

Genetic sequencing reveals natural origin, early spread and infectomes of SARS- CoV-2 in China

Weifeng Shi

School of Public Health & Key Laboratory of Etiology and Epidemiology of Emerging
Infectious Diseases in Universities of Shandong,
Shandong First Medical University & Shandong Academy of Medical Sciences

October 20th, 2020

Genetic sequencing reveals natural origin, early spread and infectomes of SARS-CoV-2 in China

1

Virosphere & genetic sequencing

2

Natural origin of SARS-CoV-2

3

Early spread of SARS-CoV-2 in China

4

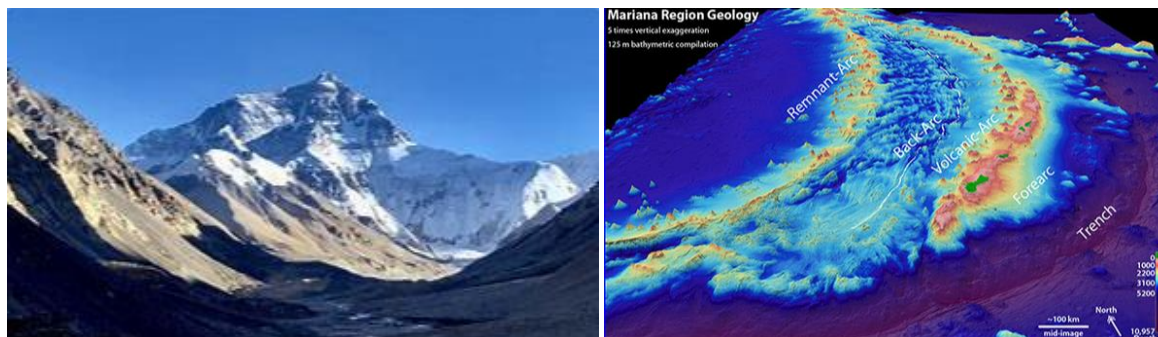
Infectomes of SARS-CoV-2 in China

1. Virosphere & genetic sequencing



► Viruses are the most abundant form of life on earth.

- They exist everywhere on earth.
- They infect everything alive.



Mt. Qomolangma

Mariana Trench



Tropical rainforest

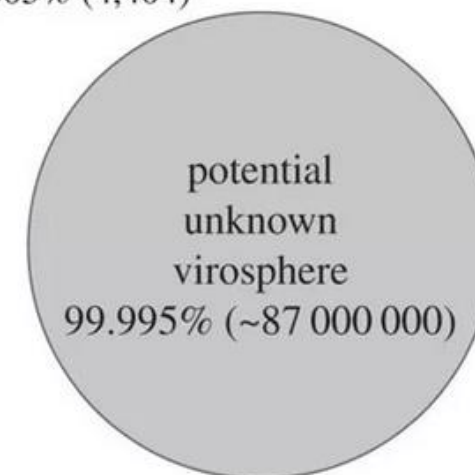
South Pole

known
classified
viruses → 0.005% (4,404)

← human
viruses
0.0003% (219)

► Viruses are 'dark matters' in life sciences.

- It is estimated that there are **millions** of viruses on earth, among which **~0.73 million** could infect mammals and even humans. A few scientists claim that the number of viral species could reach even 87 million.
- So far, only **~9000** viruses have been identified and only **~300** are able to infect humans.



Anthony, S.J., et al. (2013). *MBio* 4(5), e00598-00513. doi: 10.1128/mBio.00598-13.

Geoghegan, J.L. and Holmes, E.C. (2017). *Open Biol* 7(10) doi: 10.1098/rsob.170189.

1. Virosphere & genetic sequencing



Genomic analysis of uncultured marine viral communities

Mya Breitbart*, Peter Salamon¹, Bjarne Andresen^{1,†}, Joseph M. Mahaffy¹, Anca M. Segall*, David Mead³, Farooq Azam³, and Forest Rohwer*¹

*Department of Biology, San Diego State University, San Diego, CA 92182-4614; ¹Department of Mathematical Sciences, San Diego State University, San Diego, CA 92182-7725; ²Forst Laboratory, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark; ³Lutigen, Middleton, WI 53562; and [†]Marine Biology Division, Scripps Institution of Oceanography, La Jolla, CA 92093

Communicated by Allan Campbell, Stanford University, Stanford, CA, August 14, 2002 (received for review February 22, 2002)

Viruses are the most common biological entities in the oceans by an order of magnitude. However, very little is known about their diversity. Here we report a genomic analysis of two uncultured marine viral communities. Over 65% of the sequences were not significantly similar to previously reported sequences, suggesting that much of the diversity is previously uncharacterized. The most common significant hits among the known sequences were to viruses. The viral hits included sequences from all of the major families of dsDNA tailed phages, as well as some algal viruses. Several independent mathematical models based on the observed number of contigs predicted that the most abundant viral genome comprised 2–3% of the total population in both communities, which was estimated to contain between 374 and 7,114 viral types. Overall, diversity of the viral communities was extremely high. The results also showed that it would be possible to sequence the entire genome of an uncultured marine viral community.

Marine viruses, the majority of which are phages, have enormous influences on global biogeochemical cycles (1), microbial diversity (2, 3), and genetic exchange (4). Despite their importance, virtually nothing is known about marine viral biodiversity or the evolutionary relationships of marine and nonmarine viruses (5–7). Addressing these issues is difficult because viruses must be cultured on hosts, the majority of which cannot be cultivated by using standard techniques (8). In addition, viruses do not have ubiquitously conserved genetic elements such as rDNA that can be used as diversity and evolutionary distance markers (9). To circumvent these limitations, we developed a method to shotgun clone and sequence uncultured aquatic viral communities.

Materials and Methods

Isolation of Viral Community DNA. Marine viruses were isolated from 200 liters of surface seawater from Scripps Pier (SP, La Jolla, CA; May 2001) and the channel side of Fiesta Island in Mission Bay (MB, San Diego; June 2001) by using a combination of differential filtration and density-dependent gradient centrifugation. The water at the MB site is exchanged with each tidal cycle. Both the SP and MB sites experience increased levels of pollution during the rainy season, because of runoff from the surrounding city. The MB site routinely has more eukaryotic algae than does the SP site. Once collected, the water samples were initially filtered through a 0.16- μ m Centrimate tangential flow filter (TFF; Pall) to remove bacteria, eukaryotes, and large particles. Approximately 90% of the viruses, as determined by epifluorescent microscopy (10), and most of the water, passed through the filter and were collected in a separate tank. Subsequently, the viruses in the filtrate were concentrated by using a 100-kDa TFF filter until the final sample volume was <100 ml (~5,000 \times concentration). Recovery of viruses during this step was essentially 100%. After the TFF, the phage concentrate was loaded onto a cesium chloride (CsCl) step gradient, ultracentrifuged, and the 1.35–1.5 g/ml fraction was collected. This fraction contains the majority of the viral DNA as determined

by pulse field gel electrophoresis (11); however, this method will not recover all viruses (e.g., large eukaryotic viruses and ssRNA phages). After CsCl purification, the viruses were lysed by using a formamide extraction, and the DNA was recovered by an isopropanol precipitation and a cetyltrimethylammonium bromide (CTAB) extraction (12).

Construction of the Shotgun Library. The amount of viral DNA in an environmental sample is very low (~10 μ g/100 liters). Viral genomes often contain modified nucleotides that cannot be directly cloned into *Escherichia coli*. Additionally, because viral genomes contain genes (e.g., holins, lysozyme) that must be disrupted before cloning, we have not been able to create a representative cosmid library from these communities. We have circumvented these problems by randomly shearing the total marine viral community DNA (HydroShear, GenMachine, San Carlos, CA), end-repairing, ligating dsDNA linkers to the ends, and amplifying the fragments by using the high-fidelity Vent DNA polymerase. The resulting fragments were ligated into the pSMART vector and electroporated into MC12 cells (Lucigen, Middleton, WI). We call these libraries LASLs for linker-amplified shotgun libraries. This method has been checked to ensure randomness as described (ref. 13, and our web site at www.sci.sdsu.edu/PHAGE/LASL/index.htm). A test library was constructed of coliphage λ DNA, and 100 fragments were sequenced without observing any chimeras. Additionally, we have recently sequenced two phage genomes, Vibriophage 16T and 16C, from a mixed lysate by using the LASL approach. No chimeric molecules were observed in this mixed library. Together, these three libraries represent >1,000 sequences. Therefore, it is highly unlikely that we are observing a significant number of chimeric sequences in our library (F.R. and A.M.S., unpublished results).

Analysis of Sequences: Composition Analyses. Clones from the SP library were sequenced with both forward and reverse primers, yielding a total of 1,061 sequences. Eight hundred and seventy-three clones from the MB library were sequenced only with the forward primer. These sequences were compared against GenBank by using tBLASTX (14, 15). A hit was considered significant if it had an *E* value of <0.001. Significant hits to GenBank entries were classified into the groups described in the text, based on sequence annotation. In cases where multiple significant hits were observed for a single query sequence, the sequence was preferentially classified as a phage or virus if these occurred within the top five hits. Mobile elements consisted of transposons, plasmids, insertion sequences, retrotransposons, unstable genetic elements, and pathogenicity islands. Bacterial hits

Abbreviations: SP, Scripps Pier; MB, Mission Bay; LASL, linker-amplified shotgun library; MM%, minimal mismatch percentage.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AY079522–AY080585 and BH898061–BH898933).

To whom correspondence should be addressed. E-mail: forest@unstroke.sdsu.edu.

In 2002, based on the approach of metagenome, Breitbart and colleagues enriched the viruses from sea water and performed high throughput sequencing, which led to the discovery of many new viruses. To my knowledge, this is the first report with the idea of virome.

1. Virosphere & genetic sequencing



PERSPECTIVES

Global Screening for Human Viral Pathogens

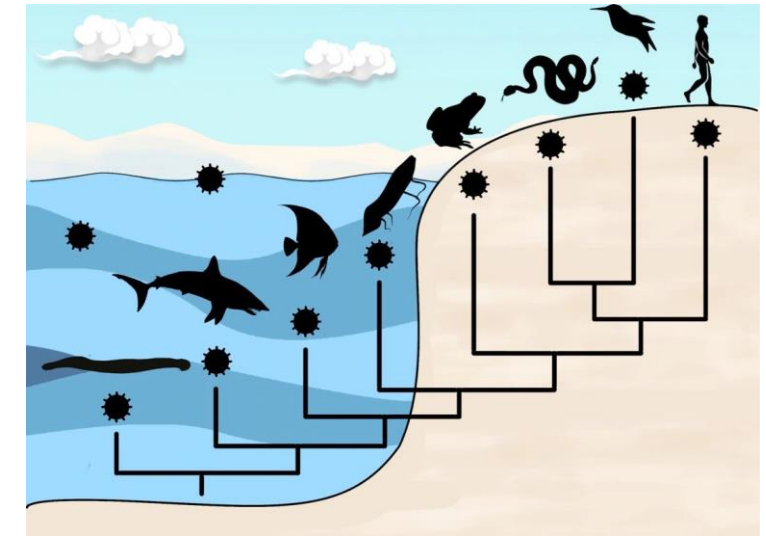
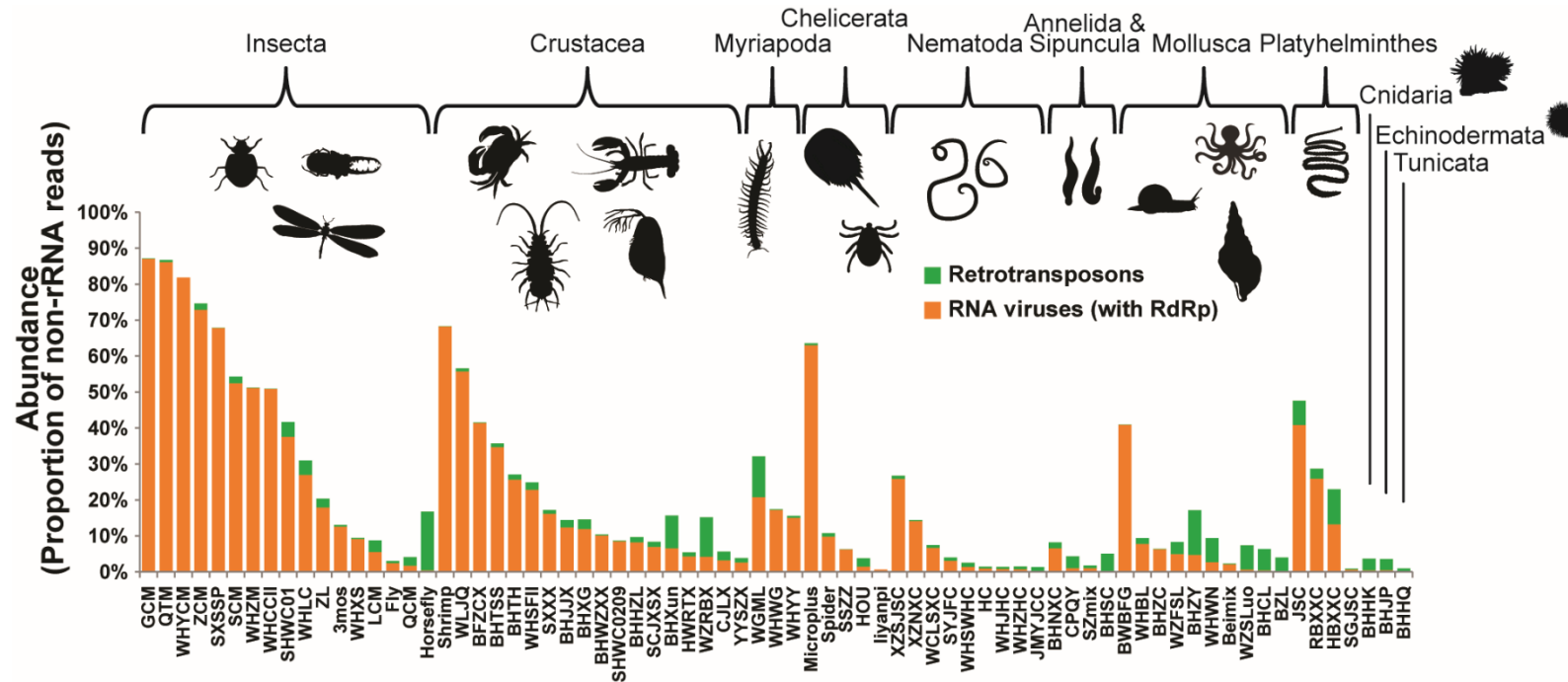
Norman G. Anderson,* John L. Gerin,† and N. Leigh Anderson‡

