickly sequence the genome of SARS-CoV-2 (8, 9). Zhou *et al.* 2020 showed that SARS-CoV-2 shared 79.6% sequence identity to SARS-CoV and 96% sequence identity to a bat CoV further supporting the theory of another spillover event (8).

Further analysis of SARS-CoV-2 genomes proposed that there are two major genotypes, known as L type and S type, based on almost complete linkage between two SNPs (10). Tang *et*

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

al. 2020 proposed that the L type may be more aggressive in replication rates and spreads more quickly, and that human intervention efforts in China may have put selective pressure on L genotypes of SARS-CoV-2 (10). The authors showed that the L type was dominant (96.3%) in Wuhan but significantly less prevalent (61.6%) outside of Wuhan, whereas the S type increased from only 3.7% in Wuhan to 38.4% outside of Wuhan. The same phenomenon is also observed temporally when comparing the S:L type distribution in the early phase of the outbreak (before January 7, 2020) versus after January 7, 2020. In order to test this hypothesis, more work needs to done that combines not only epidemiologic data but also in-depth genomic analysis of these patient samples.

Here we present a comprehensive epidemiological and genomic analysis of SARS-CoV-2 genomes from 10 patients in Shaoxing, a mid-sized city about 500 miles away from Wuhan. This work is one of the first to estimate the mutation rates of SARS-CoV-2 which is a critical component to understanding the sequence evolution of a virus, and an essential step to developing future vaccines and therapies.

MATERIALS AND METHODS

Study design and Ethics

Ten remnant nasopharyngeal swab samples collected between 1/27/2020 and 2/7/2020, and tested positive by a SARS-CoV-2 real-time PCR assay with cycle threshold (C_t) values of less than 28, were included in this study. The samples were de-identified except the associated epidemiological data were retained. The institutional review boards (IRB) approval was waived for this project by Shaoxing Center for Disease Control and Prevention.

SARS-CoV-2 PCR & RNA Sequencing

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Total nucleic acid was extracted from the nasopharyngeal swabs using the Total Nucleic Acid Extraction Kit (IngeniGen XMK Biotechnologies, Inc. Zhejiang, China). Real-time PCR was performed by using the IngeniGen XMKbio 2019-nCoV (SARS-CoV-2) RNA Detection kit, which targets the highly specific sequences in the *ORF1ab* and *N* genes of the virus, on the ABI 7500 system (ThermoFisher Scientific, Inc. MA, USA). The RNA libraries were constructed using the Ingenigen XMKbio RNA-seq Library Prep Kit (IngeniGen XMK Biotechnologies, Inc. Zhejiang, China). Briefly, DNase was used to remove residual human DNA and the RNA was fragmented, followed by double-strand cDNA synthesis, end-repair, dA-tailing and adapter ligation. Sequencing was performed by using the 2X75bp protocol on the Nextseq 550 system (Illumina, Inc. CA, USA).

Data Analysis

Quality control and trimming of paired-end reads was performed using custom Python scripts as follows: 1) trim 3' adapters; 2) trim reads at ambiguous bases; 3) filter reads shorter than 40bp; 4) filter reads with average quality score < 20. Host-derived reads were removed by alignment against the GRCh38.p13 genome reference using bowtie2 (v2.3.4.3) with default parameters. The retained reads were then mapped to 163 published SARS-CoV-2 reference genomes obtained from GISAID (https://www.gisaid.org/CoV2020/, accessed March 2, 2020) by bowtie2 (v2.3.4.3) with default parameters. snippy (v4.5.0) was used for variant calling and core SNP alignment against the Wuhan-Hu-1 reference, FastTree (v2.1.3) was used for tree construction using default parameters, and Figtree (v1.4.4) was used to visualize the resulting phylogenetic tree. Additional statistical analyses and visualizations were performed using custom Python scripts with the pandas (v0.25.0) and matplotlib (v3.1.1) modules. The S/L type was determined according to methods previously described (10). Briefly, C at position 8782 and T at

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

position 28144 was determined to be L type, and T at position 8782 and C at position 28144 was determined to be S type.

RESULTS

Epidemiology of Shaoxing Patients

All ten patients presented with symptoms consistent with COVID-19 in late January and early February of 2020. The majority of patients were male (60%) and the average age was 44 (**Table 1**). The patients can be categorized into two epidemiologic groups with either a travel history to the Hubei province or contact with a confirmed case (**Table 1**). There was one case where we were unable to obtain a travel or social history (Shaoxing-8).

