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](images)

**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.


1. Virosphere & genetic sequencing

Genomic analysis of uncultured marine viral communities

Miya Breitbart¹,², Peter Saltonstall¹,³, Bjørn Andreassen¹,³, Joseph M. Mahaffy¹,¹, Anna M. Segal²,¹, David Mendelson¹, Firas A. Azzam¹, and Forest Rohwer¹

¹Department of Biology, San Diego State University, San Diego, CA 92182-1081. ²Department of Mathematics, San Diego State University, University City, CA 92123-7520. ³Cary Institute of Ecosystem Studies, Millbrook, NY 12546. ⁴Harvard University, Cambridge, MA 02138. ⁵University of Chicago, Chicago, IL 60637.

Communicated by Allen Campbell, Stanford University, Stanford, CA. Accepted August 14, 2002; revised for publication February 22, 2003.

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. 90% of the sequences were not found in the GenBank database, suggesting that much of the diversity is presently unknown. The most common significant hits among the known sequences were to viruses. The viral hit included sequences from all of the major families of dsDNA viruses, as well as some novel viruses. Several independent mathematical models based on the observed number of contigs predicted that the total nucleic acid genome comprised 2.3·10²³ in all communities, which are equalled in content between 5% and 7% of all known genomes. The proportion of these sequences that were not detected by the BLAST search suggests that 30% to 70% of the unknown diversity is potentially available. These 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 influence on global biogeochemical cycles (1, 2), global change (3), and global climate (4). Despite their enormous importance, very little is known about the diversity or the phylogenetic relationships of marine and terrestrial viruses because they cannot be cultured and require a cyanobacterial host (5). However, we now present a new ecological tool to discover viruses that must be cultured on hosts, the majority of which cannot be cultured by using standard techniques (6). In addition, viruses do not have ubiquitous conserved genetic elements such as tRNA that can be used as diversity and evolutionary distance markers (7). To circumvent these limitations, we developed a method to design oligo-rich and sequence-unrelated oligonucleotides.

Materials and Methods

Isolation of Viral Community DNA. Marine viruses were isolated from 200 cm² of surface seawater from Siboga Endeavor (96°W; 3°15’S) and selected samples from the Indian Ocean (20°W; 1°40’S) and the Pacific Ocean (132°W; 2°55’S) by using a combination of filtration and cell lysis. Marine viral DNA was isolated from E. coli K-12 cells infected with a T4 phage to a concentration of 10¹⁰ plaque forming units (PFU) per liter of cell. This concentration was the maximum level of contamination in the next steps of the isolation process. The concentration was determined using the absorbance at 260 nm and the known absorbance of the T4 phage. The concentration of the viral DNA was determined using a standard curve for each concentration of viral DNA.

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.
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

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

<table>
<thead>
<tr>
<th>Databases and Software</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data quality control</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FastQC</td>
<td><a href="http://www.bioinformatics.babraham.ac.uk/projects/fastqc/">http://www.bioinformatics.babraham.ac.uk/projects/fastqc/</a></td>
<td>- [38]</td>
</tr>
<tr>
<td>FastQ Screen</td>
<td><a href="http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/">http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/</a></td>
<td>-</td>
</tr>
<tr>
<td>Trimmomatic</td>
<td><a href="http://www.usadellab.org/cms/index.php?page=trimmomatic">http://www.usadellab.org/cms/index.php?page=trimmomatic</a></td>
<td>-</td>
</tr>
<tr>
<td>Fastx-toolkit</td>
<td><a href="http://bannonlab.cslu.edu/fastx_toolkit/">http://bannonlab.cslu.edu/fastx_toolkit/</a></td>
<td>-</td>
</tr>
<tr>
<td>Skewer</td>
<td><a href="https://sourceforge.net/projects/skewer">https://sourceforge.net/projects/skewer</a></td>
<td>- [40]</td>
</tr>
<tr>
<td>AfterQC</td>
<td><a href="https://github.com/OpenGene/AfterQC">https://github.com/OpenGene/AfterQC</a></td>
<td>- [41]</td>
</tr>
<tr>
<td>SOAPnuke</td>
<td><a href="https://github.com/BGI-flexlab/SOAPnuke">https://github.com/BGI-flexlab/SOAPnuke</a></td>
<td>- [42]</td>
</tr>
<tr>
<td><strong>Fastp</strong></td>
<td><a href="https://github.com/OpenGene/fastp">https://github.com/OpenGene/fastp</a></td>
<td>- [43]</td>
</tr>
<tr>
<td>NanoPack</td>
<td><a href="https://github.com/wdecoster/nanopack">https://github.com/wdecoster/nanopack</a></td>
<td>- [63]</td>
</tr>
<tr>
<td>Porechop</td>
<td><a href="https://github.com/rripp/Covechop">https://github.com/rripp/Covechop</a></td>
<td>-</td>
</tr>
<tr>
<td><strong>Read Mapping</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hisat2</td>
<td><a href="https://daehwankimlab.github.io/htsat2/">https://daehwankimlab.github.io/htsat2/</a></td>
<td>- [44]</td>
</tr>
<tr>
<td>BWA</td>
<td><a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a></td>
<td>- [45]</td>
</tr>
<tr>
<td>STAR</td>
<td><a href="https://github.com/alexdobin/STAR">https://github.com/alexdobin/STAR</a></td>
<td>-</td>
</tr>
<tr>
<td>KMA</td>
<td><a href="https://bitbucket.org/genomicscience/kma">https://bitbucket.org/genomicscience/kma</a></td>
<td>- [47]</td>
</tr>
<tr>
<td>SortmRNA</td>
<td><a href="http://bioinfo.lif.fr/RNA/sortmerna">http://bioinfo.lif.fr/RNA/sortmerna</a></td>
<td>- [48]</td>
</tr>
<tr>
<td>Minimap2</td>
<td><a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a></td>
<td>- [64]</td>
</tr>
<tr>
<td><strong>Geneome Visualization</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGV</td>
<td><a href="http://software.broadinstitute.org/software/igv/">http://software.broadinstitute.org/software/igv/</a></td>
<td>- [61]</td>
</tr>
<tr>
<td>Geneious</td>
<td><a href="https://www.geneious.com/">https://www.geneious.com/</a></td>
<td>-</td>
</tr>
<tr>
<td>QUAST</td>
<td><a href="https://sourceforge.net/projects/quast/">https://sourceforge.net/projects/quast/</a></td>
<td>- [62]</td>
</tr>
<tr>
<td>SEQuAMAN</td>
<td><a href="https://www.dna-star.com/software/molecular-biology/">https://www.dna-star.com/software/molecular-biology/</a></td>
<td>-</td>
</tr>
<tr>
<td><strong>Database</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global Initiative on</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sharing All Influenza</td>
<td><a href="https://www.epicov.org/">https://www.epicov.org/</a></td>
<td>- [133]</td>
</tr>
<tr>
<td>Data (GISAID)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>National bioinformatics Center (CNBC)</td>
<td><a href="https://bigd.big.ac.cn/ncov/">https://bigd.big.ac.cn/ncov/</a></td>
<td>-</td>
</tr>
<tr>
<td>Genomics Data Center</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NGDC) database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Center for</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biotechnology Information (NCBI)</td>
<td><a href="https://www.ncbi.nlm.nih.gov/">https://www.ncbi.nlm.nih.gov/</a></td>
<td>-</td>
</tr>
<tr>
<td>Genome Warehouse (GWH)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sequence alignment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLUSTALW</td>
<td><a href="https://www.genome.jp/tools-bin/clustalw">https://www.genome.jp/tools-bin/clustalw</a></td>
<td>- [70]</td>
</tr>
<tr>
<td>MAFFT</td>
<td><a href="https://mafft.cbrc.jp/alignment/software/">https://mafft.cbrc.jp/alignment/software/</a></td>
<td>- [71]</td>
</tr>
<tr>
<td>MUSCLE</td>
<td><a href="http://drive5.com/muscle/">http://drive5.com/muscle/</a></td>
<td>- [72]</td>
</tr>
</tbody>
</table>

