

**Appendix for**  
**Genomic epidemiology of SARS-CoV-2 in the United Arab Emirates reveals novel virus mutation, patterns of co-infection and tissue specific host innate immune response**

Rong Liu<sup>1,2</sup>, Pei Wu<sup>1,2</sup>, Pauline Ogrodzki<sup>1</sup>, Sally Mahmoud<sup>1</sup>, Ke Liang<sup>3</sup>, Pengjuan Liu<sup>3</sup>, Stephen S. Francis<sup>4,5</sup>, Hanif Khalak<sup>1</sup>, Denghui Liu<sup>6</sup>, Junhua Li<sup>2,7</sup>, Tao Ma<sup>3</sup>, Fang Chen<sup>3</sup>, Weibin Liu<sup>2</sup>, Xinyu Huang<sup>3</sup>, Wenjun He<sup>6</sup>, Zhaorong Yuan<sup>6</sup>, Nan Qiao<sup>6</sup>, Xin Meng<sup>6</sup>, Budoor Alqarni<sup>1</sup>, Javier Quilez<sup>1</sup>, Vinay Kusuma<sup>1</sup>, Long Lin<sup>2</sup>, Xin Jin<sup>2</sup>, Chongguang Yang<sup>8</sup>, Xavier Anton<sup>1</sup>, Ashish Koshy<sup>1</sup>, Huanming Yang<sup>2</sup>, Xun Xu<sup>2</sup>, Jian Wang<sup>2</sup>, Peng Xiao<sup>1</sup>, Nawal Ahmed Mohamed Al Kaabi<sup>9</sup>, Mohammed Saifuddin Fasihuddin<sup>9</sup>, Francis Amirtharaj Selvaraj<sup>9</sup>, Stefan Weber<sup>9</sup>, Farida Ismail Al Hosani<sup>10</sup>, Siyang Liu<sup>2#</sup>, Walid Abbas Zaher<sup>1#</sup>

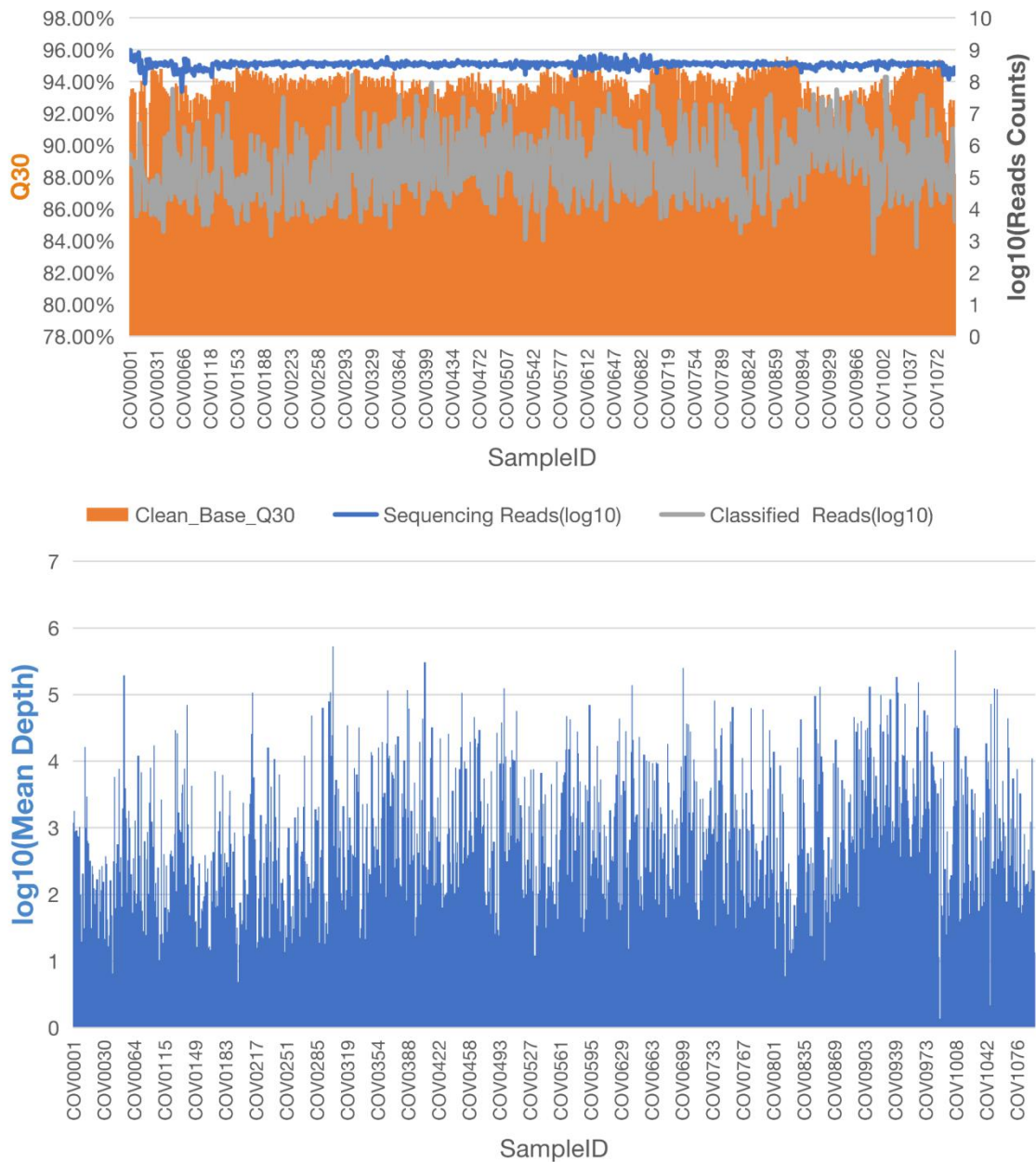
1. Group42 Healthcare, Abu Dhabi, United Arab Emirates
2. BGI-Shenzhen, Shenzhen 518083, Guangdong, China
3. MGI, BGI-Shenzhen, Shenzhen 518083, Guangdong, China
4. Department of Neurological Surgery, University of California, San Francisco
5. Department of Epidemiology and Biostatistics, University of California, San Francisco
6. Laboratory of Health Intelligence, Huawei Technologies Co., Ltd, Shenzhen, 518100, China
7. Shenzhen Key Laboratory of Unknown Pathogen Identification, BGI-Shenzhen, Shenzhen 518083, China
8. Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven
9. SEHA, Abu Dhabi Health Services Co, Abu Dhabi, United Arab Emirates
10. Department of Health, Abu Dhabi, United Arab Emirates

Correspondence to  
Siyang Liu: [liusiyang@bgi.com](mailto:liusiyang@bgi.com)  
Walid Abbas Zaher: [Walid.Zaher@g42.ai](mailto:Walid.Zaher@g42.ai)