Table 1. Epidemiologic Data of the 10 Shaoxing Patients

ID	Sex	Age Range	History of Travel or Sick Contact	Date of Symptom Onset	Date of Sample Collection
Shaoxing-01	F	30-39	Family members traveled together to Hubei province (1/15 – 1/24)	1/24/20	1/27/20
Shaoxing-02	M	70-79		1/29/20	1/30/20
Shaoxing-03	F	60-66		1/28/20	1/30/20
Shaoxing-04	M	50-59		1/29/20	1/31/20
Shaoxing-05	F	50-59	Traveled to Hubei (1/16 – 1/23)	1/29/20	1/31/20
Shaoxing-06	M	39-39	Resident of Wuhan; Traveled to Shaoxing on 1/17	1/29/20	1/31/20
Shaoxing-07	M	<18	Traveled to Hubei (1/11 – 1/24); Two family members were confirmed cases	1/30/20	1/30/20
Shaoxing-08	M	50-59	Unknown	1/31/20	2/7/20
Shaoxing-09	F	30-39	Family members living together no travel history; Contact with a confirmed case from Ningbo, Zhejiang on 1/27	2/2/20	2/5/20
Shaoxing-10	M	30-39		2/5/20	2/6/20

There are two apparent clusters in these ten patients. The first cluster involves four patients that are relatives who traveled together to the Hubei province in the late January. The first patient in this cluster had symptom onset on their last day in Hubei province while the other

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

three patients had symptom onset 4-5 days after their trip (**Table 1**). The second cluster involves two patients who are family members that live together and did not travel to the Hubei province. Patient Shaoxing-09 developed symptoms a few days before Shaoxing-10, and Shaoxing-09 had a sick contact with a confirmed case from Ningbo, a more populated city in the same province (Zhejiang) as Shaoxing (**Table 1**).

Metagenomic Sequencing

The patients were confirmed to have SARS-CoV-2 infection by a commercial Real-time PCR assay. The average C_t values for the 10 patient samples were 23.17 for *ORF1ab* and 24.54 for *N* (**Table 2**). Metagenomic sequencing was performed to recover the full viral genome. The total number of sequence reads for the samples ranged from 10.4 million to 27.5 million reads with an average of 17.1 million reads. A small percentage of these reads mapped to SARS-CoV-2 RNA (**Table 2**). The range of sequence reads that mapped to SARS-CoV-2 RNA was 2,413 reads to 163,158 reads with an average of 49,066 reads. We observed a clear negative correlation between the C_t values of each gene (*ORF1ab* and *N*) and the log value of SARS-CoV-2 RNA reads (**S1 Fig**). However, the linearity is not significant (R^2=0.6628, 0.5595 for *ORF1ab* and *N*, respectively), indicating that the number of RNA reads measured by metagenomics sequencing are only semi-quantitative and cannot be interpreted directly as viral loads.

Table 2. Summary of Sequencing Results of 10 Shaoxing Patient Samples.

ID	Ct Value (ORF1ab)	Ct Value (N)	Total Sequencing Reads (PE 75)	2019-nCoV RNA (Raw Reads)	2019-nCoV RNA (Log Value)	Genome Coverage (%)	Mean Depth (X)
Shaoxing-01	21.57	23.62	17,158,277	40,057	4.60	99.4	219
Shaoxing-02	18.86	20.93	13,602,710	149,682	5.18	99.9	929
Shaoxing-03	20.09	22.25	24,769,343	163,158	5.21	99.8	1024
Shaoxing-04	24.02	24.68	21,509,477	15,424	4.19	100.0	81

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Shaoxing-05	21.81	23.81	14,043,326	99,521	5.00	99.9	591
Shaoxing-06	25.88	27.08	18,909,299	3,535	3.55	99.9	18
Shaoxing-07	23.11	23.87	10,480,051	5,063	3.70	99.8	26
Shaoxing-08	23.34	24.53	11,506,909	2,413	3.38	99.9	12
Shaoxing-09	26.81	27.85	27,517,291	8,897	3.95	99.9	47
Shaoxing-10	26.24	26.8	11,071,595	2,911	3.46	99.7	15
Min	18.86	20.93	10,480,051	2,413	3.38	99.4	12
Max	26.81	27.85	27,517,291	163,158	5.21	100.0	1024
Mean	23.17	24.54	17,056,828	49,066	4.22	99.8	296