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.
2. Natural origin of SARS-CoV-2

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

- 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

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

- 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.

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**

Tommy Tsan-Yui Lam1,2,3,4, Na Raia,5, Ya Wei Zhang1,5, Marcus Ho-Hin Shum4,5, Xia Fu Jiang1,6, Hue-Chen Zhu1, Yi-Gang Tong1,5, Yong-Xia Shi1, Yue-Bing Ni7, Yun Shi Liao1, Wen-Juan Li6, Bang Gui Jiang1, Wei Wei1, Ting Ting Yuan1, Kai Zheng2, Xiao Ming Cui1, Jie Li7, Guang-Qian Pei7, Xin Qiang7, William Yu Men Cheung5, Lian-Feng Li7, Fang-Feng Sun7, Si Qin8, Ji-Cheng Huang9, Gabriel M. Leung4, Edward C. Holmes8, Yan-Ling Hu10, Yi Guan10 & Woe-Chun Cao11

https://doi.org/10.1038/s41586-020-2169-z

Received: 7 February 2020
Accepted: 17 March 2020
Published online: 26 March 2020

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-CoV).

**Article**

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

Kangpeng Xiao1,2,3, Junqiong Zhou1,2, Yuyao Feng1,2, Niu Zhou1,2, Xu Zhang1,2, He-Jian Zou1,2, Na Li3, Yaqong Guo2,3, Xiaobing Li7, Xuejuan She3, Zhipeng Zhang1, Fanfan Shu3, Wanyi Huang3,4, Yu Li5, Ziding Zhang5, Rui-Ai Chen6, Yi-Jiang Wu7, Shi-Ming Peng7, Minen Huang2, Wei-Jun Xie8, Qin-Hui Cai2, Fang-Hui Hou9, Wu Chen1,2, Lihu Xia1,2 & Yongxi She1,2

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

Received: 16 February 2020
Accepted: 28 April 2020
Published online: 7 May 2020

---

**Fig. 1. 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 described. In the Guangdong pangolin-CoV sequence the codon positions encoding the amino acids Pro337, Asn420, Pro499 and Asn839 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 1000 bootstrap replicates are shown. Branch scale bars are shown as 0.1 substitutions per site.
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.

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 23rd, 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

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.
Genomic deletions identified in the SARS-CoV-2 from Shandong and validation of the genome deletions by Sanger sequencing.

- 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, Yuyun Deng, Fangyuan Zhou, Wei Wu, Yan Zhang, Zhihua Lv, Jingtao Huang, Xiaoqian Guo, Lina Feng, Zunen Xia, Di Li, Zhiliang Xu, Tiangang Liu, Pingen Zhang, Yongqing Tong, Yan Li
doi: https://doi.org/10.1101/2020.02.12.20022327

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

Dachuan Lin¹, Lei Lin², Mingxia Zhang¹, Yunlong Hu¹, Qianting Yang¹, Jitiaoao Guo¹, Yongchao Guo¹, Youchao Dai¹, Yuzhong Xu¹, Yi Cai¹, Xinchun Chen¹, Zheng Zhang² & Kaisong Huang¹

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

Guqin Zhang³, Chang Hu³, Linjie Luo³, Fang Fang³, Yongfeng Chen³, Jianguo Li³, Zhiyong Peng³, Huaqin Pan³ &
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.
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.
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?