Anderson, N.G., Gerin, J.L., and Anderson, N.L. (2003) Global screening for human viral pathogens. *Emerg Infect Dis* 9, 768-774.

Virome refers to the collection of nucleic acids, both RNA and DNA, that make up the viral community associated with a particular ecosystem or holobiont. The word is derived from virus and genome and is used to describe viral shotgun metagenomes. All macro-organisms have viromes that include bacteriophage and viruses. Viromes are important in the nutrient and energy cycling, development of immunity, and a major source of genes through lysogenic conversion.

<https://en.wikipedia.org/wiki/Virome>

1. Virosphere & genetic sequencing

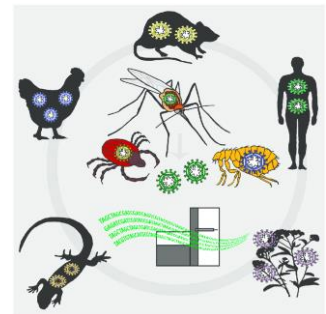


<https://www.sydney.edu.au/news-opinion/news/2018/04/05/ancient-origins-of-viruses-discovered.html>

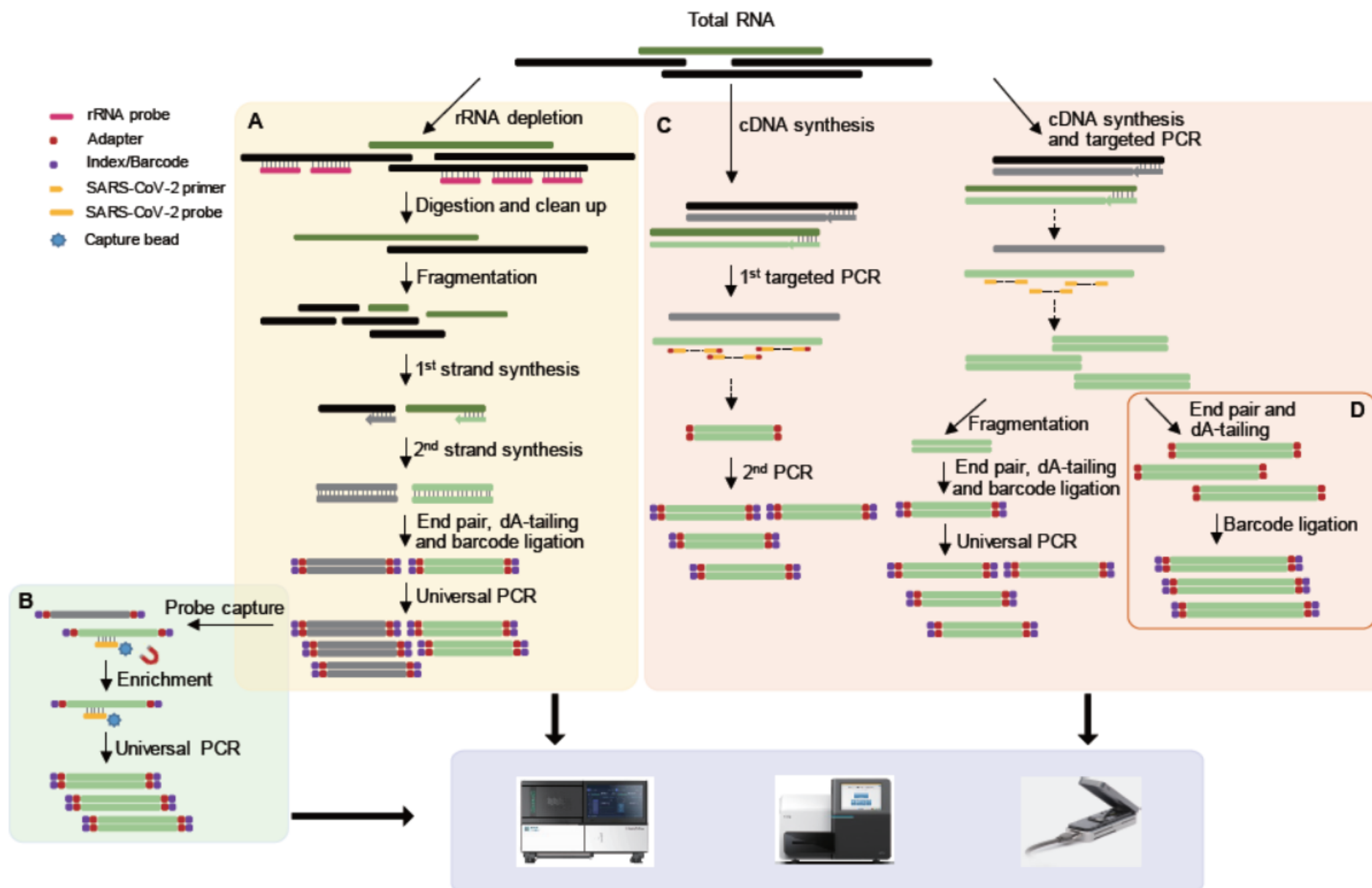
Shi M, et al. *Nature*. 2016;540(7634):539-543. doi:10.1038/nature20167

Virosphere (a term coined by professor Curtis Suttle, University of British Columbia)

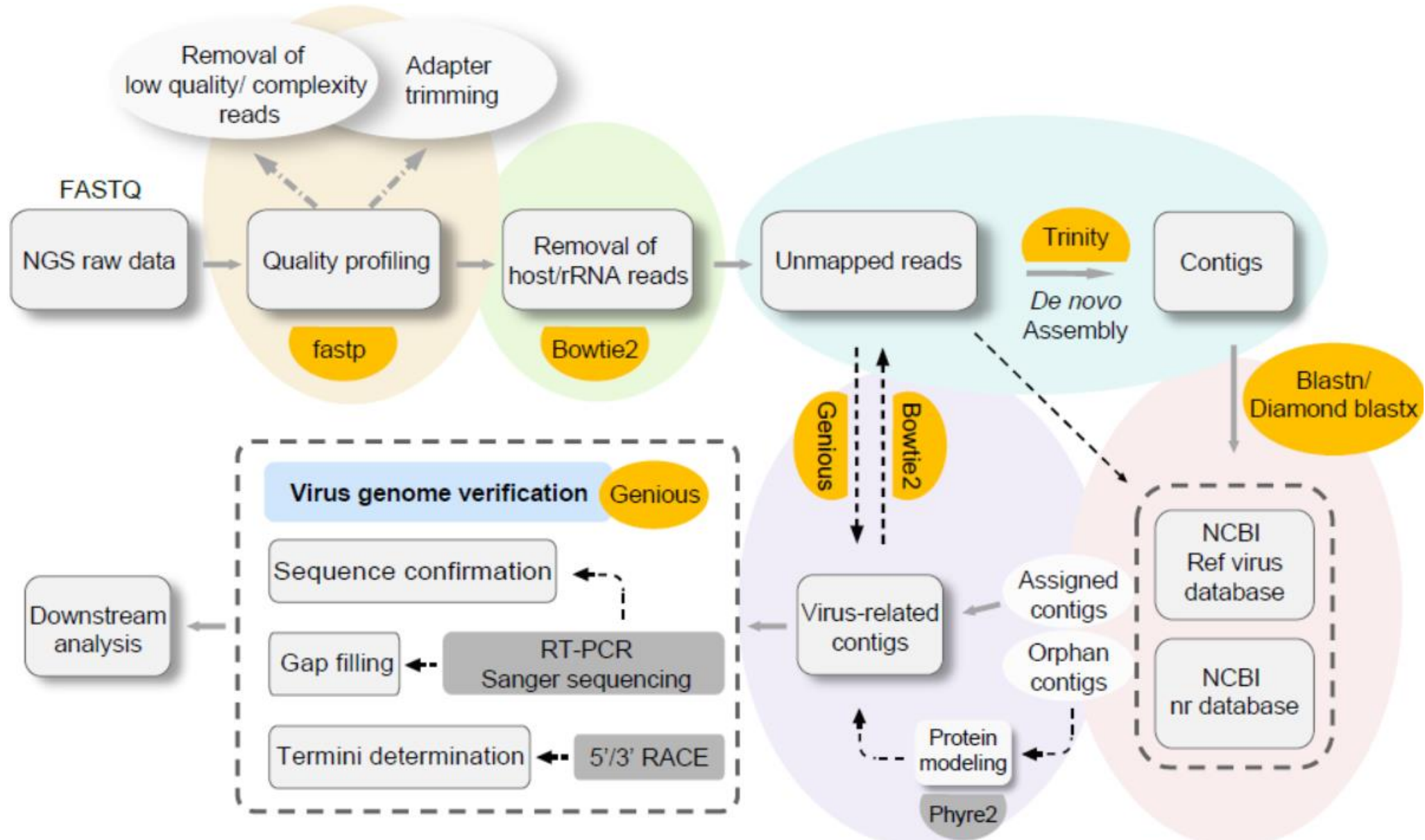
*All those places where viruses are found or in which they interact with their hosts.