With a large variation in the SARS-CoV-2 RNA mapped reads, we were still able to obtain excellent coverage and depth when each genome was mapped to the first SARS-CoV-2 genome, Wuhan-Hu-1 (6) (**Fig 1A**). The coverage for all genomes was above 99% and the mean depth for the genomes ranged from 12X to 1024X (**Table 2, Fig 1B**). Genomes sequenced to a relatively low mean depth (12X to 47X) were still able to be genotyped successfully (see Results below) but our results suggest that SARS-CoV-2 read counts of at least 15,000 yield sufficiently high depth to characterize even low prevalence or rare mutations.

Fig 1. Coverage and Depth

(A) Coverage and Depth Map. The coverage and depth at each base are depicted by the dark red shading along the circle. (**B) Depth Ratio.** The X-axis plots the log value of the depth for each genome while the Y-axis plots the cumulative percentage of bases covered to the specified depth.

<u>Mutation Rate</u>

To determine the single nucleotide polymorphisms (SNPs) of SARS-CoV-2 in these 10 patients, we mapped each genome to the original Wuhan-Hu-1 reference which was collected on

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

December 31, 2019 (6). The genomes contained a fairly moderate number of SNPs (mean of 4 SNPs, range 1-9) (**Table 3**), consistent with previous reports of relatively low mutation rates (11). The genomes with the largest number of SNPs came from individuals who had contact with a confirmed case from Ningbo, Zhejiang and no travel history to the Hubei province (**Table 3, Shaoxing-9 and 10**).

Table 3. Summary of Genomic Descriptions for the Shaoxing SARS-CoV-2 Genomes

ID	Haplotype	No. of SNP[a]	No. of Days[b]	Mutation Rate (#SNP/day)	Mutation Rate (#SNP/day/nt)	Mutation Rate (#SNP/yr/nt)
Shaoxing-01	S	2	27	0.07	2.48E-06	9.04E-04
Shaoxing-02	S	2	30	0.07	2.23E-06	8.14E-04
Shaoxing-03	S	2	30	0.07	2.23E-06	8.14E-04
Shaoxing-04	S	2	31	0.06	2.16E-06	7.87E-04
Shaoxing-05	L	2	31	0.06	2.16E-06	7.87E-04
Shaoxing-06	S	3	31	0.10	3.24E-06	1.18E-03
Shaoxing-07	L	5	30	0.17	5.57E-06	2.03E-03
Shaoxing-08	L	1	38	0.03	8.80E-07	3.21E-04
Shaoxing-09	S	9	36	0.25	8.36E-06	3.05E-03
Shaoxing-10	S	9	37	0.24	8.13E-06	2.97E-03
Min		1	27	0.03	8.80E-07	3.21E-04
Max		9	38	0.25	8.36E-06	3.05E-03
Mean		4	32	0.11	3.74E-06	1.37E-03

[a] SNP calculated by mapping each genome to the genome of Wuhan-Hu-1 (6)

[b] Number of days between the date that the sample was collected and the date the Wuhan-Hu-1 genome was published (12/31/2019)

Using the SNP analysis, we calculated the various mutation rates using the number of days between the date that the sample was collected and the date the Wuhan-Hu-1 sample was collected. The mutation rate (SNP per day) ranged from 0.03 to 0.25 (**Table 3**). We used this

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

mutation rate to calculate the nucleotide substitution per site per day and the nucleotide substitution per site per year. We saw an average mutation rate of 3.74×10^{-6} nucleotide substitution per site per day and an average mutation rate of 1.37×10^{-3} nucleotide substitution per site per year (**Table 3**).

We looked into deeper into each SNP to determine if there were any non-synonymous mutations in genes important to the virus lifecycle (**Table 4**). No non-synonymous mutations were found in the S gene, which encodes the spike protein that's critical for viral binding to human receptor ACE2 (8). Notably in the cluster where the two family members lived together, the two viruses are closely related but not identical, suggesting a sequential transmission between them. Shaoxing-9 was infected first and then transmitted to Shaoxing-10, whose virus gained a non-synonymous mutation (**Table 4**).