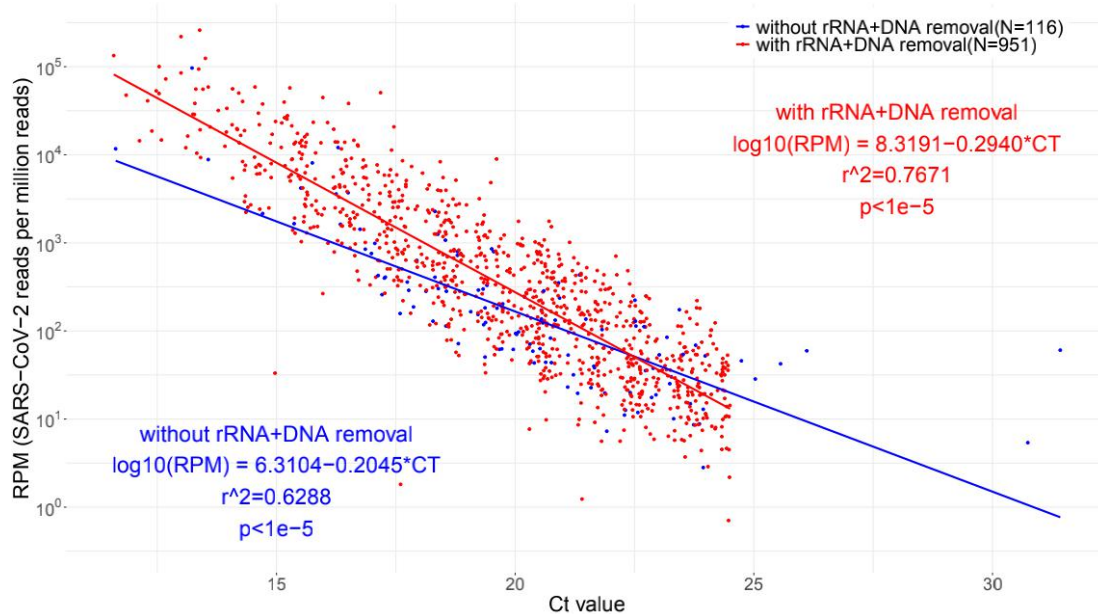
## Table of Contents

<i>Supplementary Figures</i> .....	3
Figure S1. Sequencing quality check .....	3
Figure S2. Correlation between the relative viral load (RPM) estimated by meta-transcriptome sequencing and qRT-PCR Ct value using linear regression .....	4
Figure S3. Gap distribution over the SARS-CoV-2 reference genome .....	5
Figure S4. Density distribution of the number of consensus variants per sample .....	6
Figure S5. Number of consensus variants identified from the 1,067 samples by variant type..	7
Figure S6. Concordance rate of the intra-host variant detection as a function of increasing alternative allele frequency threshold .....	8
Figure S7. Number of consensus variants identified per sample as a function of number of sequenced reads aligned to SARS-CoV-2 .....	9
Figure S8. Identity distribution of the 52 closest relative strains .....	10
Figure S9. Conservativeness of the eleven UAE sub-clade definitive genetic variation .....	11
Figure S10. Molecular dynamics analysis of the E583D and the I780V mutants in the S protein and the Q57H mutant in the ORF3a protein .....	12
Figure S11. Allele frequency changes of the subclade-definitive variants as a function of time .....	13
Figure S12. Allele frequency changes of the mainclade-definitive variants as a function of time .....	13
Figure S13. Density distribution of number of intra-host variation per individual .....	14
Figure S14. Patterns of intra-host genetic variation .....	15
Figure S15. Density distribution of number of intra-host variation per individual .....	16
Figure S16. Base changes as a function of time .....	17
Figure S17. Gene expression of ADAR gene among several issues from GTEX .....	18
<i>Supplementary Tables</i> .....	20
<i>Supplementary Notes</i> .....	21

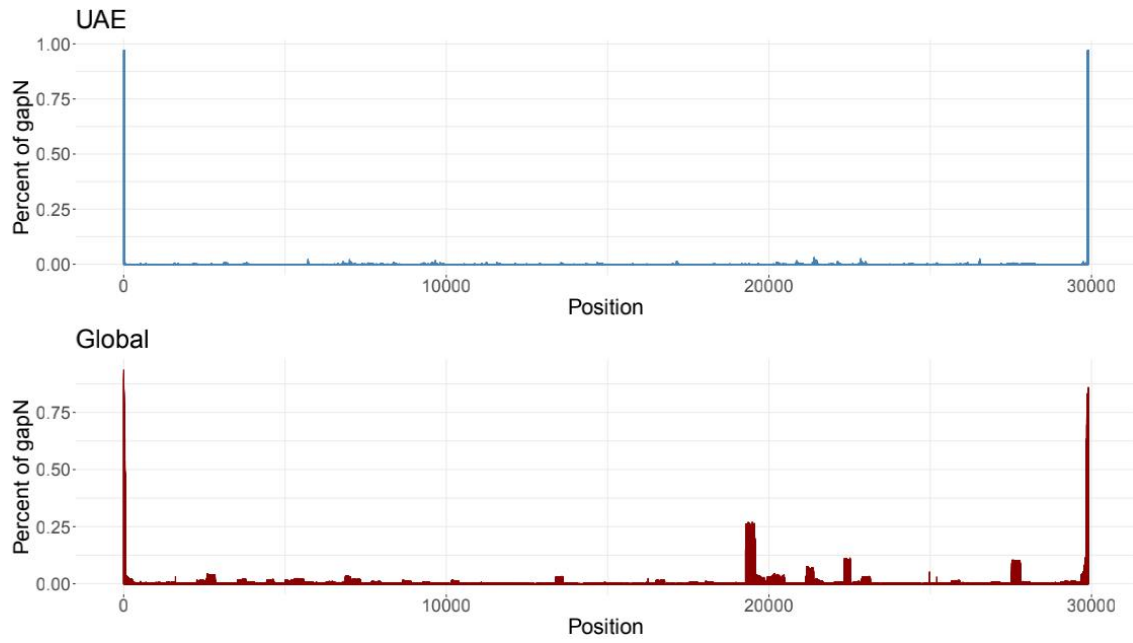
## Supplementary Figures



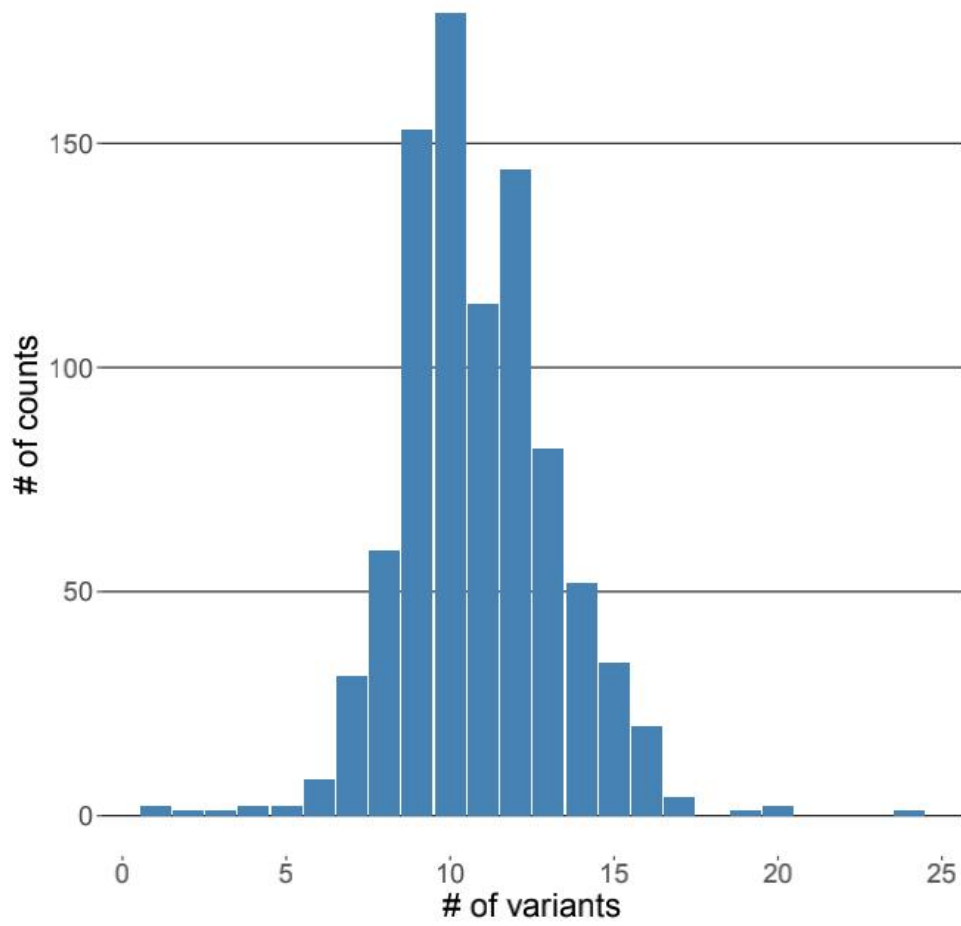
**Figure S1. Sequencing quality check.** (A) Percentage of sequencing bases reaching Q30 for each of the 1,067 samples is shown in orange bar. The base 10 logarithm of the number of sequencing reads and that of the number of reads classified as originated from SARS-CoV-2 is shown in blue and gray line, respectively. (B) The base 10 logarithm of the average sequencing depth of SARS-CoV-2 for each sample is shown in blue bar. The average sequencing depth ranges from 44 to 20,000 fold. X-axis shows the names of the 1,067 samples.