1. Virosphere & genetic sequencing



1. Virosphere & genetic sequencing



1. Virosphere & genetic sequencing

Bioinformatics resources for SARS-CoV-2 discovery

Databases and Software	URL	Reference
Data quality control		
FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	-
FastQ Screen	http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/	[38]
TrimGalore	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/	-
Trimmomatic	http://www.usadellab.org/cms/index.php?page=trimmomatic	[39]
Fastx-toolkit	http://hannonlab.cshl.edu/fastx_toolkit/	-
Skewer	https://sourceforge.net/projects/skewer	[40]
BBduk	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/	-
AfterQC	http://www.github.com/OpenGene/AfterQC	[41]
SOAPnuke	https://github.com/BGI-flexlab/SOAPnuke	[42]
Fastp	https://github.com/OpenGene/fastp	[43]
NanoPack	https://github.com/wdecoster/nanopack	[63]
Porechop	https://github.com/rrwick/Porechop	-
Read Mapping		
Hisat2	https://daehwankimlab.github.io/hisat2/	[44]
BWA	http://bio-bwa.sourceforge.net/	[45]
Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml	[46]
STAR	https://github.com/alexdobin/STAR	-
KMA	https://bitbucket.org/genomicepidemiology/kma	[47]
SortmeRNA	http://bioinfo.lifl.fr/RNA/sortmerna	[48]
Minimap2	https://github.com/lh3/minimap2	[64]

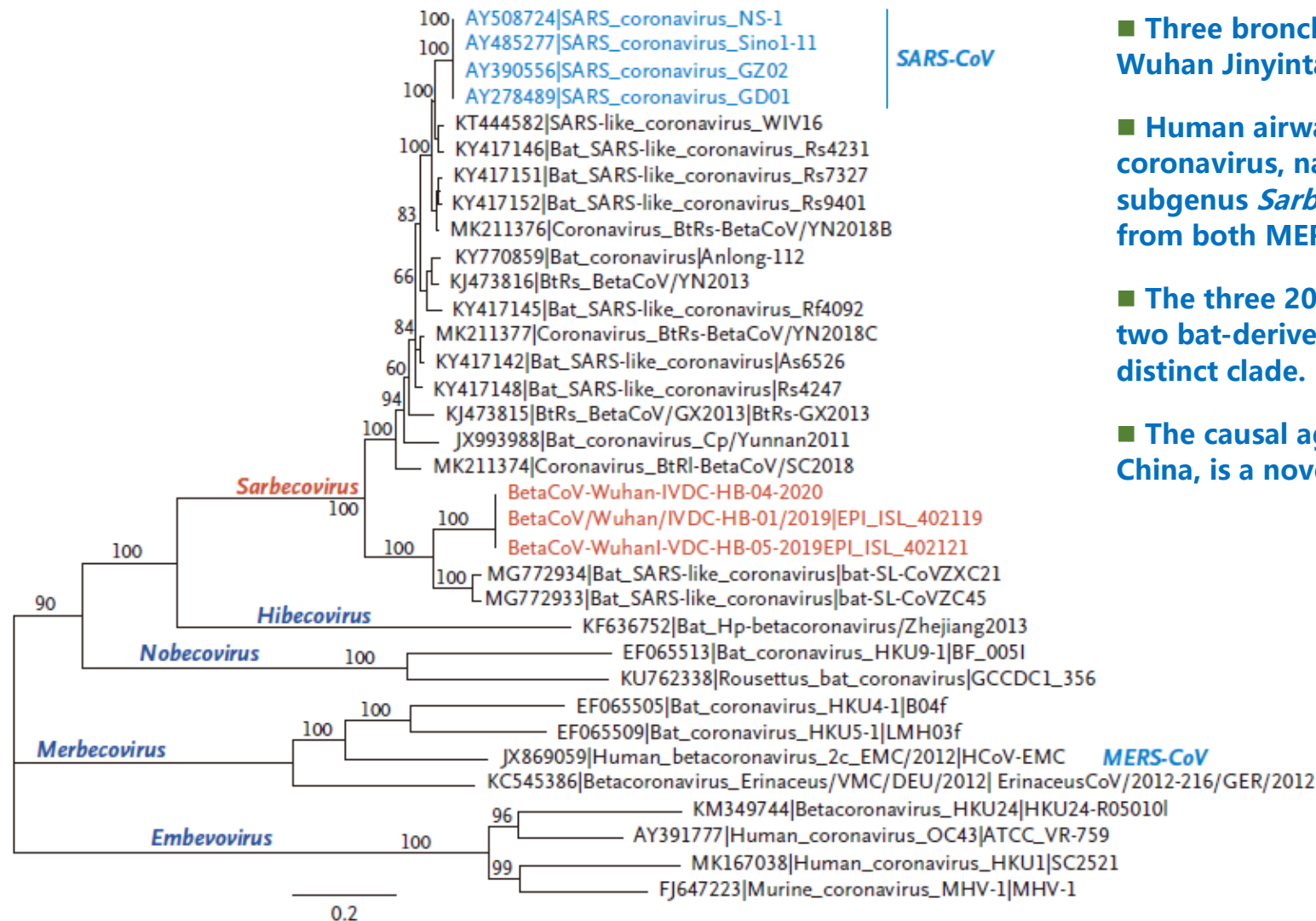
Diamond	https://www.wsi.uni-tuebingen.de/lehrstuehle/algorithms-in-bioinformatics/software/diamond/	[57]
Phyre2	http://www.sbg.bio.ic.ac.uk/phyre2/html	[60]
Canu	https://github.com/marbl/canu	[67]
Falcon	https://github.com/PacificBiosciences/falcon	-
Miniasm	https://github.com/lh3/miniasm	[68]
Genome Visualization		
IGV	http://software.broadinstitute.org/software/igv/	[61]
Geneious	https://www.geneious.com/	-
QUAST	https://sourceforge.net/projects/quast/	[62]
SEQMAN	https://www.dnastar.com/software/molecular-biology/	-
Database		
Global Initiative on Sharing All Influenza Data (GISAID)	https://www.epicov.org/	[133]
National bioinformatics Center (CNCB) / National Genomics Data Center (NGDC) database	https://bigd.big.ac.cn/ncov/	-
National Center for Biotechnology Information (NCBI)	https://www.ncbi.nlm.nih.gov/	-
Genome Warehouse (GWH)	https://bigd.big.ac.cn/gwh/	-
Sequence alignment		
CLUSTALW	https://www.genome.jp/tools-bin/clustalw	[70]
MAFFT	https://mafft.cbrc.jp/alignment/software/	[71]
MUSCLE	http://drive5.com/muscle/	[72]

Hu T, Li J, Zhou H, Li C, Holmes EC, Shi W. *Briefings in Bioinformatics*, under revision.

2. Natural origin of SARS-CoV-2



NGS identified SARS-CoV-2



■ Three bronchoalveolar-lavage samples were collected from Wuhan Jinyintan Hospital on December 30, 2019.

■ Human airway epithelial cells were used to isolate a novel coronavirus, named 2019-nCoV, which formed a clade within the subgenus *Sarbecovirus*, *Orthocoronavirinae* subfamily, different from both MERS-CoV and SARS-CoV.

■ The three 2019-nCoV coronaviruses from Wuhan, together with two bat-derived SARS-like strains, ZC45 and ZXC21, form a distinct clade.

■ The causal agent of an outbreak of severe pneumonia in Wuhan, China, is a novel coronavirus.

The NEW ENGLAND JOURNAL of MEDICINE

BRIEF REPORT

A Novel Coronavirus from Patients with Pneumonia in China, 2019

Na Zhu, Ph.D., Dingyu Zhang, M.D., Wenling Wang, Ph.D., Xingwang Li, M.D., Bo Yang, M.S., Jingdong Song, Ph.D., Xiang Zhao, Ph.D., Baoying Huang, Ph.D., Weifeng Shi, Ph.D., Roujian Lu, M.D., Peihua Niu, Ph.D., Faxian Zhan, Ph.D., Xuejun Ma, Ph.D., Dayan Wang, Ph.D., Wenbo Xu, M.D., Guizhen Wu, M.D., George F. Gao, D.Phil., and Wenjie Tan, M.D., Ph.D., for the China Novel Coronavirus Investigating and Research Team

2. Natural origin of SARS-CoV-2

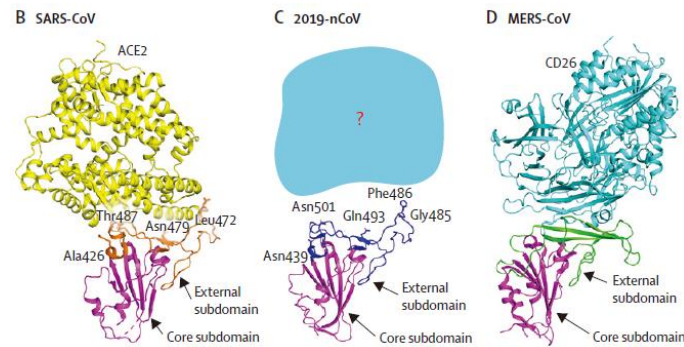
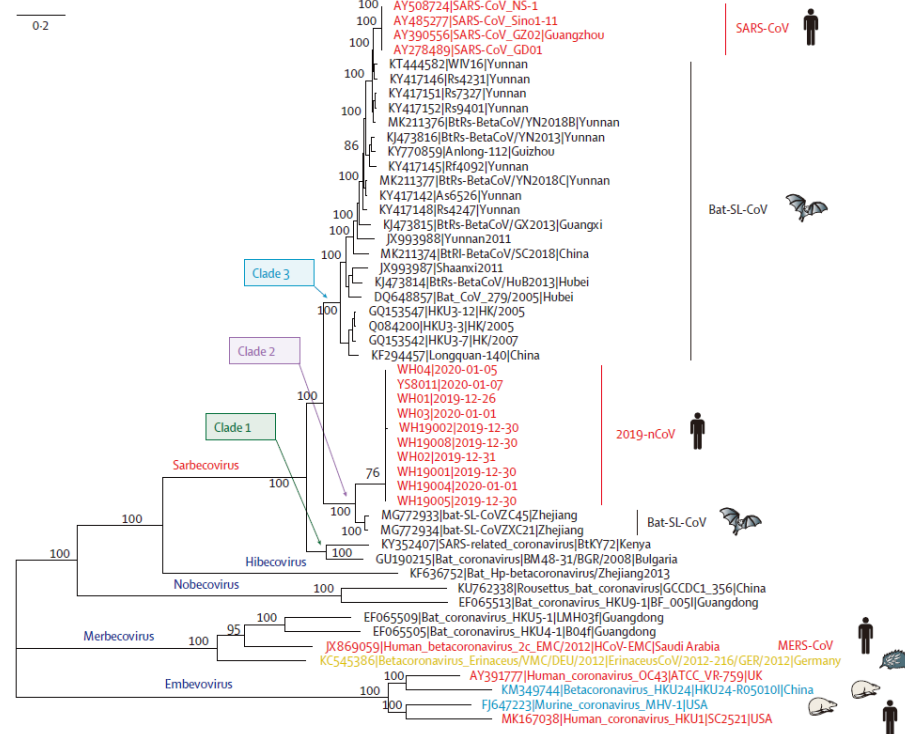
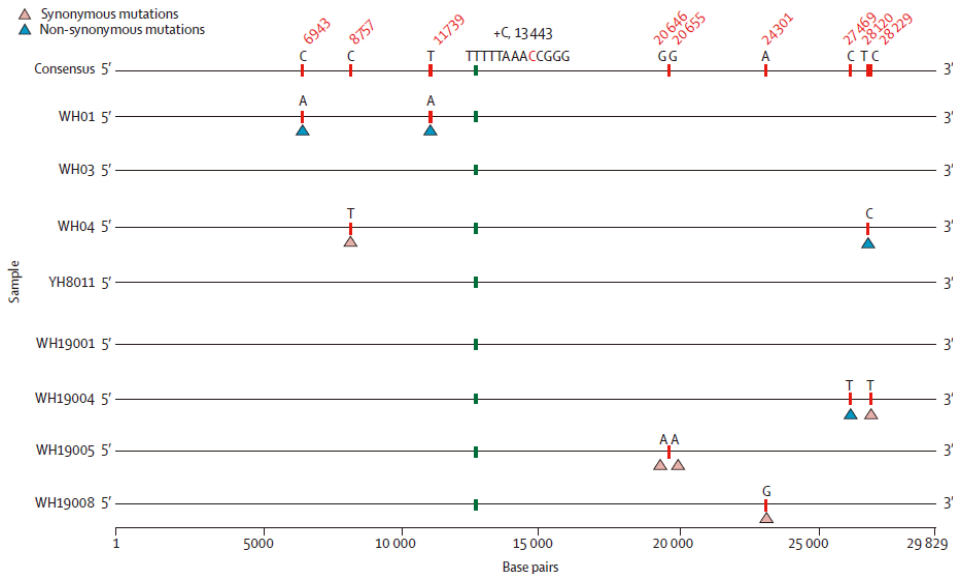


Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding

	Patient information		Sample information		Genome sequence obtained	
	Exposure to Huanan seafood market	Date of symptom onset	Admission date	Sample type	Collection date	Ct value
Samples WH19001 and WH19005	Yes	Dec 23, 2019	Dec 29, 2019	BALF and cultured virus	Dec 30, 2019	30-23
Sample WH19002	Yes	Dec 22, 2019	NA	BALF	Dec 30, 2019	30-50
Sample WH19004	Yes	NA	NA	BALF	Jan 1, 2020	32-14
Sample WH19008	Yes	NA	Dec 29, 2019	BALF	Dec 30, 2019	26-35
Sample YS8011	Yes	NA	NA	Throat swab	Jan 7, 2020	22-85
Sample WH01	Yes	NA	NA	BALF	Dec 26, 2019	32-60
Sample WH02	Yes	NA	NA	BALF	Dec 31, 2019	34-23
Sample WH03	Yes	Dec 26, 2019	NA	BALF	Jan 1, 2020	25-38
Sample WH04	No*	Dec 27, 2019	NA	BALF	Jan 5, 2020	25-23

Ct=threshold cycle. BALF=bronchoalveolar lavage fluid. NA=not available. 2019-nCoV=2019 novel coronavirus. *Patient stayed in a hotel near Huanan seafood market from Dec 23 to Dec 27, 2019, and reported fever on Dec 27, 2019.