Table 4. Summary of SNPs in the Ten SARS-CoV-2 Genomes

SNP#	Position	Gene	Reference nt	Shaoxing-01	Shaoxing-02	Shaoxing-03	Shaoxing-04	Shaoxing-05	Shaoxing-06	Shaoxing-07	Shaoxing-08	Shaoxing-09	Shaoxing-10
1	207	non-coding	C									T (non-coding)	
2	889	orf1ab	T							C (A)			
3	946	orf1ab	T									C (G)	C (G)
4	5099	orf1ab	T									A (S->T)	A (S->T)
5	7420	orf1ab	C									T (I)	T (I)
6	8344	orf1ab	C								T (D)		
7	8782	orf1ab	C	T (S)	T (S)	T (S)	T (S)		T (S)			T (S)	T (S)
8	9962	orf1ab	C										T (H->Y)
9	11430	orf1ab	A									G (Y->C)	G (Y->C)
10	11916	orf1ab	C							T (S->L)			
11	15324	orf1ab	C					T (N)					
12	21676	S	C									T (Y)	T (Y)
13	22081	S	G							A (Q)			
14	25672	ORF3a	C							A (L->I)			
15	28000	ORF8	C							T (P->L)			
16	28144	ORF8	T	C (L->S)	C (L->S)	C (L->S)	C (L->S)		C (L->S)			C (L->S)	C (L->S)
17	29095	N	C						T (F)				
18	29303	N	C					T (P->S)					
19	29625	ORF10	C									T (S->F)	T (S->F)

L Type | S Type | Non- Synonymous | Synonymous

SNP analysis was based on NC_045512.2 (Wuhan-Hu-1) as the reference genome (6). Variants are denoted as nucleotides versus the reference base. Amino acid changes listed in parentheses with synonymous mutations listed as a single residue. The non-synonymous mutations are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

SARS-CoV-2 Genotype and Phylogenetic Characteristics

Previous reports demonstrate that SARS-CoV-2 has two genotypes known as L type and S type (10). The majority of the Shaoxing patients in this study have the S type (70%). We decided to compare our ten SARS-CoV-2 genomes to 163 other published SARS-CoV-2 genomes obtained from GISAID (9). The majority of the SARS-CoV-2 genomes obtained from GISAID are the L type (**Fig 2, red**). The Shaoxing SARS-CoV-2 genomes are distributed throughout the other genomes (**Fig 2, green dots**).

Fig 2 Phylogenetic Comparison of SARS-CoV-2 Genomes.

Phylogenetic comparison of 162 published genomes from GISAID (9) and the ten Shaoxing genomes (green dots). Genomes are color-coded based on their haplotype (Red=L type, Blue=S type).

Interestingly, four of the Shaoxing SARS-CoV-2 genomes were identical to six other GISAID SARS-CoV-2 genomes (**Fig 2, Cluster 1**). These six other genomes were isolated from patients all over the world: two from Wuhan, two from Taiwan, one from Belgium and one from Australia (**Fig 2, Cluster 1**). The cluster with family members who lived together (Shaoxing-9 & -10) is closely related to another cluster found in England (**Fig 2, Cluster 2**). Shaoxing-6 is identical to five other genomes isolated in Shenzhen, Guangdong Province, China (**Fig 2, Cluster 3**).

The three Shaoxing L type genomes are phylogenetically distributed throughout the other L type genomes (**Fig 2**). Shaoxing-5 is closed related to two other genomes from China (**Fig 2,** Guangzhou_20SF206_2020 and China_IQTC01_2020), while Shaoxing-7 is closed related to a

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

genome isolated from Singapore (**Fig 2,** Singapore_1_2020). Shaoxing-8 is not closely related to any other L type genome (**Fig 2**).

DISCUSSION

In this study, we sequenced the SARS-CoV-2 genome from ten patient samples from Shaoxing, Zhejiang, China. Using metagenomic sequencing, we were able to obtain above 99% coverage and an average depth of 296X for all 10 SARS-CoV-2 genomes. Although not statistically significant, there does appear to be a clear negative correlation between the C_t values of both gene targets and the log count of SARS-CoV-2 RNA sequence reads acquired by metagenomics sequencing. This suggests that the log value of RNA sequence reads by metagenomics sequencing may be used as a semi-quantitative, but not accurate, measurement for SARS-CoV-2 viral loads.

The rapid spread of this virus is highlighted by the fact that four SARS-CoV-2 genomes from Shaoxing individuals were identical to six other SARS-CoV-2 genomes from patients all over the world. We were unable to obtain epidemiologic data from the other six SARS-CoV-2 genomes but it would be interesting to see if these patients shared the same contact or not.