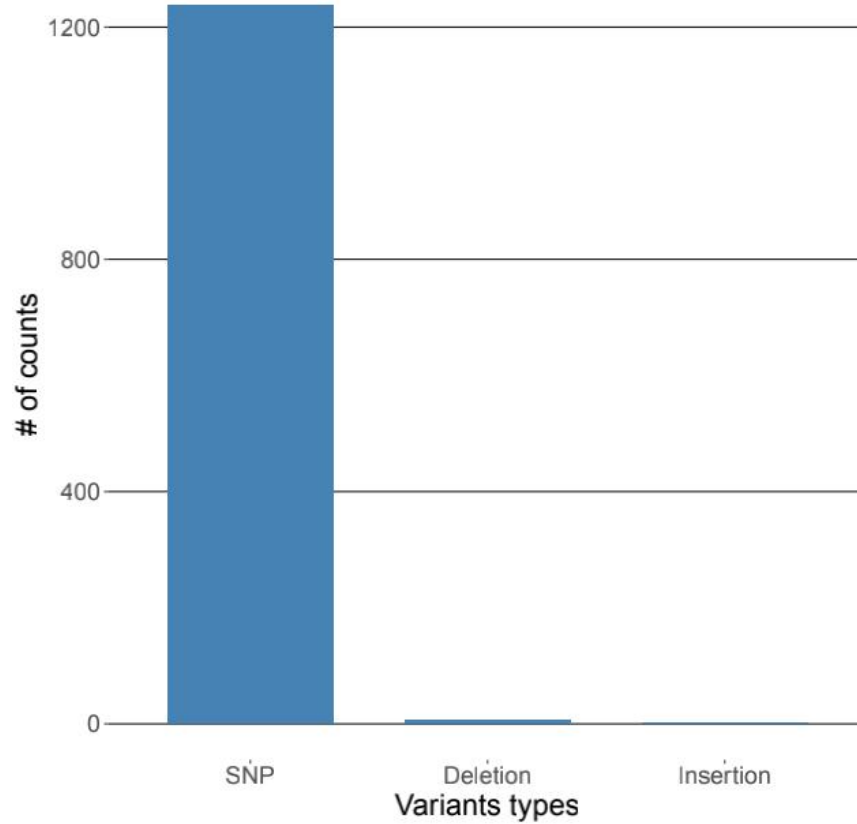
**Figure S2. Correlation between the relative viral load (RPM) estimated by meta-transcriptome sequencing and qRT-PCR Ct value using linear regression.** RPM indicates the number of reads aligned to SARS-CoV-2 reference per million total sequencing reads for each sample. There were two protocols adopted in the meta-transcriptomic sequencing. One without removal of rRNA and depletion of human DNA (N=116) while the other removes rRNA and deplete the human DNA during sequencing library prep (N=951). There is a high correlation between the Ct value and with the SARS-CoV-2 RPM estimated by both protocols ( $R^2 \sim 0.62-0.76$ ). All the 1,067 positive samples have sequencing reads unambiguously aligned to the SARS-CoV-2 reference. All the 3 negative samples do have SARS-CoV-2 reads detected.



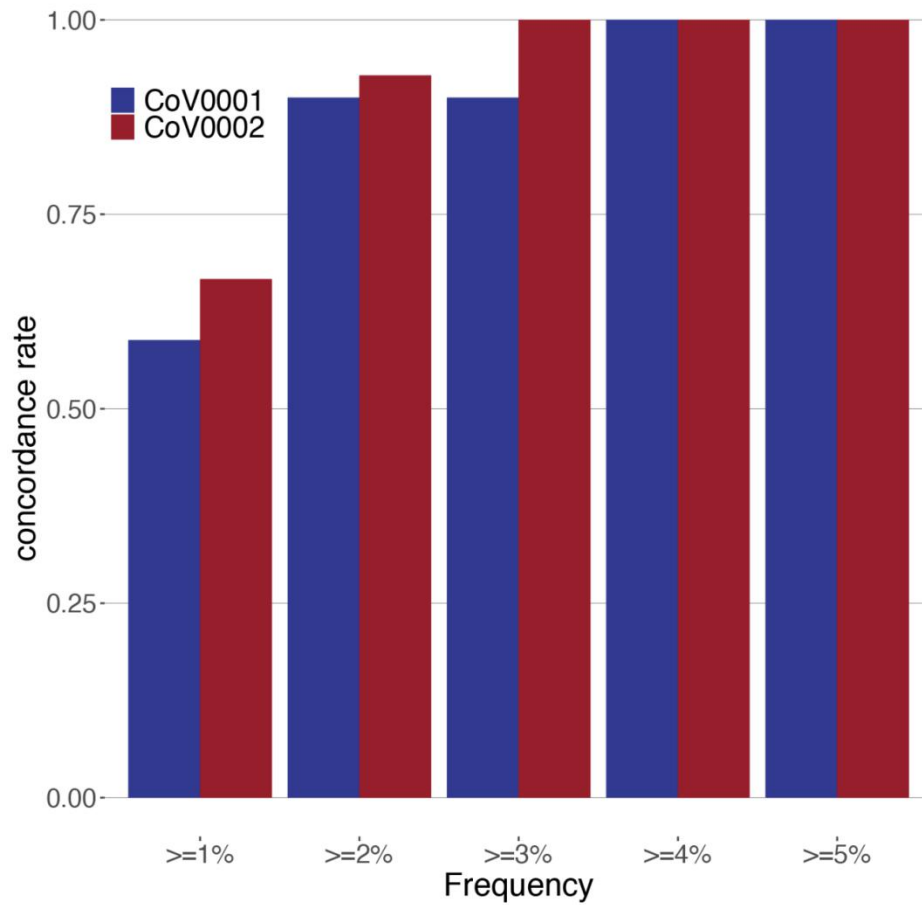
**Figure S3. Gap distribution over the SARS-CoV-2 reference genome.** (Top): gap distribution of the UAE SARS-CoV-2 assemblies reported in this study. (Bottom): gap distribution of the global assemblies reported in GISAID.



**Figure S4. Density distribution of the number of consensus variants per sample.** The variants were compared to the SARS-CoV-2 genome reference (BetaCoV/Wuhan/IVDC-HB-01/2019|EPI\_ISL\_402119).

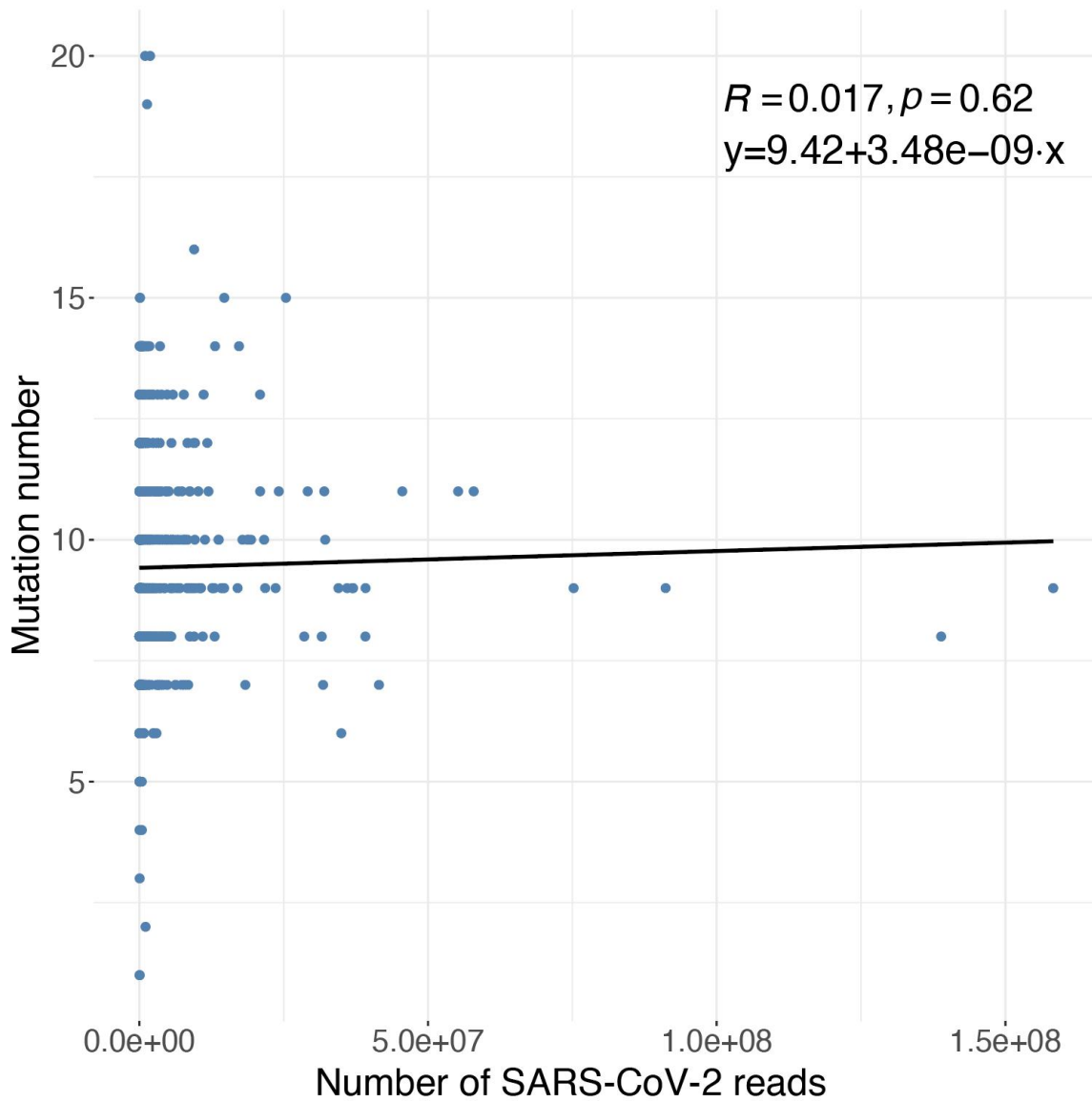


**Figure S5. Number of consensus variants identified from the 1,067 samples by variant type.** The consensus variants were detected by comparison to the SARS-CoV-2 genome reference (BetaCoV/Wuhan/IVDC-HB-01/2019|EPI\_ISL\_402119).