Table: Information about samples taken from nine patients infected with 2019-nCoV



- Genome sequences of 2019-nCoV sampled from nine patients who were among the early cases of this severe infection are almost genetically identical, which suggests very recent emergence of this virus in humans and that the outbreak was detected relatively rapidly.

- 2019-nCoV is most closely related to other betacoronaviruses of bat origin, indicating that these animals are the likely reservoir hosts for this emerging viral pathogen.

- Structural analysis suggests that 2019-nCoV had a similar receptor-binding domain structure to that of SARS-CoV, and it might be able to bind to the ACE2 receptor in humans.

2. Natural origin of SARS-CoV-2

Article

A pneumonia outbreak associated with a new coronavirus of probable bat origin


<https://doi.org/10.1038/s41586-020-2012-7>

Received: 20 January 2020

Accepted: 29 January 2020

Published online: 3 February 2020

Open access

 Check for updates

Peng Zhou^{1,5}, Xing-Lou Yang^{1,5}, Xian-Guang Wang^{2,5}, Ben Hu¹, Lei Zhang¹, Wei Zhang¹, Hao-Rui Si^{1,3}, Yan Zhu¹, Bei Li¹, Chao-Lin Huang², Hui-Dong Chen², Jing Chen^{1,3}, Yun Luo^{1,3}, Hua Guo^{1,3}, Ren-Di Jiang^{1,3}, Mei-Qin Liu^{1,3}, Ying Chen^{1,3}, Xu-Rui Shen^{1,3}, Xi Wang^{1,3}, Xiao-Shuang Zheng^{1,3}, Kai Zhao^{1,3}, Quan-Jiao Chen¹, Fei Deng¹, Lin-Lin Liu⁴, Bing Yan¹, Fa-Xian Zhan⁴, Yan-Yi Wang¹, Geng-Fu Xiao¹ & Zheng-Li Shi^{1,2,5}

Since the outbreak of severe acute respiratory syndrome (SARS) 18 years ago, a large number of SARS-related coronaviruses (SARSr-CoVs) have been discovered in the

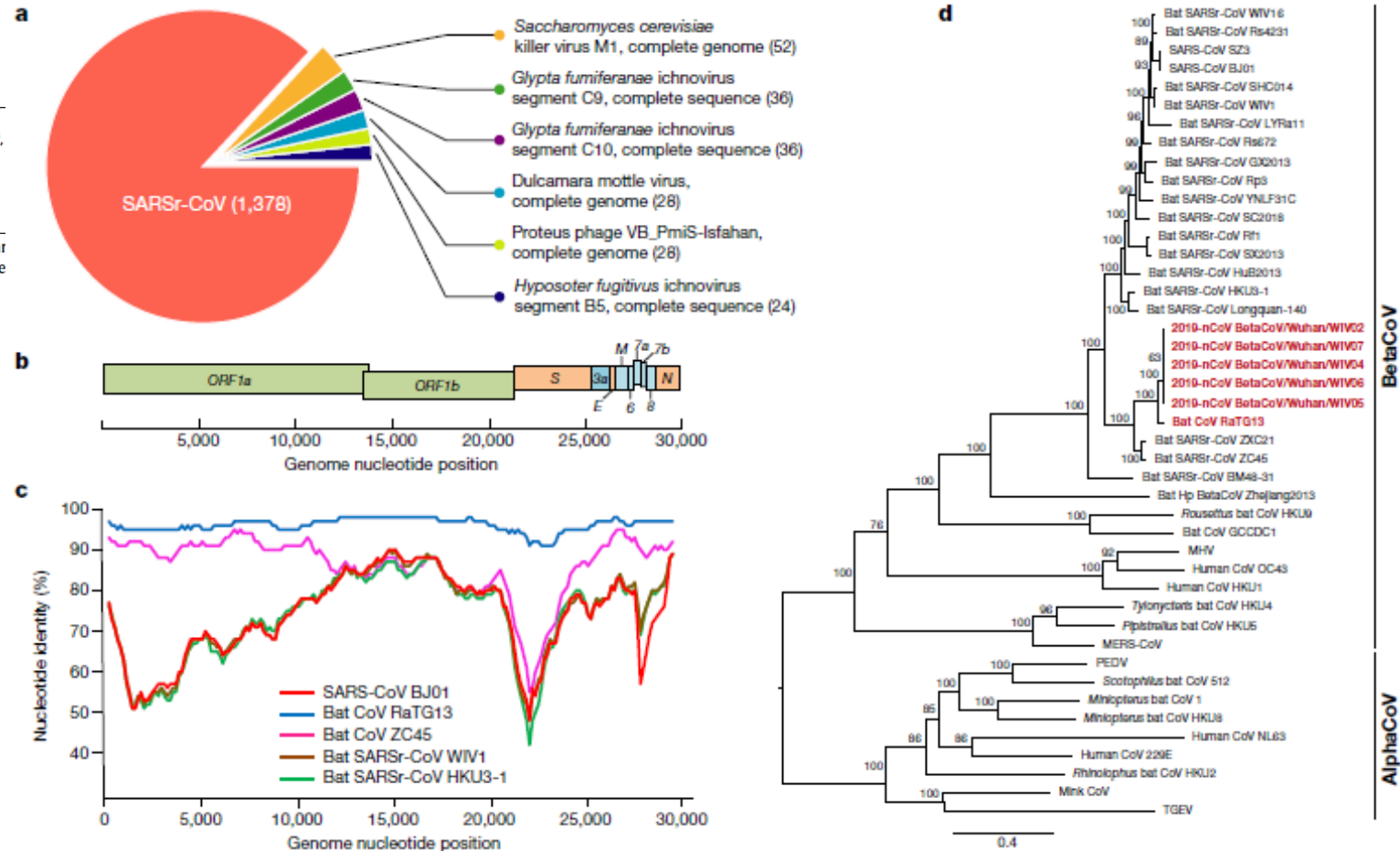
■ Zhou et al. reported a bat coronavirus (BatCoV RaTG13), which was previously detected in *Rhinolophus affinis* from Yunnan province, with an overall genome sequence identity of 96.2% to 2019-nCoV.

■ The receptor-binding spike protein encoded by the S gene was with a 93.1% nucleotide identity to RaTG13.

■ One of its six key amino acid residues involved in the interaction with human ACE2 are same with 2019-nCoV.

■ Simplot analysis showed that 2019-nCoV was highly similar to RaTG13 throughout the genome.

■ 2019-nCoV may have originated in bats.



2. Natural origin of SARS-CoV-2



Check for updates

correspondence

The proximal origin of SARS-CoV-2

To the Editor — Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China^{1,2}, there has been considerable discussion on the origin of the causative virus, SARS-CoV-2³ (also referred to as HCoV-19)⁴. Infections with SARS-CoV-2 are now widespread, and as of 11 March 2020, 121,564 cases have been confirmed in more than 110 countries, with 4,373 deaths⁵.

SARS-CoV-2 is the seventh coronavirus known to infect humans; SARS-CoV, MERS-CoV and SARS-CoV-2 can cause severe disease, whereas HKU1, NL63, OC43 and

While the analyses above suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal⁷ and that the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding^{7,11}. Thus, the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 that permits another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is not the product of purposeful manipulation.

low-pathogenicity avian influenza viruses into highly pathogenic forms¹⁶. The acquisition of polybasic cleavage sites by HA has also been observed after repeated passage in cell culture or through animals¹⁷.

The function of the predicted O-linked glycans is unclear, but they could create a 'mucin-like domain' that shields epitopes or key residues on the SARS-CoV-2 spike protein¹⁸. Several viruses utilize mucin-like domains as glycan shields involved in immunoevasion¹⁸. Although prediction of O-linked glycosylation is robust, experimental studies are needed

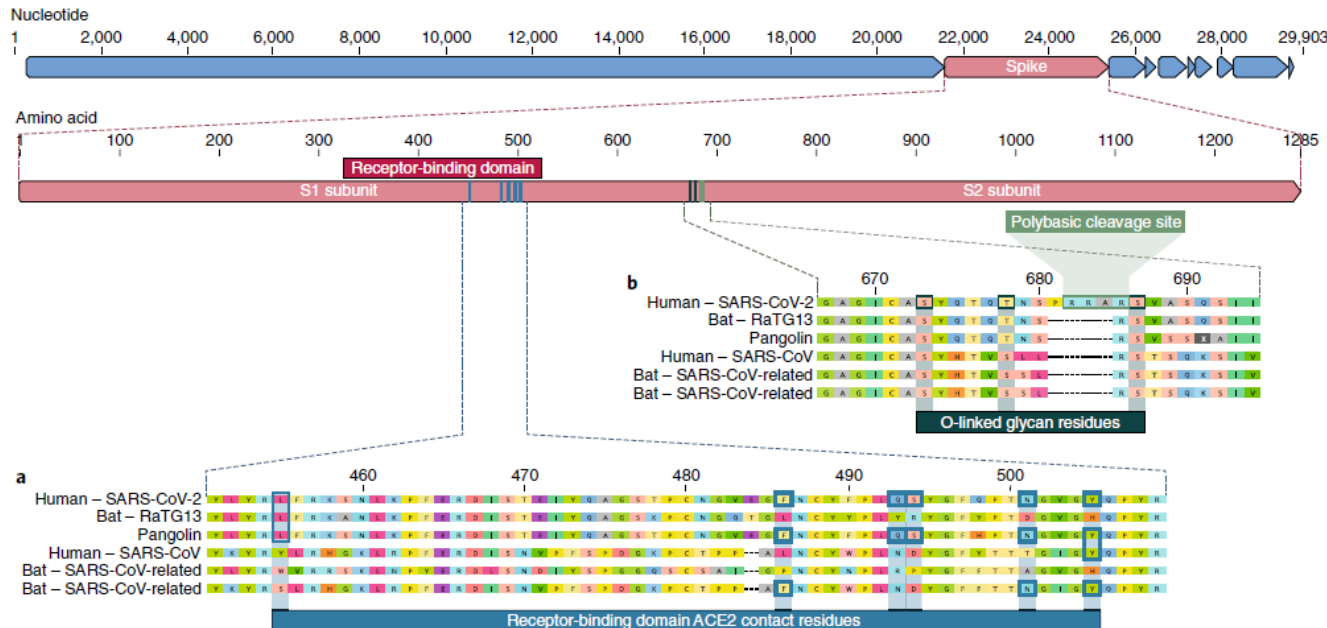


Fig. 1 | Features of the spike protein in human SARS-CoV-2 and related coronaviruses. a, Mutations in contact residues of the SARS-CoV-2 spike protein. The

■ SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor;

■ The highly variable spike (S) protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

■ The origin of SARS-CoV-2: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer.

■ Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

Kristian G., A., Andrew Rambaut, W. Ian Lipkin, Holmes, E. C. & Garry, R. F. The Proximal Origin of SARS-CoV-2. (2020) Nature Medicine, *in press*.

2. Natural origin of SARS-CoV-2



Pangolin may play an important role in the community ecology of coronaviruses.