Overall, we did not see a large number of SNPs in these SARS-CoV-2 genomes. The greatest number of SNPs seen was 9 and these two SARS-CoV-2 genomes were from individuals with no travel history to Hubei province (**Table 3, Shaoxing-9 and 10**). Instead, Shaoxing-9 and 10 had contact with a confirmed case from Ningbo, another city outside of Hubei. We can use these data to infer that the virus accumulated more mutations when it was spread to another city outside of Hubei first before coming to Shaoxing, compared to the virus spread to Shaoxing directly from Hubei.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

We combined epidemiologic data with the SNP analysis to estimate the mutation rate of the SARS-CoV-2 from these ten patients. We saw an average mutation rate of 1.37×10^{-3} nucleotide substitution per site per year for SARS-CoV-2, which is similar to SARS-CoV-1 with a reported mutation rate of 0.80-2.38 $\times 10^{-3}$ nucleotide substitution per site per year (12). These data demonstrate that SARS-CoV-2 is similar in the mutation rate as other coronaviruses.

Our data support the hypothesis put forward by Tang et al. 2020, which states that human intervention efforts in China may have put selective pressure on both the S and L genotypes of SARS-CoV-2 (10). The less aggressive form of the SARS-CoV-2 (S type) was allowed to increase in prevalence due to relatively weaker selective pressure. Although a small sample size, 70% (7/10) of our patients were infected with the S type and the majority of which (71%, 5/7) traveled to or from Wuhan within 14 days of symptom onset. The dynamics of S and L genotype distribution may have a role in assessment of the severity of the outbreak as it is still rampaging the world as we write this manuscript. Our study adds the growing body of evidence that the mutation rate of SARS-CoV-2 is not any different from other coronavirus, which is important for vaccine development (11, 13).

The major limitation of this study is that we only had 10 samples analyzed due to the requirement of sufficient SARS-CoV-2 RNA from a metagenomic sample. However, with the development of a SARS-CoV-2 probe enrichment kit, this type of deep sequencing analysis may be applied to samples with lower viral loads, thereby enabling more complete molecular epidemiological surveillance. In addition, the C_t value cut-off of 28 established in this study may not be directly applicable to other real-time PCR assays due to the technical differences.

In summary, we showed that a full viral genomic analysis is feasible via metagenomics sequencing on nasopharyngeal samples with SARS-CoV-2 Real-time PCR C_t values less than 28.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Our analysis demonstrated that the virus spread extremely quickly around the globe as early as late January with few mutations. The mutation rate of the virus is similar to that of other coronaviruses, lending hope that development of an effective, long-lasting vaccine may be possible.

ACKNOWLEDGEMENTS

We would like to thank Yong-Zhen Zhang (Fudan University) and Eddie Holmes (University of Sydney) for sharing the sequence of the first SARS-CoV-2 isolate in a very timely manner. We would also like to thank Fanchao Meng, Bin Hu, Haihao Shou and Yuanyuan Cai from Shaoxing IngeniGen XMK Biotechnologies, Inc. for their technical assistance. This study is funded by Shaoxing IngeniGen XMK Biotechnologies, Inc.

References

1. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, Evaluation and Treatment Coronavirus (COVID-19). StatPearls. Treasure Island (FL): StatPearls Publishing StatPearls Publishing LLC.; 2020.

2. Weiss SR, Leibowitz JL. Coronavirus pathogenesis. Adv Virus Res. 2011;81:85-164.

3. Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and Sources of Endemic Human Coronaviruses. Adv Virus Res. 2018;100:163-88.

4. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science. 2003;300(5624):1394-9.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

5. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. The New England journal of medicine. 2012;367(19):1814-20.

6. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.

7. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet (London, England). 2020;395(10223):470-3.

8. 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;579(7798):270-3.

9. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22(13).

10. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020.

11. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020.

12. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol. 2004;4:21.

13. Li X, Wang W, Zhao X, Zai J, Zhao Q, Li Y, et al. Transmission dynamics and evolutionary history of 2019-nCoV. J Med Virol. 2020;92(5):501-11.