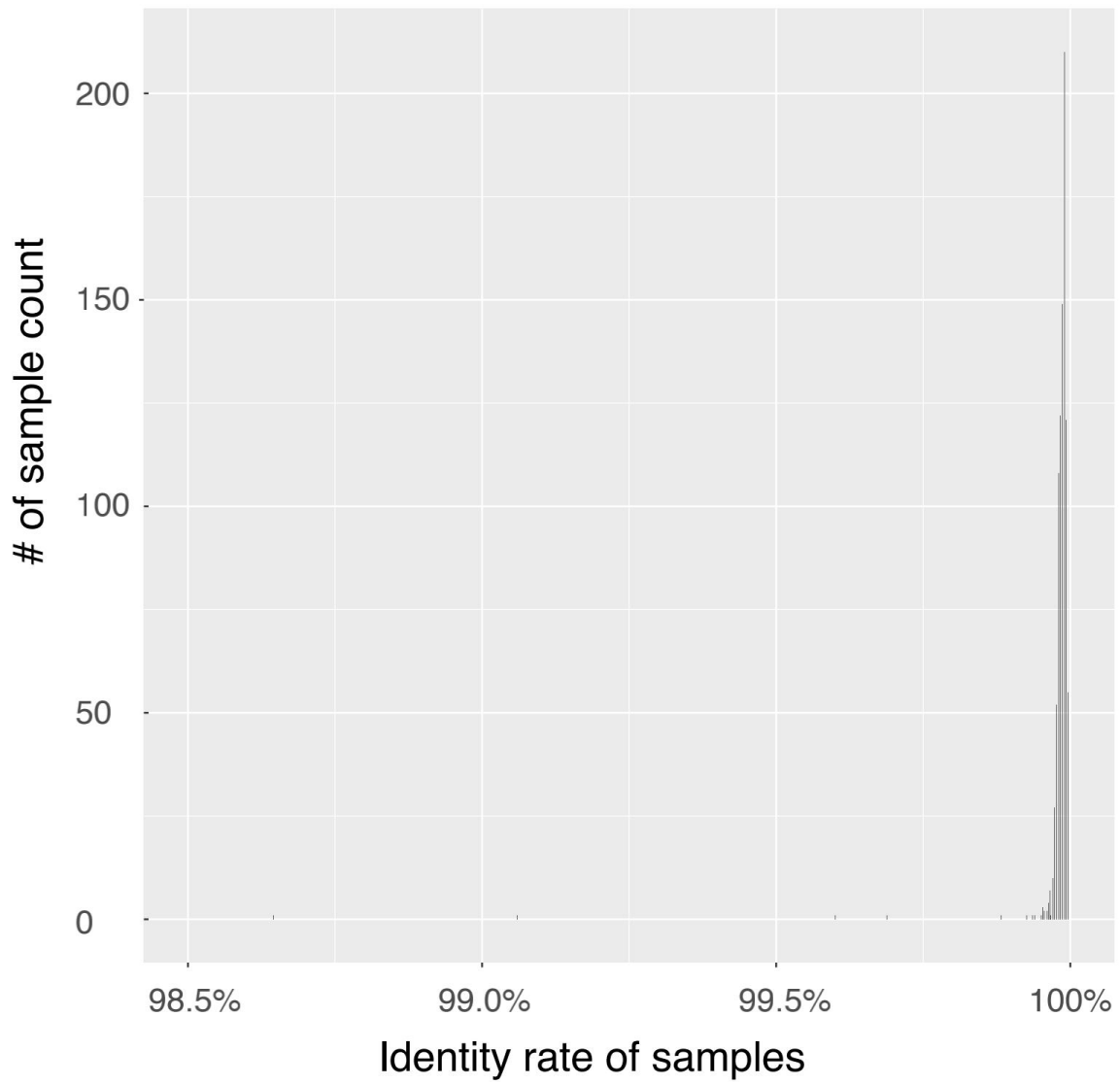


**Figure S6. Concordance rate of the intra-host variant detection as a function of increasing alternative allele frequency threshold.** Concordance rate was computed for the three technical replicates for the two randomly selected samples. Detailed statistics corresponding to this figure can be found in Table S2.



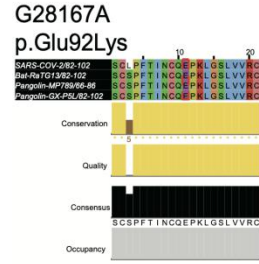
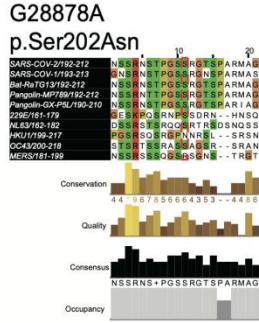
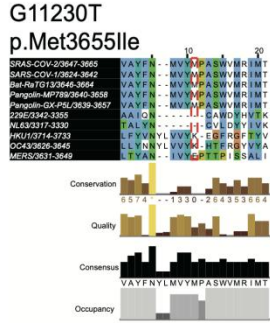


**Figure S7. Number of consensus variants identified per sample as a function of number of sequenced reads aligned to SARS-CoV-2.** The formula was derived from linear regression and the pearson correlation coefficient R was displayed.

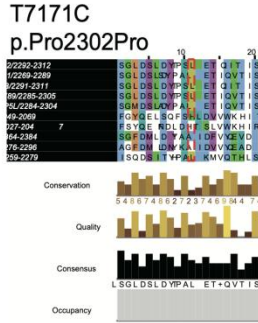
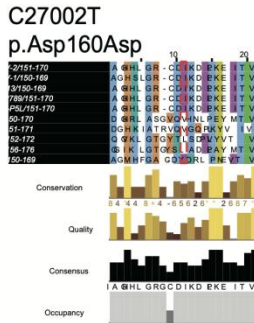


**Figure S8. Identity distribution of the 52 closest relative strains.** Shown is the density distribution of 52 closest relative strains compared with the 637 unique genomes assembled in our study.

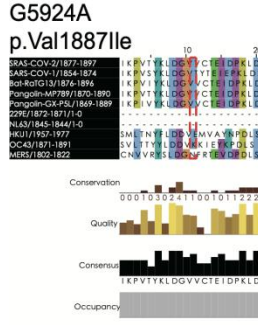
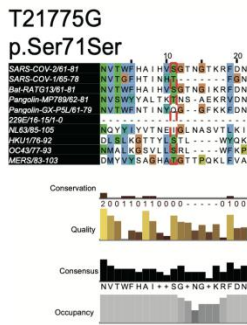
Sub19B.1



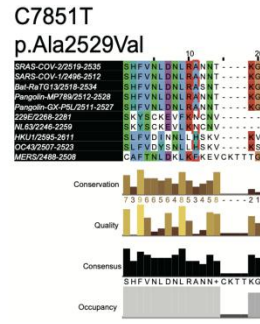
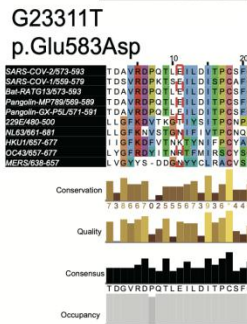
Sub20B.1