Article

Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins

<https://doi.org/10.1038/s41586-020-2169-0>

Received: 7 February 2020

Accepted: 17 March 2020

Published online: 26 March 2020

Tommy Tsan-Yuk Lam^{1,2,10}, Na Jia^{3,10}, Ya-Wei Zhang^{3,10}, Marcus Ho-Hin Shum^{2,10}, Jia-Fu Jiang^{3,10}, Hua-Chen Zhu^{1,2}, Yi-Gang Tong^{4,10}, Yong-Xia Shi⁵, Xue-Bing Ni², Yun-Shi Liao², Wen-Juan Li⁴, Bao-Gui Jiang³, Wei Wei⁶, Ting-Ting Yuan³, Kui Zheng⁵, Xiao-Ming Cui³, Jie Li³, Guang-Qian Pei³, Xin Qiang³, William Yiu-Man Cheung², Lian-Feng Li⁷, Fang-Fang Sun⁵, Si Qin³, Ji-Cheng Huang⁶, Gabriel M. Leung², Edward C. Holmes⁸, Yan-Ling Hu^{8,9} & Yi Guan^{1,2} & Wu-Chun Cao³

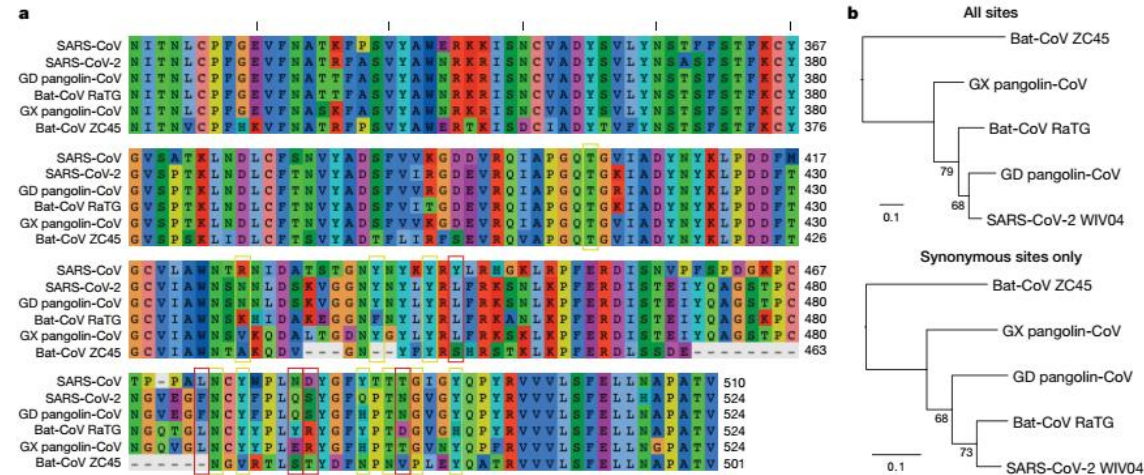


Fig. 3 | Analysis of the RBD sequence. a, Sequence alignment showing the RBD in human, pangolin and bat coronaviruses. The five critical residues for binding between SARS-CoV RBD and human ACE2 protein are indicated in red boxes, and ACE2-contacting residues are indicated by yellow boxes as previously described⁸. In the Guangdong pangolin-CoV sequence, the codon positions encoding the amino acids Pro337, Asn420, Pro499 and Asn519 have ambiguous nucleotide compositions, resulting in possible alternative amino acids at these

sites (threonine, glycine, threonine and lysine, respectively). Sequence gaps are indicated with dashes. The short black lines at the top indicate the positions of every 10 residues. GD, Guangdong; GX, Guangxi. **b**, Phylogenetic trees of the SARS-CoV-2-related lineage estimated from the entire RBD region (top) and synonymous sites only (bottom). Branch supports obtained from 1,000 bootstrap replicates are shown. Branch scale bars are shown as 0.1 substitutions per site.

- Indeed, the Guangdong pangolin coronaviruses and SARS-CoV-2 possess identical amino acids at the five critical residues of the RBD, whereas RaTG13 only shares one amino acid with SARS-CoV-2 (residue 442, according to numbering of the human SARS-CoV9).

Article

Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins

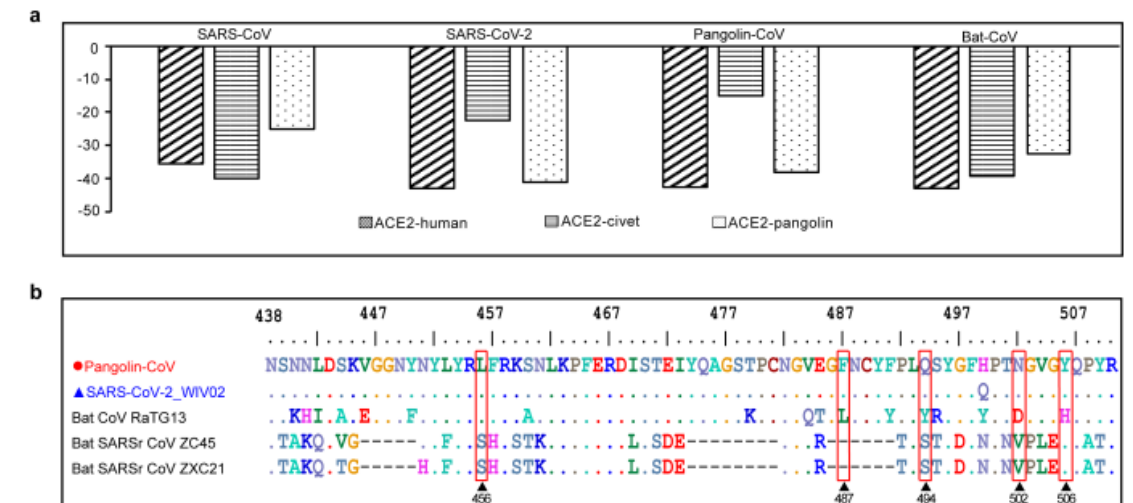
<https://doi.org/10.1038/s41586-020-2313-x>

Received: 16 February 2020

Accepted: 28 April 2020

Published online: 7 May 2020

Kangpeng Xiao^{1,2,7}, Junqiong Zhai^{3,7}, Yaoyu Feng^{1,2}, Niu Zhou³, Xu Zhang^{1,2}, Jie-Jian Zou⁴, Na Li^{1,2}, Yaqiong Guo^{1,2}, Xiaobing Li¹, Xuejuan Shen¹, Zhipeng Zhang¹, Fanfan Shu^{1,2}, Wanyi Huang^{1,2}, Yu Li³, Ziding Zhang⁵, Rui-Ai Chen^{1,6}, Ya-Jiang Wu³, Shi-Ming Peng³, Mian Huang³, Wei-Jun Xie³, Qin-Hui Cai³, Fang-Hui Hou³, Wu Chen³ & Yongyi Shen^{1,2}



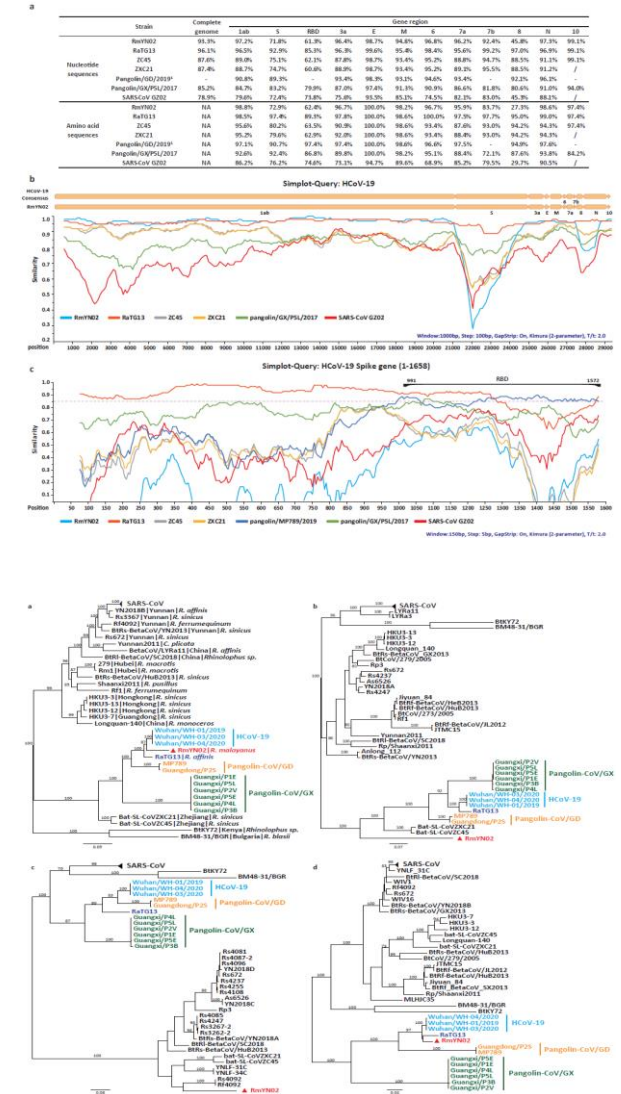
2. Natural origin of SARS-CoV-2



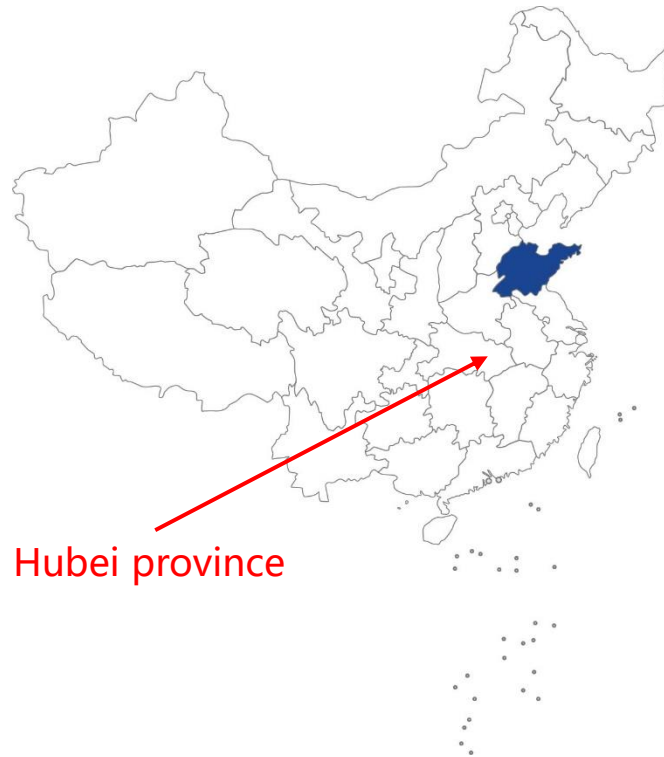
A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein

- Here, we report a novel bat-derived coronavirus, denoted RmYN02, identified from a metagenomic analysis of samples from 227 bats collected from Yunnan Province in China between May and October 2019.
- Notably, RmYN02 shares 93.3% nucleotide identity with SARS-CoV-2 at the scale of the complete virus genome and 97.2% identity in the 1ab gene, in which it is the closest relative of SARS-CoV-2 reported to date.
- In contrast, RmYN02 showed low sequence identity (61.3%) to SARS-CoV-2 in the receptor-binding domain (RBD) and might not bind to angiotensin-converting enzyme 2 (ACE2).
- Critically, and in a similar manner to SARS-CoV-2, RmYN02, was characterized by the insertion of multiple amino acids at the junction site of the S1 and S2 subunits of the spike (S) protein. This provides strong evidence that such insertion events can occur naturally in animal *betacoronaviruses*.

Zhou *et al.* A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Current Biology*, 2020, 30: 2196-2203.