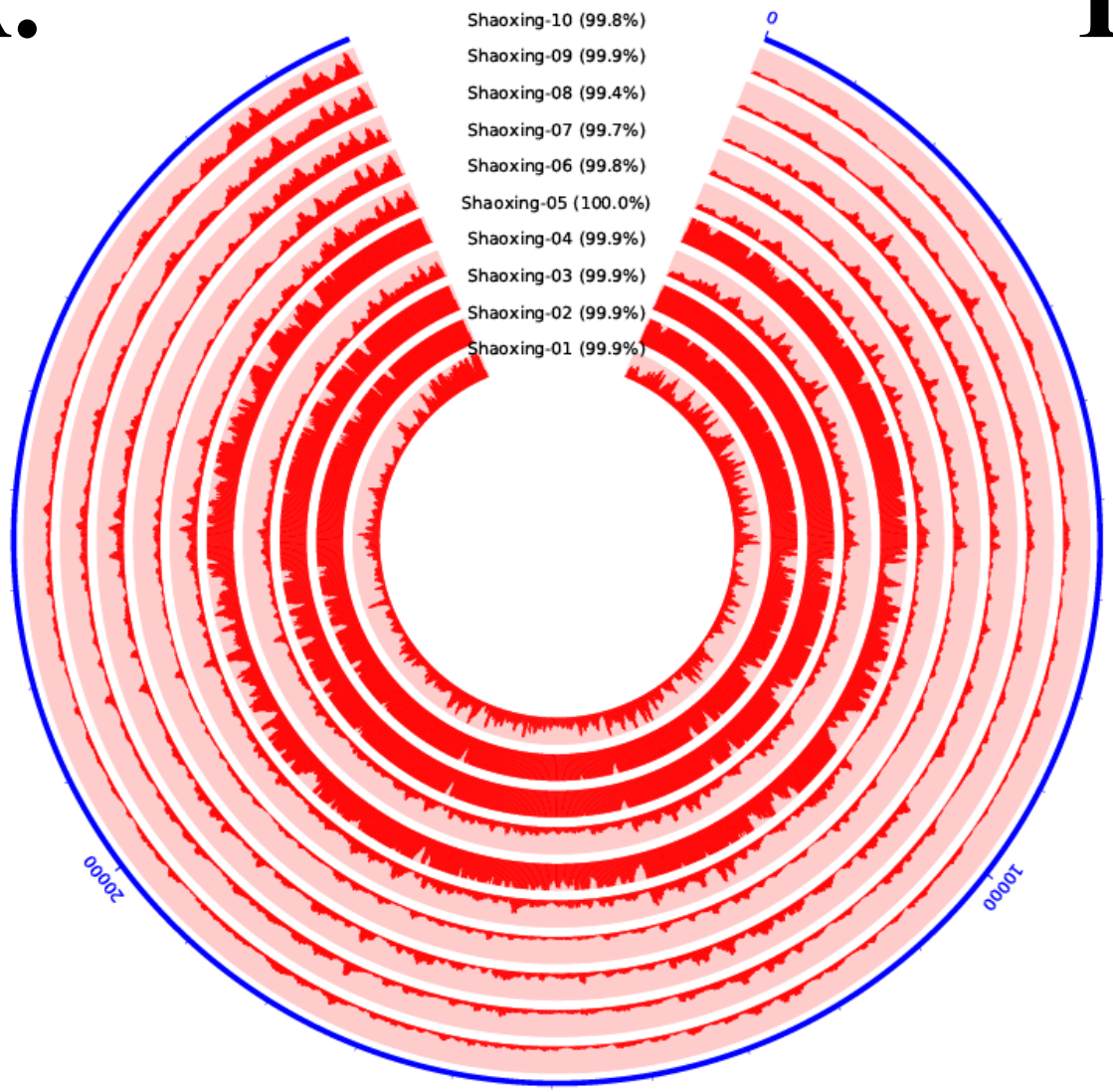
medRxiv preprint doi: https://doi.org/10.1101/2020.04.16.20058560; this version posted April 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY INFORMATION CAPTIONS