Sub20B.2



Sub20B.3



Sub20B.4

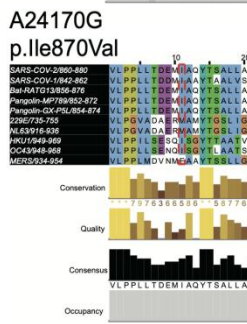
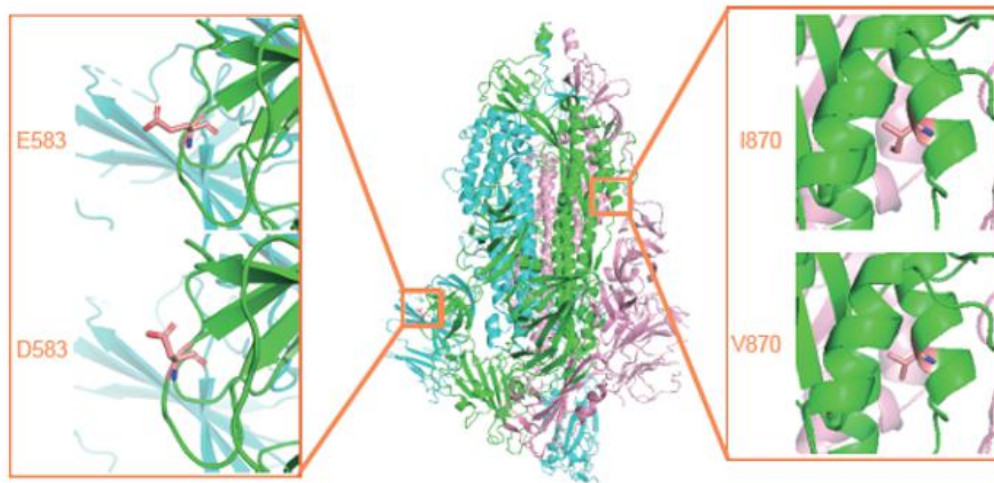
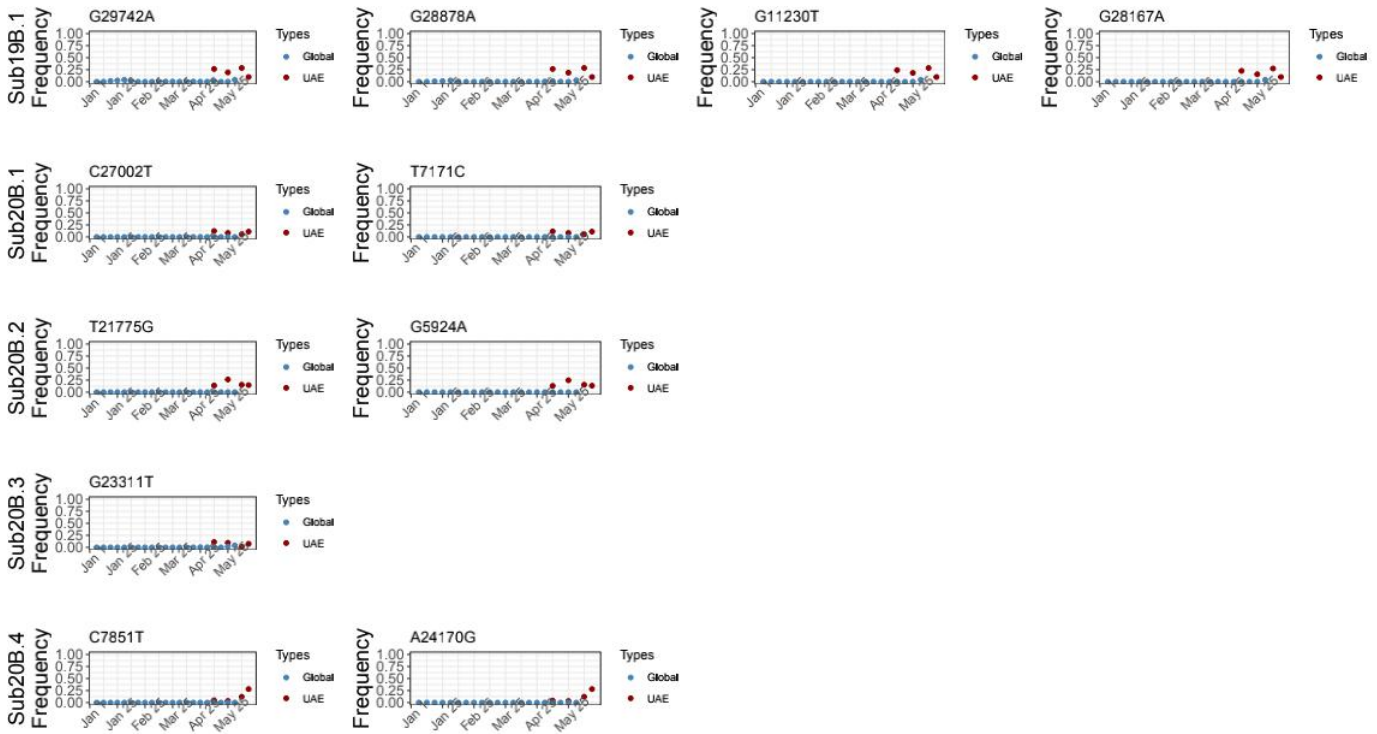


Figure S9. Conservativeness of the eleven UAE sub-clade definitive genetic variation. Shown is the conservation, quality, consensus and occupancy score

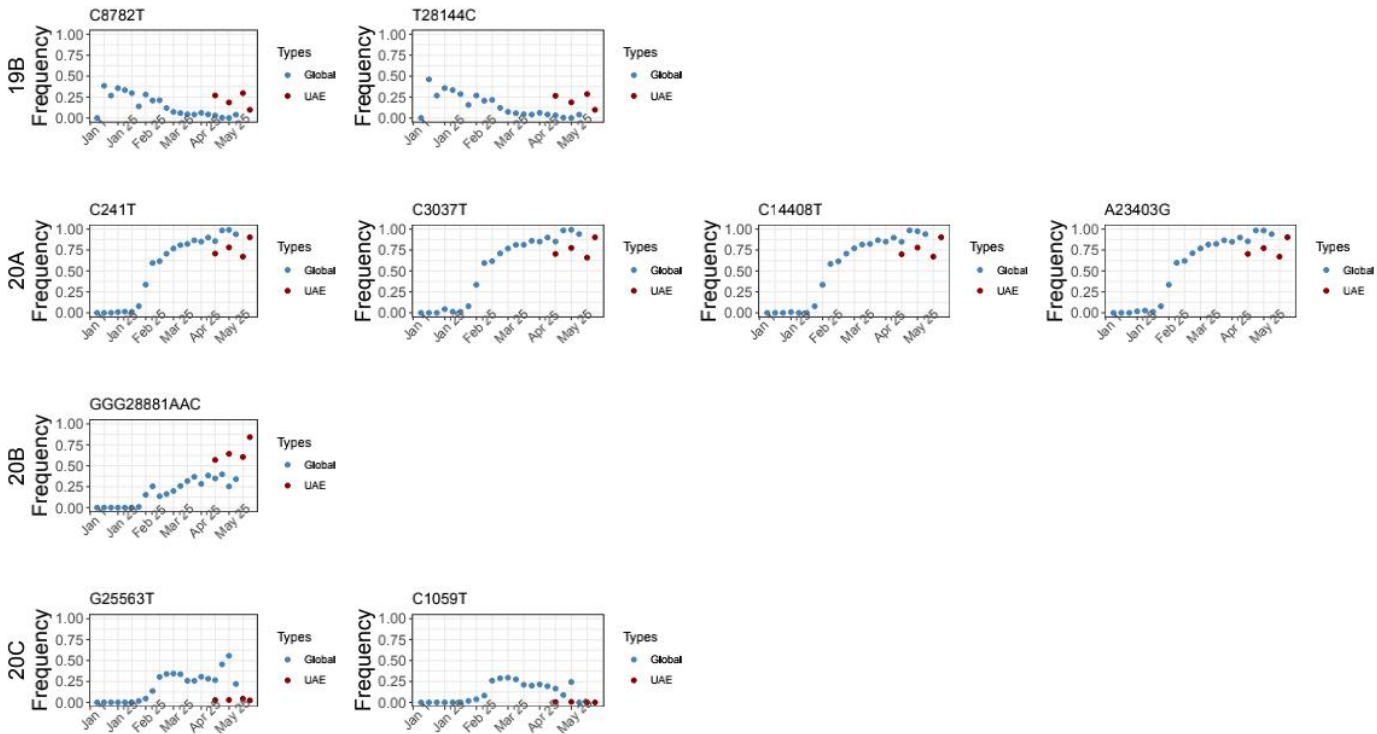
for the amino acid sequence centering on the mutation. The calculation of the score was based on the multiple alignment of seven coronaviruses reference sequences using Jalview.