3. Early spread of SARS-CoV-2 in China



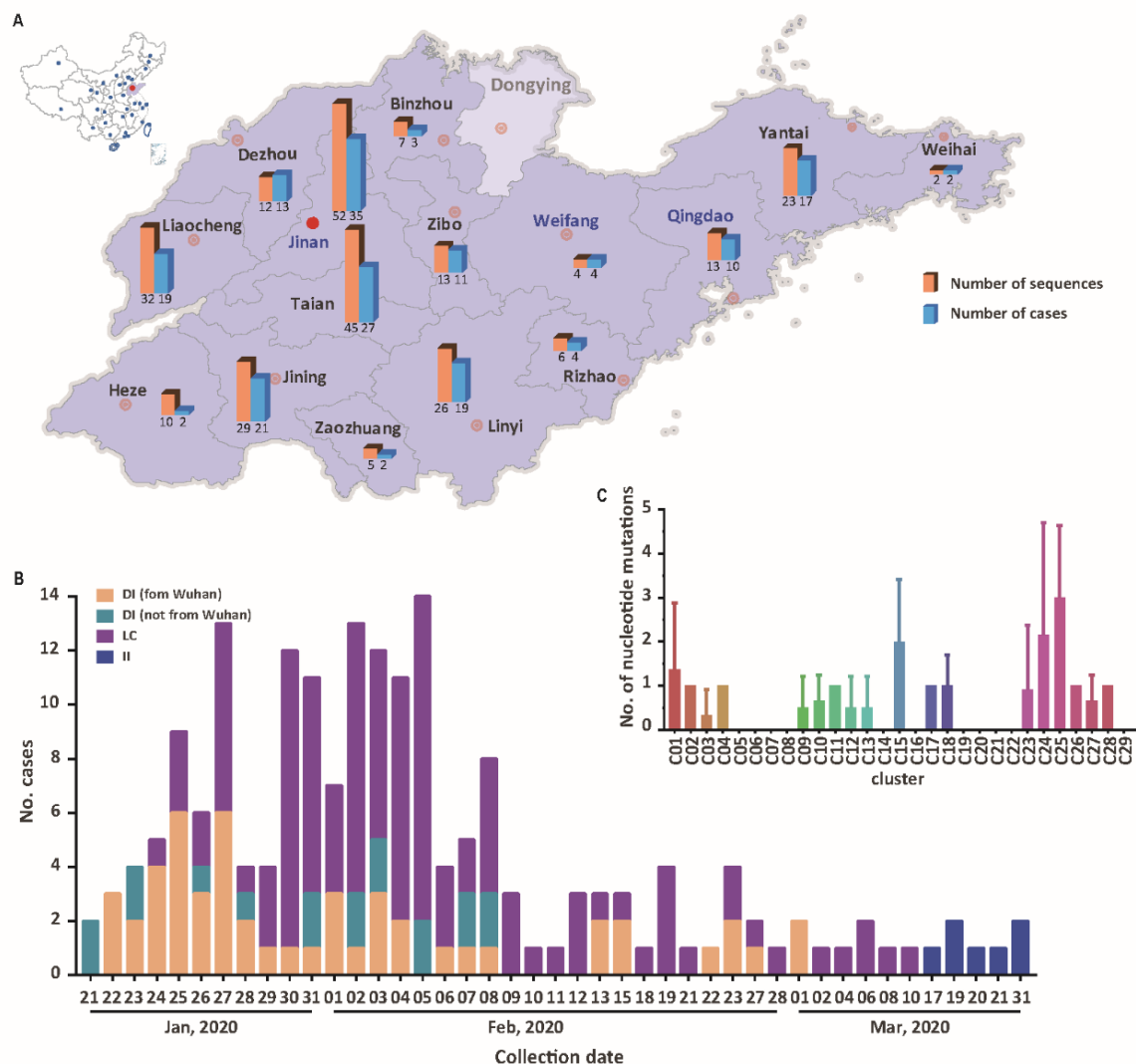
- Shandong, located in Eastern China, is a highly populous province, with a population of 100.7 million at the end of 2019. As of September 23th, 2020, Shandong province has reported a total of 763 SARS-CoV-2 cases including 102 transmission chain events, beginning on January 21st, 2020. In addition, 69 cases were internationally imported into Shandong from 16 countries.
- We performed next generation sequencing (NGS) of 390 clinical/cell culture samples from 292 confirmed COVID-19 cases, covering ~ 35% of all reported cases in Shandong province.
- From these, we obtained 196 full-length genome sequences from 165 COVID-19 cases, including 150 respiratory tract samples, 17 fecal samples, 15 samples from cell culture and the remaining 14 samples of unknown type.

Phylogenetic and genomic analysis of 196 full-length SARS-CoV-2 genome sequences, combined with detailed epidemiological data, revealed the genomic epidemiology of COVID-19 during the duration of the outbreak in Shandong province.

3. Early spread of SARS-CoV-2 in China

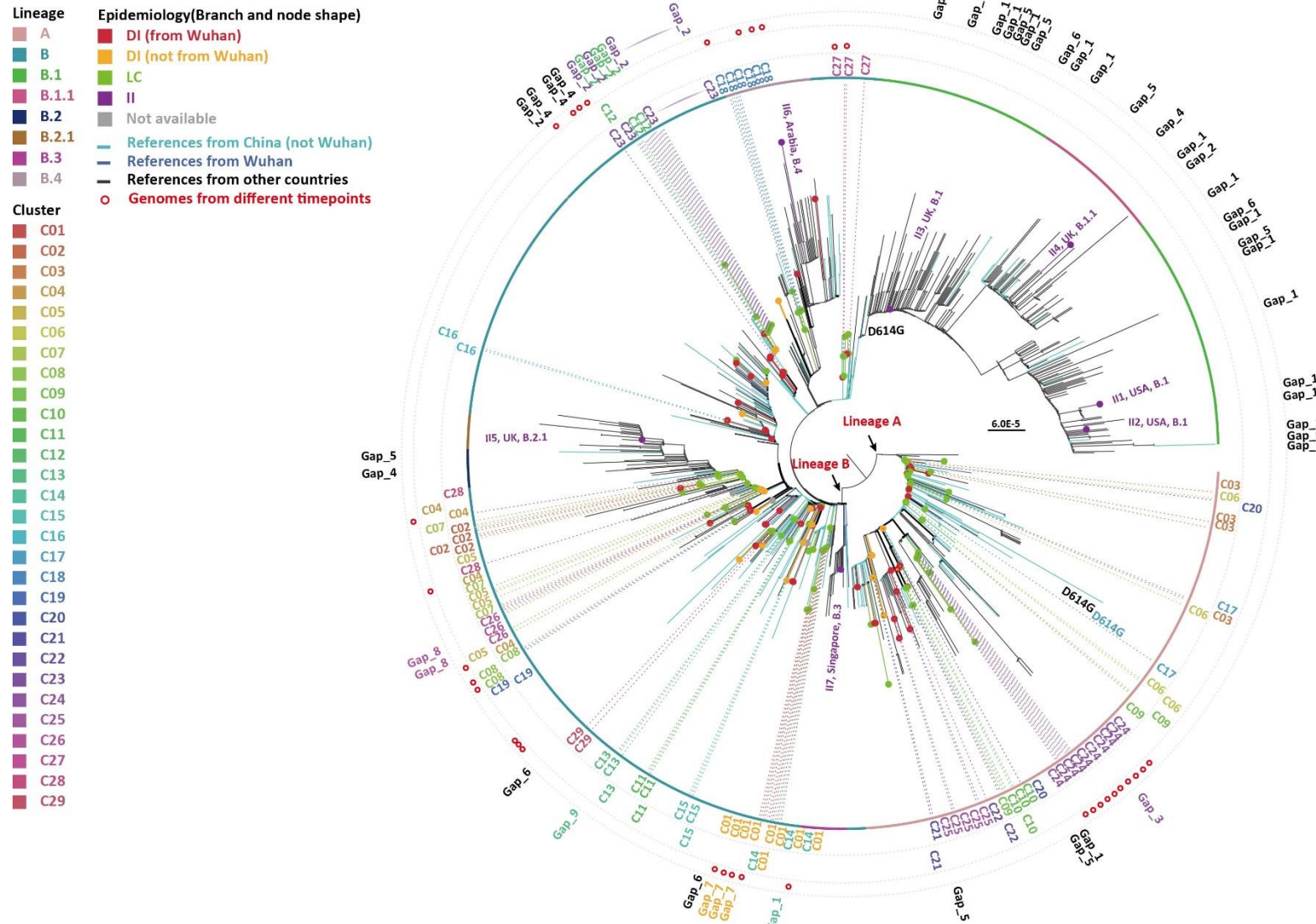


Spatiotemporal distribution of 196 SARS-CoV-2 genome sequences across Shandong province, China.

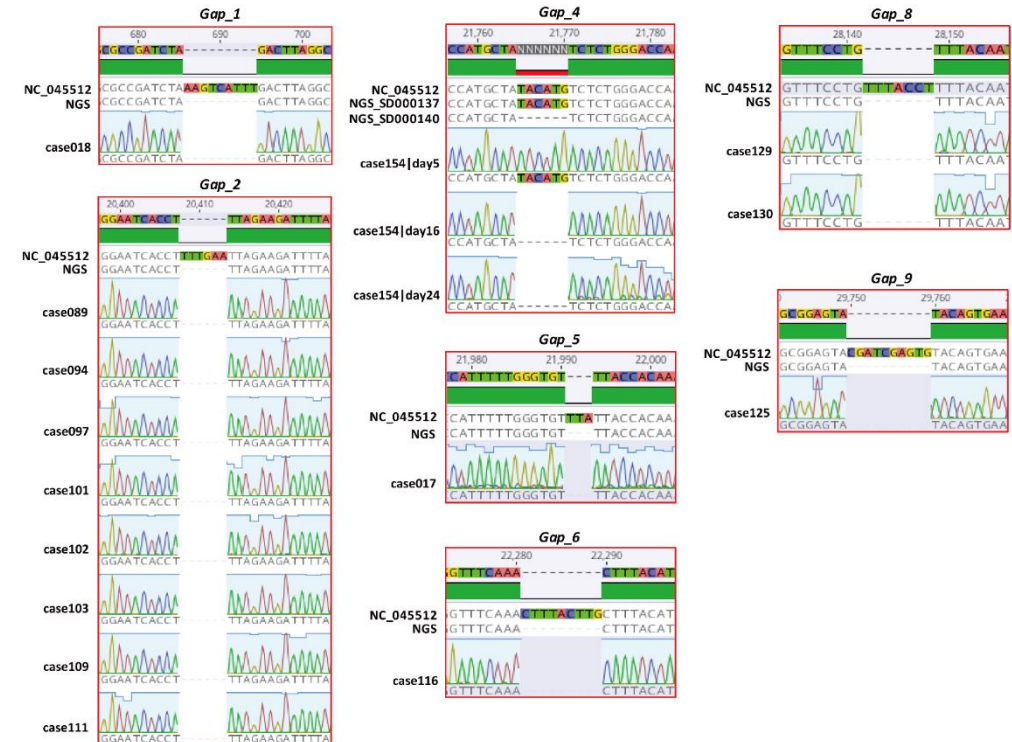


- The 196 sequences from 165 positive cases were distributed in all 15 cities that reported COVID-19 cases in Shandong province and most of viruses sequenced were collected during the peak period covering January to early February (79.59%, 156/196).
- 94 cases (with 114 sequences) were acquired through Local Community transmission linked to 29 different transmission chains (C1-C29). The largest cluster comprised 14 full-length genomes related to the nosocomial person-to-person transmission.
- There was no genomic variations observed within 10 clusters, and the largest number of nucleotide differences between the index case and its subsequent infected individuals was three.

3. Early spread of SARS-CoV-2 in China

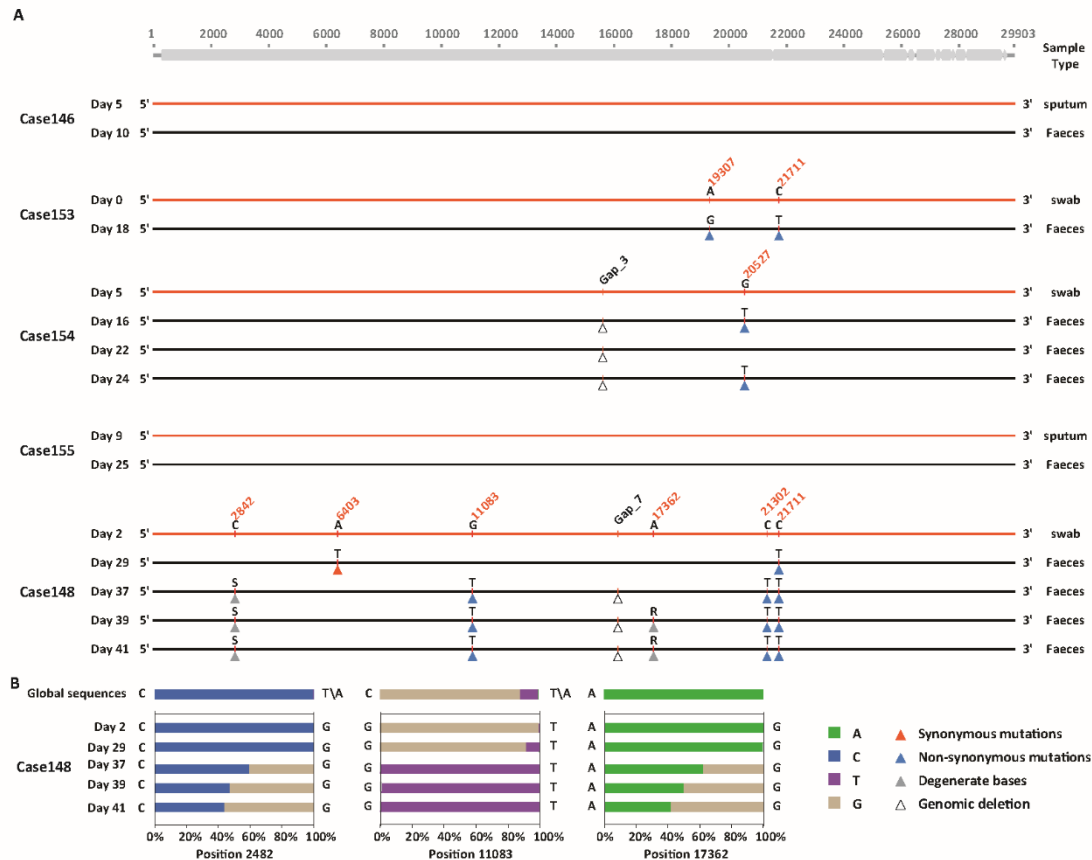


- **Phylogenetic analysis of 196 SARS-CoV-2 genome sequences from Shandong and representative strains worldwide.**
- The great majority (188/189) of the Domestically Imported and Local Community transmission infection sequences from Shandong province belonged to the basal lineages A (8782T:28144C) and B (8782C:28144T), while all the International Importation sequences comprised sublineages of lineage B.
- Subsequent viral lineage diversification had happened in Wuhan before multiple independent importation events transmitted into Shandong province in late January and February.