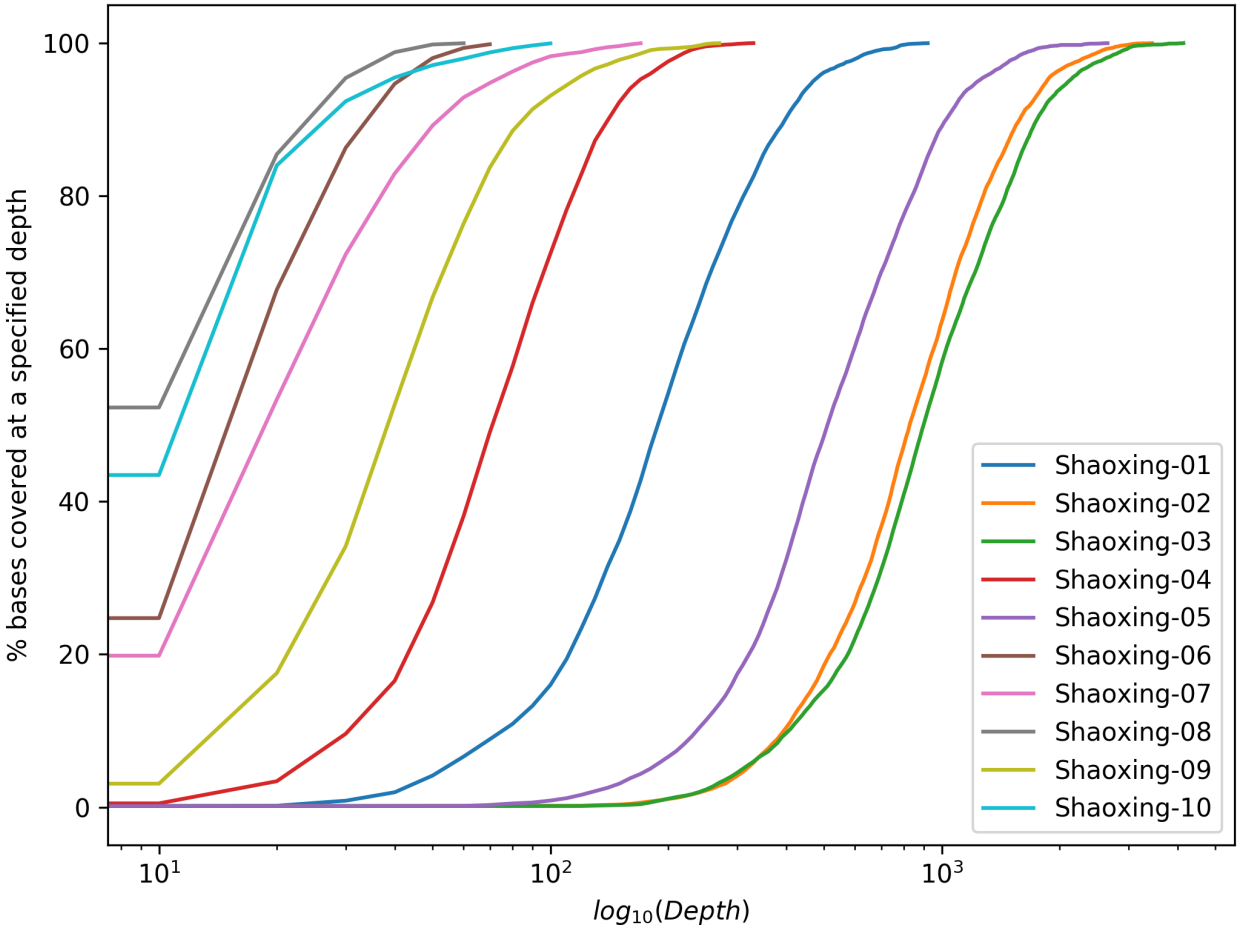
S1 Fig Correlation of C_t Values and SARS-CoV-2 RNA Reads.