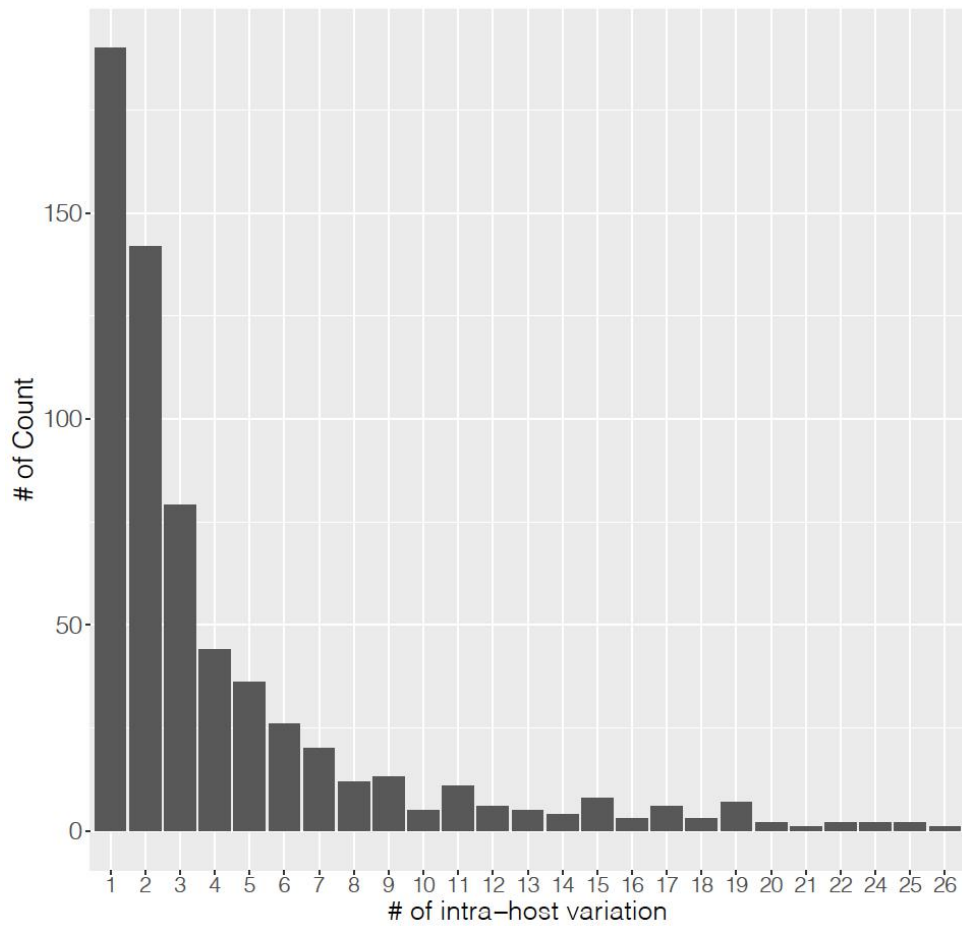
**Figure S10. Molecular dynamics analysis of the E583D and the I780V mutants in the S protein.** Shown are the changes of the protein structure comparing the wild types and the mutants.



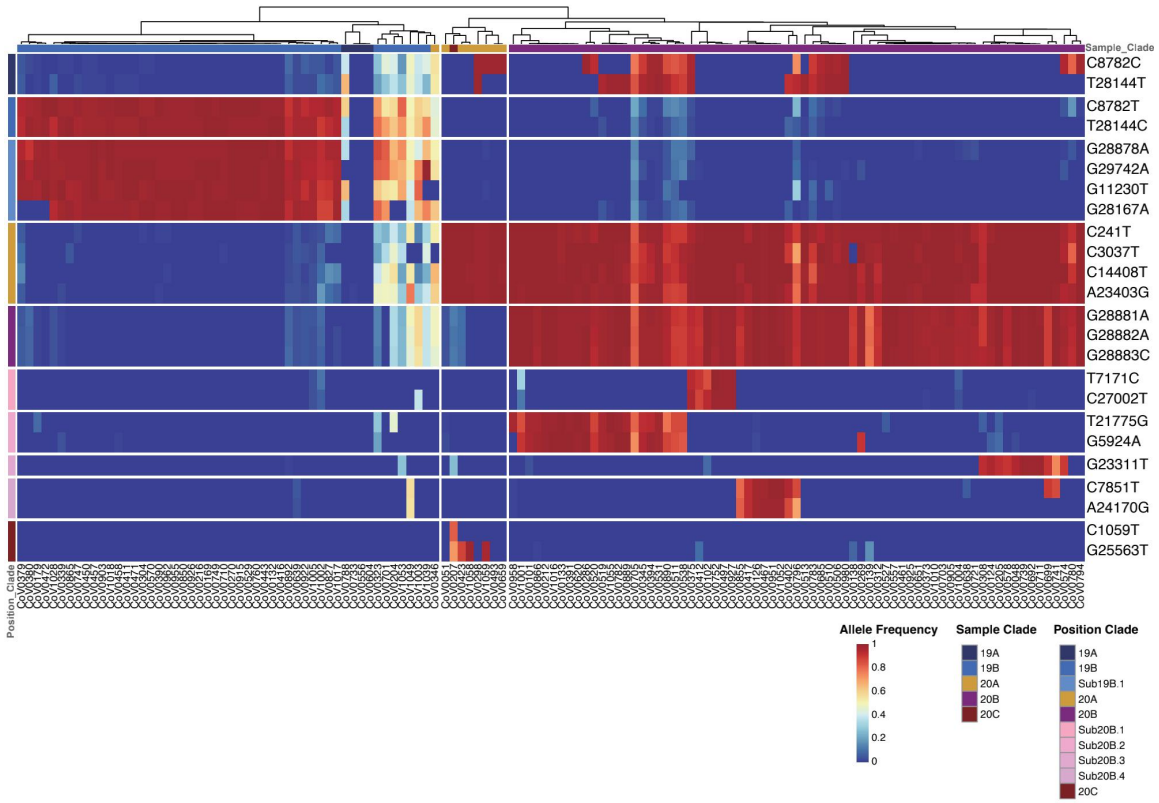
**Figure S11. Allele frequency changes of the subclade-definitive variants as a function of time**



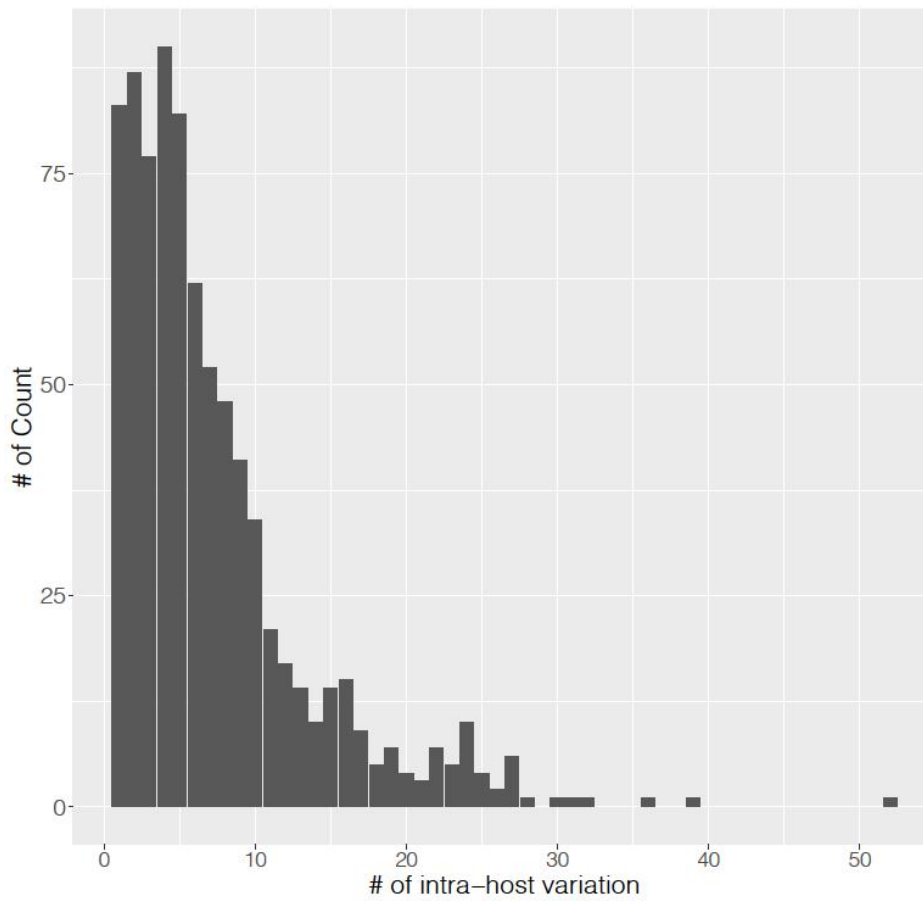
**Figure S12. Allele frequency changes of the clade-definitive variants as a function of time**



**Figure S13. Density distribution of number of intra-host variation per individual.** Number of iSNV with a 5% minor allele frequency threshold and a minimum of 4 minor allele counts was calculated for each individual.