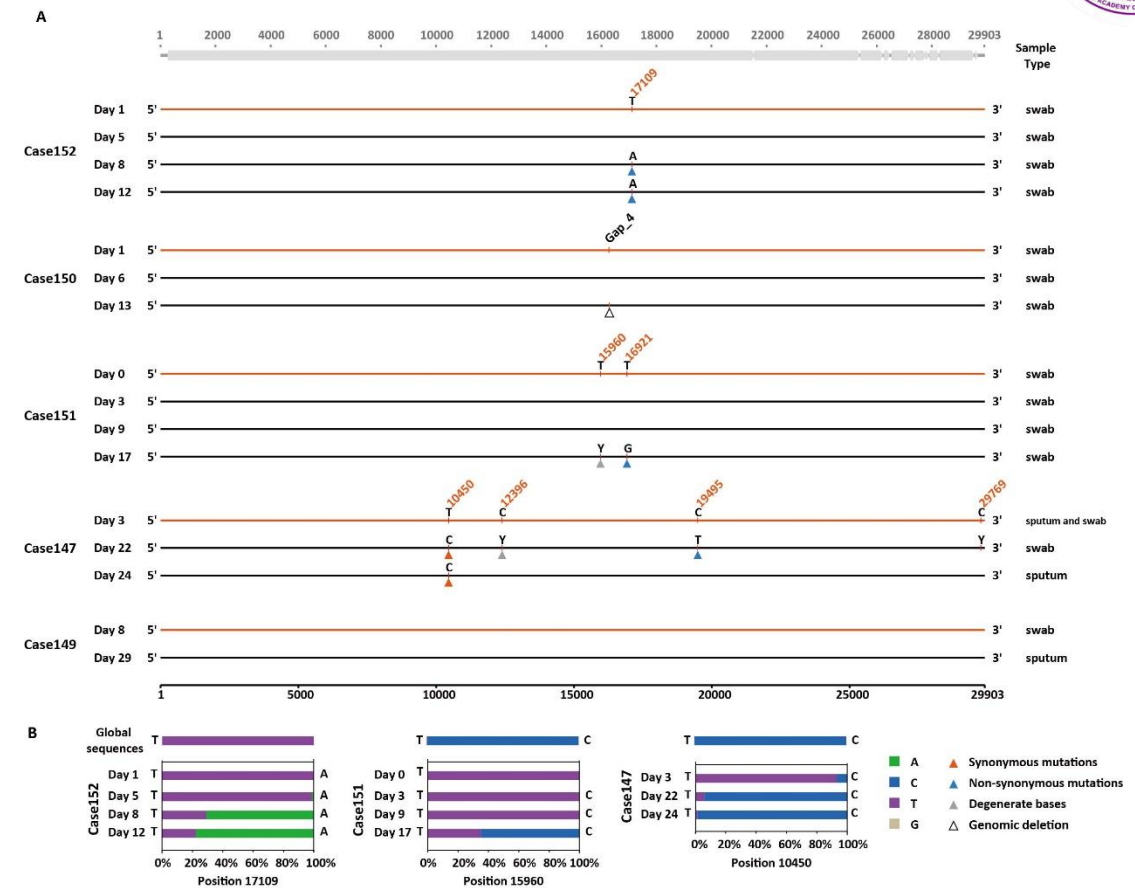
- **Nine different genomic deletions in 31 viruses** are shown from the 5' to the 3' part of the genome.
- All the deletions identified in Shandong were found in other locations with the exception of gap_3 and gap_8 by investigating the presence of these deletion mutations in the SARS-CoV-2 genomes available in the GISAID database.
- Gap_2 and gap_6 were respectively found in cases returning from Wuhan indicating the two gaps might have first emerged in Wuhan.
- Gap_5 was found in one individual returning from Yunnan province.
- Gap 4, gap 7, gap 1, gap 9, gap 3 and gap 8 likely arose locally in Shandong.

3. Early spread of SARS-CoV-2 in China



Comparison of SARS-CoV-2 genomes sampled from feces at different time points.

- SARS-CoV-2 genome nucleotide variations in cases 146, 153, 154, 155, and 148 from different time points. The iSNVs observed at positions 2482, 11083, and 17362 of SARS-CoV-2 genomes of case 148 from different time points.



Comparison of SARS-CoV-2 genomes from respiratory samples at different time points.

- Three iSNVs were notable: 17109 (T→A) in case 152, 15960 (T→C) in case 151, and 10450 (T→C) in case 147.

4. Infectomes of SARS-CoV-2 in China



*Large-scale surveillance studies → PCR-based methods

Clinical diagnosis of 8274 samples with 2019-novel coronavirus in Wuhan

Ming Wang, Qing Wu, Wanzhou Xu, Bin Qiao, Jingwei Wang, Hongyun Zheng, Shupeng Jiang, Junchi Mei, Zegang Wu, Yayun Deng, Fangyuan Zhou, Wei Wu, Yan Zhang, Zhihua Lv, Jingtao Huang, Xiaoqian Guo, Lina Feng, Zunen Xia, Di Li, Zhiliang Xu, Tiangang Liu, Pingan Zhang, Yongqing Tong, Yan Li

doi: <https://doi.org/10.1101/2020.02.12.20022327>

- Respiratory electrophoresis fragment analysis with PCR

Co-infections of SARS-CoV-2 with multiple common respiratory pathogens in infected patients

Dachuan Lin^{1†}, Lei Liu^{2†}, Mingxia Zhang^{2†}, Yunlong Hu¹, Qianting Yang², Jiubiao Guo¹, Yongchao Guo¹, Youchao Dai¹, Yuzhong Xu³, Yi Cai¹, Xinchun Chen¹, Zheng Zhang^{2*} & Kaisong Huang^{1*}

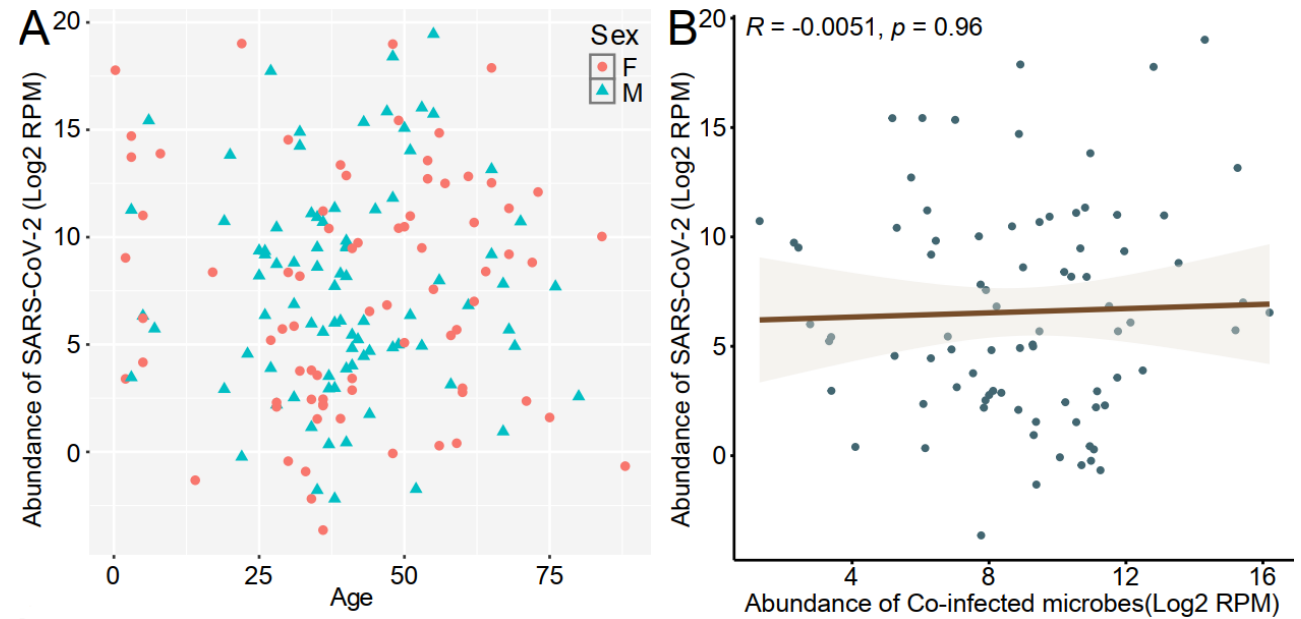
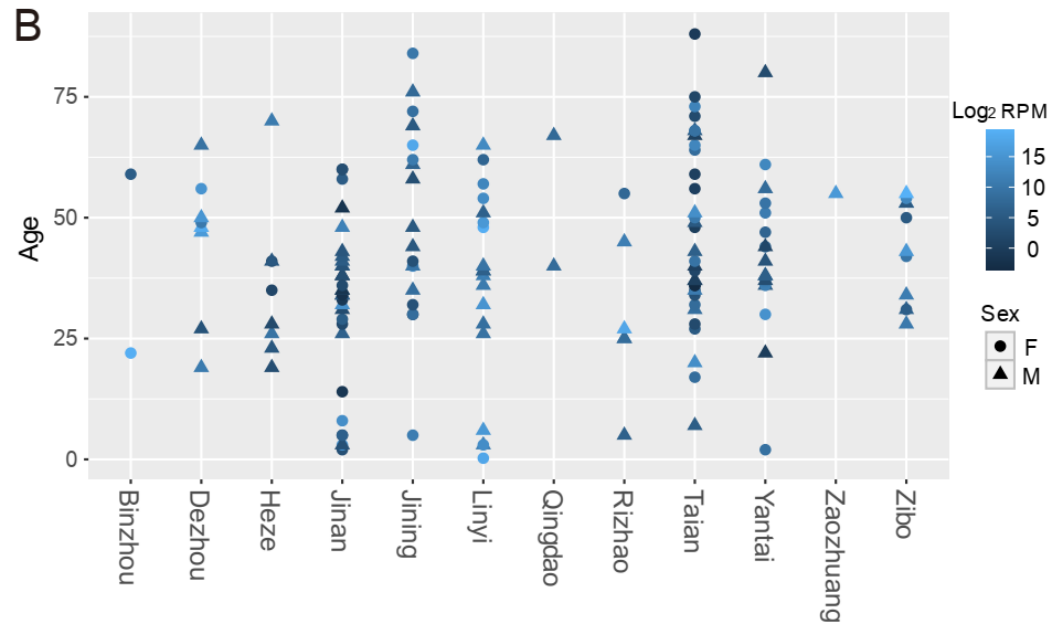
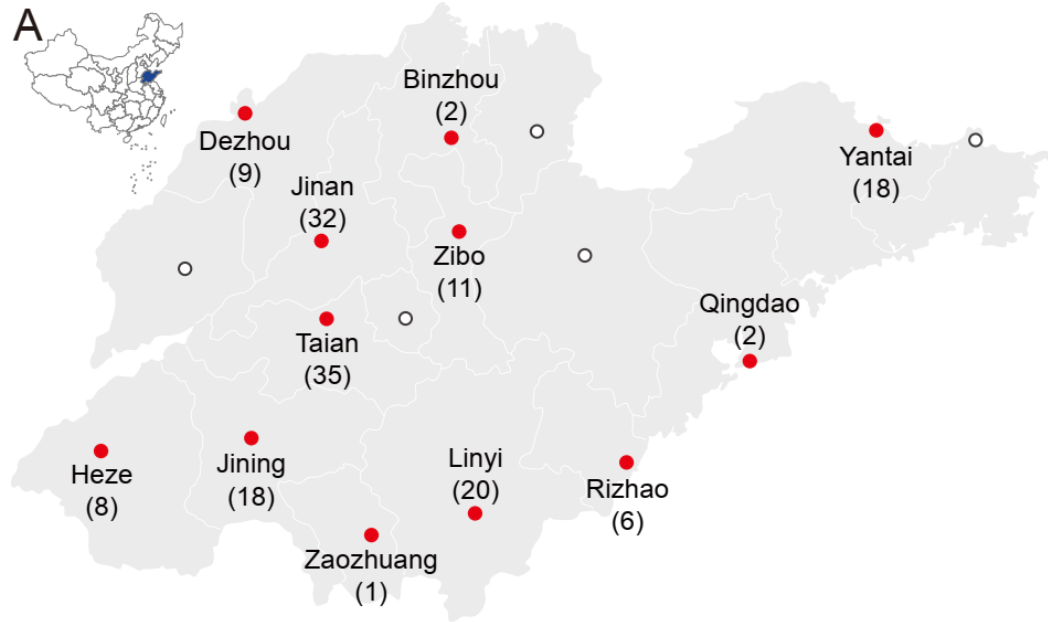
- Multiplex rapid detection kit 2.0 (Uni-MEDICATech)
- Multiplex RT-PCR method