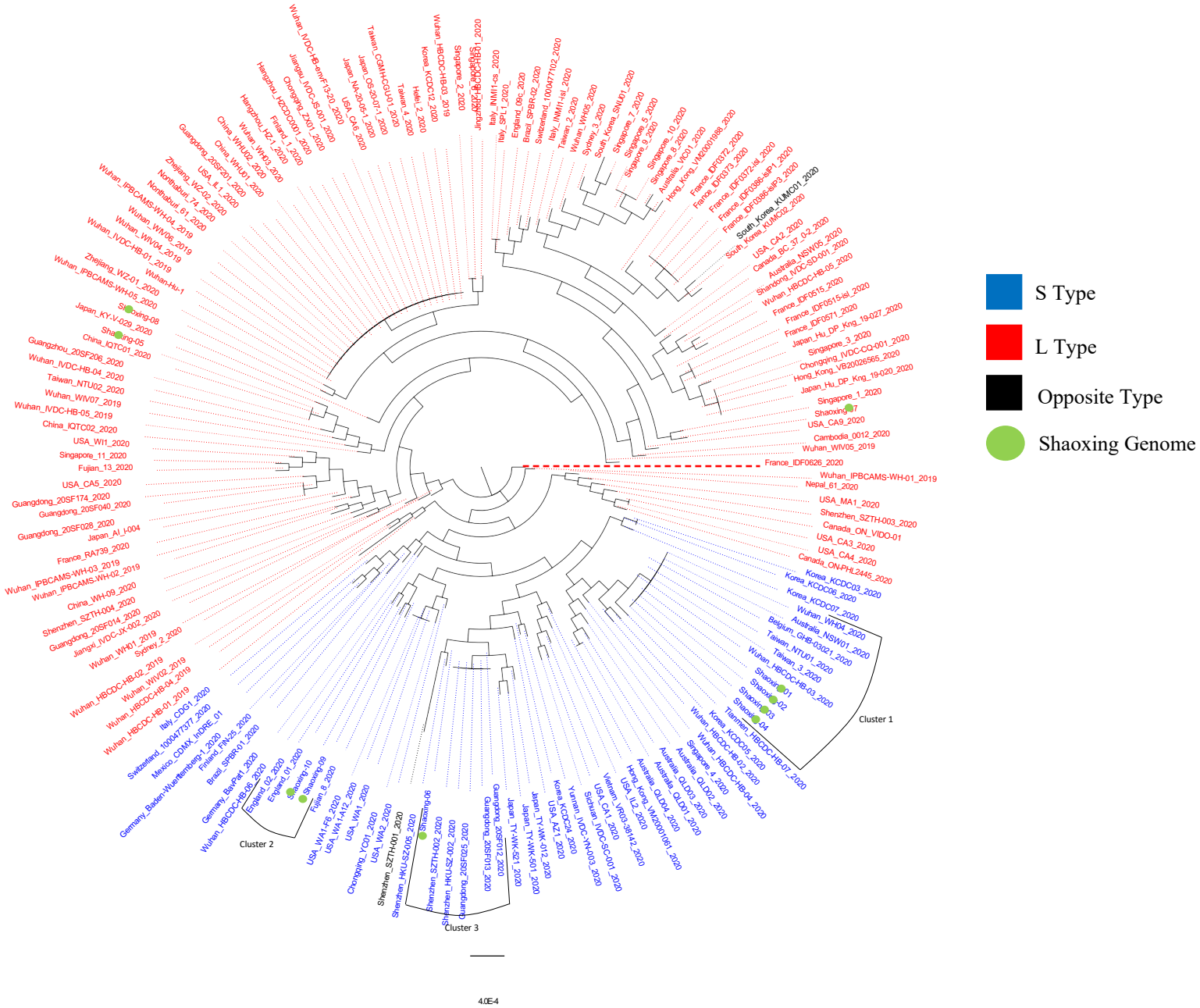
(A) *ORF1ab* Correlation. The X-axis plots the log value of the SARS-CoV-2 RNA reads while the Y-axis plots the C_t values for the *ORF1ab* gene for the ten Shaoxing patients. (**B) *N* Correlation.** The X-axis plots the log value of the SARS-CoV-2 RNA reads while the Y-axis plots the C_t values for the *N* gene for the ten Shaoxing patients.

A.


Shaoxing-10 (99.8%)
Shaoxing-09 (99.9%)
Shaoxing-08 (99.4%)
Shaoxing-07 (99.7%)
Shaoxing-06 (99.8%)
Shaoxing-05 (100.0%)
Shaoxing-04 (99.9%)
Shaoxing-03 (99.9%)
Shaoxing-02 (99.9%)
Shaoxing-01 (99.9%)
0
10000
20000

B.


% bases covered at a specified depth
100
80
60
40
20
0
10^1
10^2
10^3
$log_{10}(Depth)$
Shaoxing-01
Shaoxing-02
Shaoxing-03
Shaoxing-04
Shaoxing-05
Shaoxing-06
Shaoxing-07
Shaoxing-08
Shaoxing-09
Shaoxing-10


S Type
L Type
Opposite Type
Shaoxing Genome
Cluster 1
Cluster 2
Cluster 3
4.0E-4