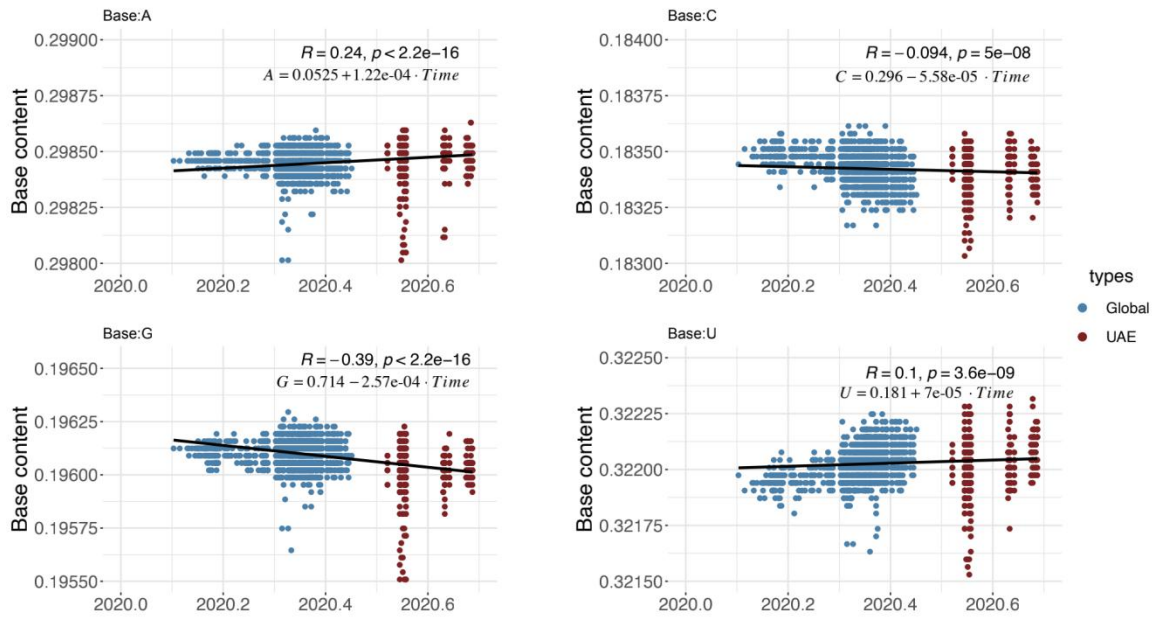


**Figure S14. Patterns of intra-host genetic variation.** The intra-host variation profile when using a 0.05% minor allele frequency threshold and a minimum of 4 minor allele counts per individual.

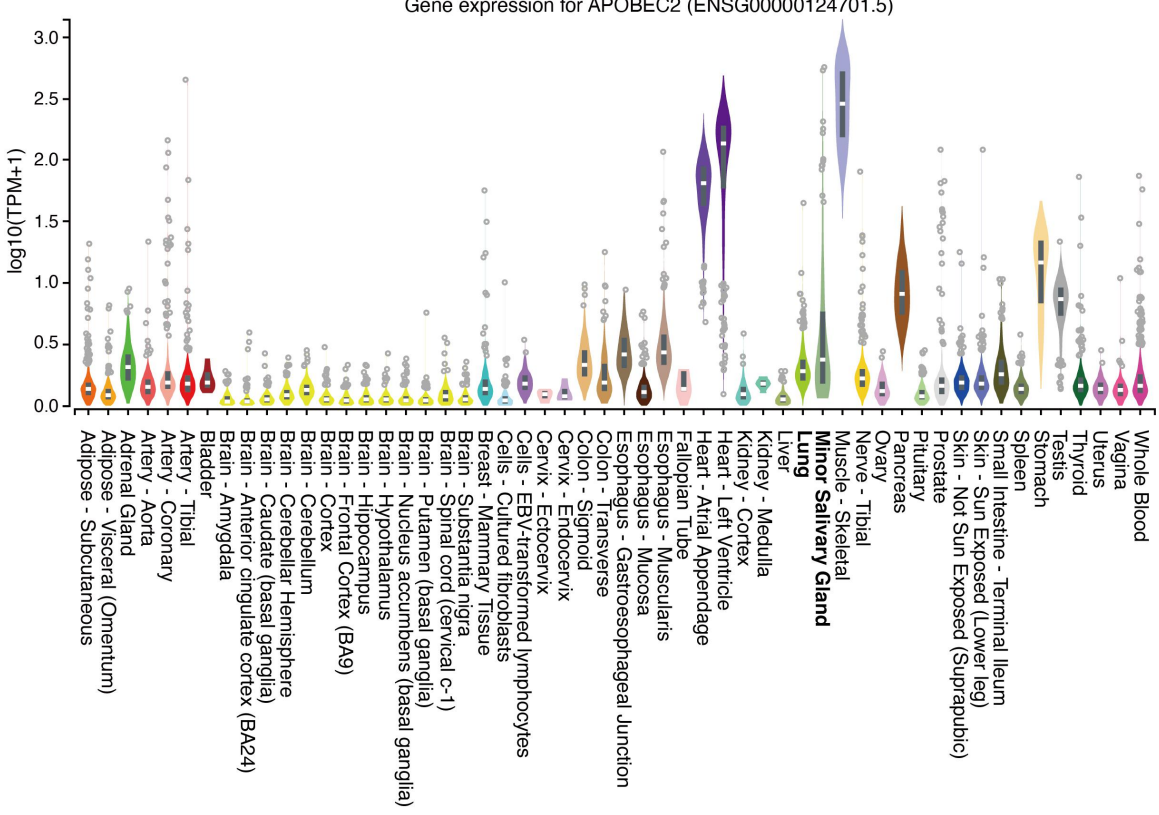
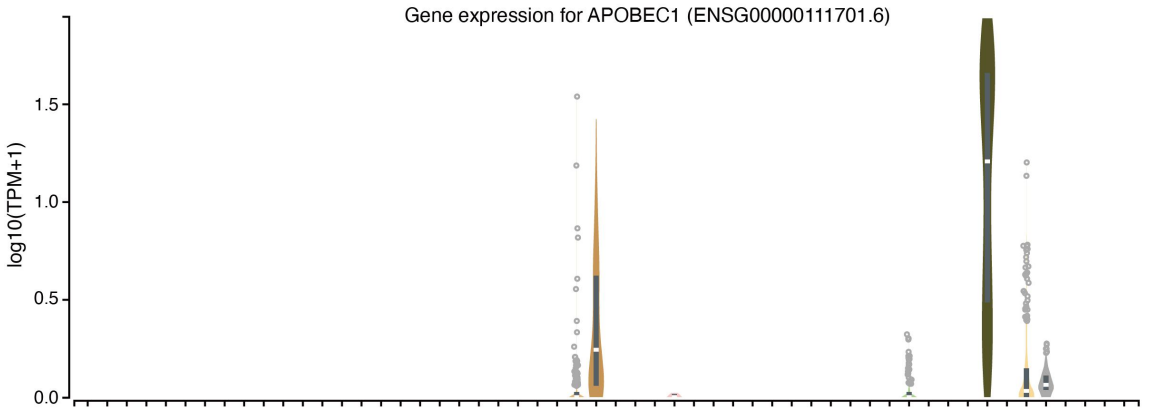
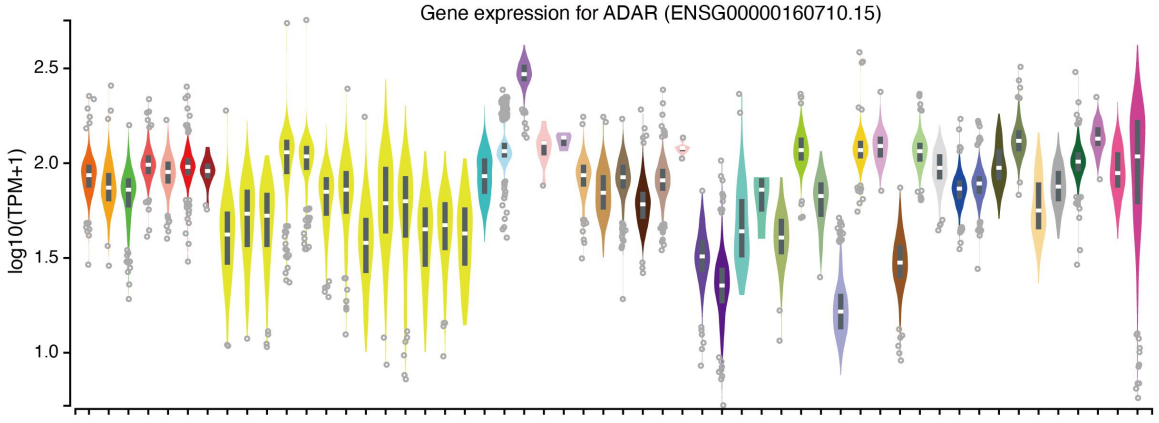


**Figure S15. Density distribution of number of intra-host variation per individual.** Number of iSNV with a 0.05% minor allele frequency threshold and a minimum of 4 minor allele counts was calculated for each individual.