Clinical features and short-term outcomes of 221 patients with COVID-19 in Wuhan, China

Guqin Zhang^{a, 1}, Chang Hu^{b, 1}, Linjie Luo^{c, 1}, Fang Fang^d, Yongfeng Chen^e, Jianguo Li^b, Zhiyong Peng^b, Huaqin Pan^b ✉

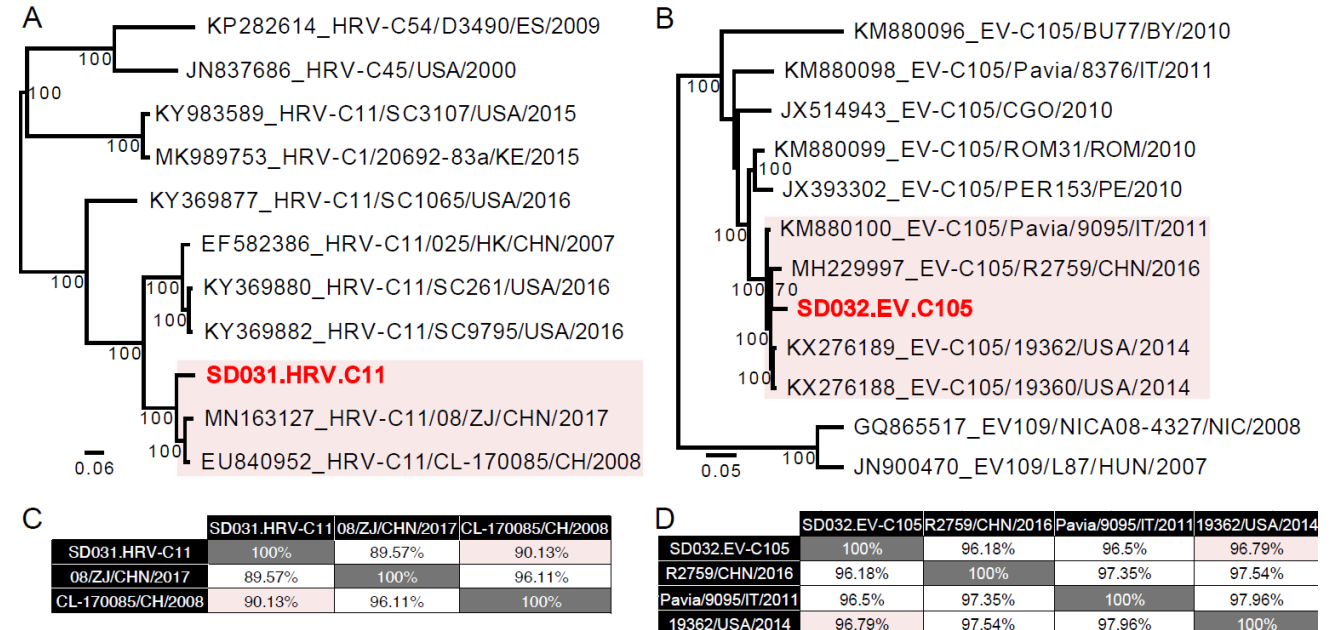
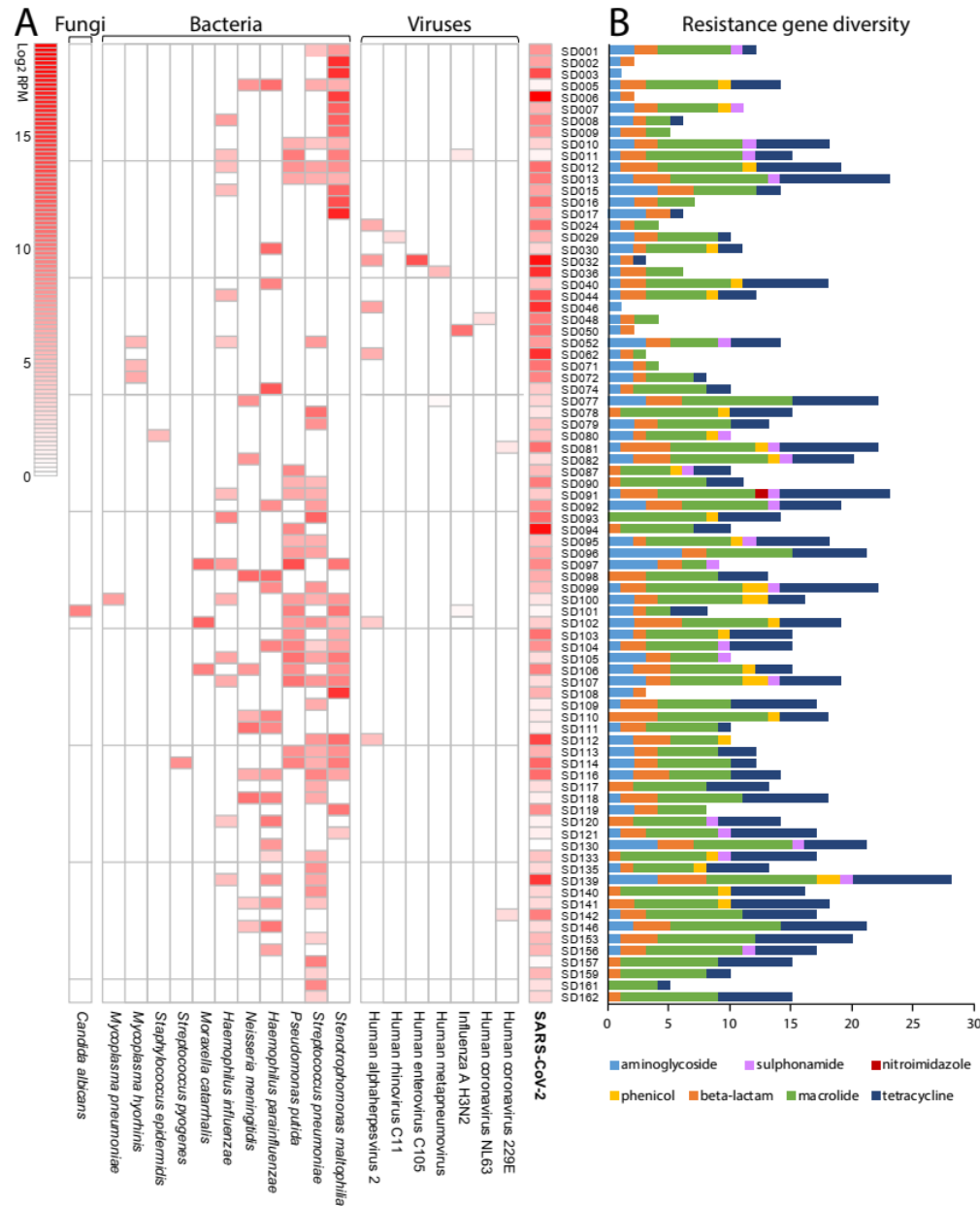
- Tested other respiratory viruses by real-time RT-PCR

4. Infectomes of SARS-CoV-2 in China



- A total of 162 SARS-CoV-2 positive patients from 12 cities in Shandong province, China were enrolled into this study.
- No positive correlations between the age or sex of the patients, nor the abundance of the co-infected microbes, and the abundance of SARS-CoV-2.

4. Infectomes of SARS-CoV-2 in China



- Seven viruses with potential pathogenicity in 15 of the 162 (9.26%) COVID-19 cases were identified , comprising one DNA virus and six RNA viruses.
- Human rhinovirus C11 and human enterovirus C105 were relatively new genotypes → the first time reported in COVID-19 cases.

4. Infectomes of SARS-CoV-2 in China

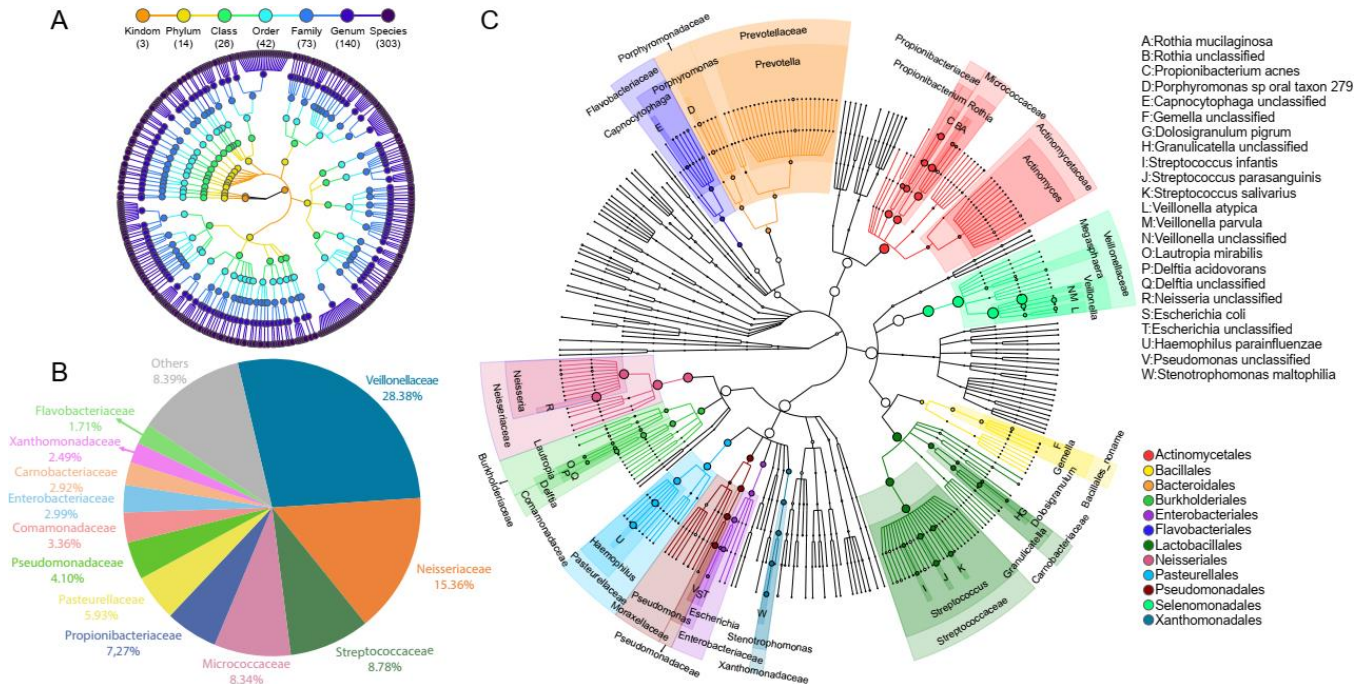


Table 1. Summary of co-infections of SARS-CoV-2 and other microbes

SARS-CoV-2	Co-infecting microbes	No. of cases	Percent
+	Virus	9	1.85%
+	Multiple viruses	1	0.62%
+	Virus + Bacteria	4	2.47%
+	Virus + Bacteria + Fungi	1	0.62%
+	One bacterial species	33	20.37%
+	Multiple bacteria	34	20.99%
Total		82	50.62%

- Several important pathogenic and/or opportunistic bacteria were also identified to the species level:
 - *Streptococcus pneumoniae* (n=37), *Stenotrophomonas maltophilia* (n=31), *Pseudomonas putida* (n=21), *Haemophilus parainfluenzae* (n=19), *Haemophilus influenzae* (n=14), *Neisseria meningitidis* (n=11), *Moraxella catarrhalis* (n=3), *Streptococcus pyogenes* (n=1), *Streptococcus epidermidis* (n=1)
 - *Mycoplasma hyorhinis* (n=3), *Mycoplasma pneumoniae* (n=1)
- Importantly, bacteria present in multiple cases were all confirmed by PCR using species-specific primers.
- **Overall, 82 of the 162 SARS-CoV-2 cases (50.62%) were co-infected by at least one additional potentially pathogenic microbe.**

Thank you for your attention.

Questions?