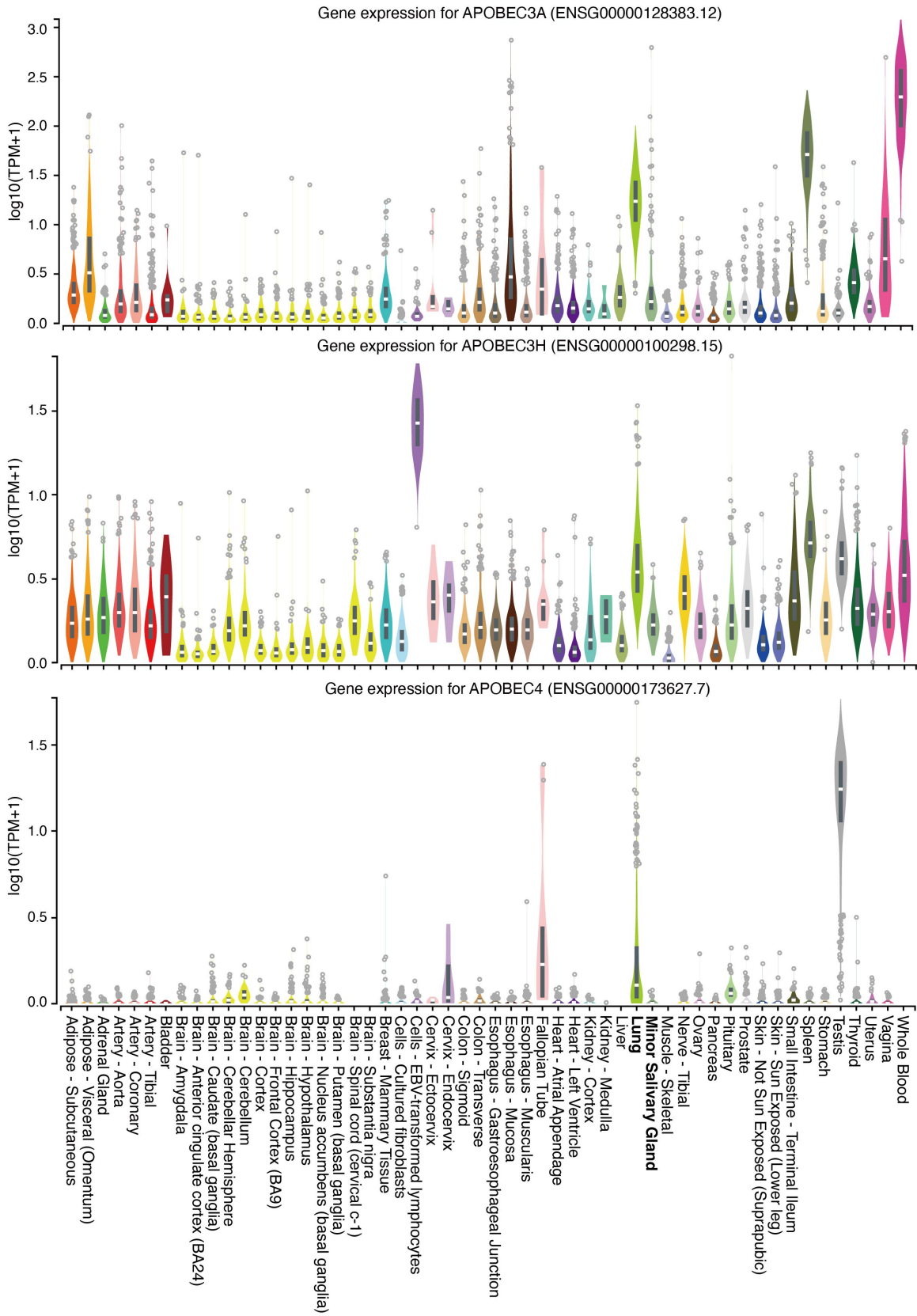




**Figure S16. Dynamic nucleotide composition at the coding regions of SARS-CoV-2 across time.** Each dot represents one sequence. The calculation was based on 2,574 unique SARS-CoV-2 strains isolated from December 24th, 2019 to April 17th, 2020 and the 896 high quality genomes in the UAE.



- Whole Blood
- Vagina
- Uterus
- Thyroid
- Testis
- Stomach
- Spleen
- Small Intestine - Terminal Ileum
- Skin - Sun Exposed (Lower leg)
- Skin - Not Sun Exposed (Suprapubic)
- Pituitary
- Prostate
- Pancreas
- Ovary
- Nerve - Tibial
- Muscle - Skeletal
- Minor Salivary Gland
- Lung
- Liver
- Kidney - Medulla
- Kidney - Cortex
- Heart - Left Ventricle
- Heart - Atrial Appendage
- Fallopian Tube
- Esophagus - Muscularis
- Esophagus - Mucosa
- Esophagus - Gastroesophageal Junction
- Colon - Transverse
- Colon - Sigmoid
- Cervix - Endocervix
- Cervix - Ectocervix
- Cells - EBV-transformed lymphocytes
- Cells - Cultured fibroblasts
- Breast - Mammary Tissue
- Brain - Substantia nigra
- Brain - Spinal cord (cervical c-1)
- Brain - Putamen (basal ganglia)
- Brain - Nucleus accumbens (basal ganglia)
- Brain - Hypothalamus
- Brain - Hippocampus
- Brain - Frontal Cortex (BA9)
- Brain - Cortex
- Brain - Cerebellum
- Brain - Cerebellar Hemisphere
- Brain - Caudate (basal ganglia)
- Brain - Anterior cingulate cortex (BA24)
- Brain - Amygdala
- Bladder
- Artery - Tibial
- Artery - Coronary
- Artery - Aorta
- Adrenal Gland
- Adipose - Visceral (Omentum)
- Adipose - Subcutaneous



**Figure S17. Gene expression of ADAR gene among several issues from GTEX.**

## Supplementary Tables

**Table S1. Number of reads filtered during the analysis and coverage of the viral genome**

**Table S2. Information of the 1245 genetic variants identified from the 896 consensus viral assemblies**

**Table S3. Comparison of variants detected from Freebayes and Pilon and allele frequency estimated from REDITools between three technical replicates from two sample COV20R000001 and COV20R000002**

**Table S4. Samples that shared identical consensus genome sequence with the others**

**Table S5. Demographic data of 152 anonymous samples**

**Table S6. Closest relative in the GISAID database for the 896 high quality viral assemblies**

**Table S7. Molecular dynamic simulation of the non-synonymous variants**

**Table S8. Wilcoxon test of gene expression difference**

## Supplementary Notes

### Nomenclature of the clades and lineages

Pangolin	GISAID	Clades
A	S	19B
B	L	19A
B.2	V	19A
B.1	G	20A
B.1.*	GH	20A/20C
B.1.1.1	GR	20B
B.1.177	GV	20A

### URL— Resource Table

	Select samples randomly	submission	URL
<b>Ruijin</b>	112/112	Submitted to NCBI by Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine	<a href="https://www.ncbi.nlm.nih.gov/bioproject/PRJNA627662">https://www.ncbi.nlm.nih.gov/bioproject/PRJNA627662</a>
<b>Virginia</b>	35/474	Submitted to NCBI by Virginia Division of Consolidated Laboratory Services Sequencing Submission Group	<a href="https://www.ncbi.nlm.nih.gov/bioproject/PRJNA625551">https://www.ncbi.nlm.nih.gov/bioproject/PRJNA625551</a>
<b>Spain</b>	36/244	Submitted to NCBI by FISABIO - Public Health	<a href="https://www.ncbi.nlm.nih.gov/bioproject/PRJEB37513">https://www.ncbi.nlm.nih.gov/bioproject/PRJEB37513</a>
<b>Wuhan</b>	11/11	Di Giorgio et al., Sci. Adv. 2020	<a href="https://www.ncbi.nlm.nih.gov/bioproject/PRJNA601736">https://www.ncbi.nlm.nih.gov/bioproject/PRJNA601736</a> <a href="https://www.ncbi.nlm.nih.gov/bioproject/PRJNA603194">https://www.ncbi.nlm.nih.gov/bioproject/PRJNA603194</a> <a href="https://www.ncbi.nlm.nih.gov/bioproject/PRJNA605907">https://www.ncbi.nlm.nih.gov/bioproject/PRJNA605907</a>
<b>GISAID</b>	23,164		<a href="https://bigd.big.ac.cn/ncov/variation/statistics">https://bigd.big.ac.cn/ncov/variation/statistics</a>