Supplementary Materials for

The twin-beginnings of COVID-19 in Asia and Europe – One prevails quickly

Yongsen Ruan[1], Haijun Wen[1], Mei Hou[1], Ziwen He[1], Xuemei Lu[2], Yongbiao Xue[3], Xionglei He[1], Ya-Ping Zhang[2*], Chung-I Wu[1, 3, 4*]

Correspondence to: wzhongyi@mail.sysu.edu.cn, ciwu@uchicago.edu (Chung-I Wu)

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

Materials and Methods

Virus sequences pre-processing. We download 2,063,459 SARS-CoV-2 genomes from the GISAID database [1] as of July 5, 2021 with the following download options: (1) "complete": genomes > 29,000nt; (2) "low coverage excel": exclude viruses with >5% Ns (undefined base).We filtered out sequences without the collection date or with an implausible date (e.g. '1900-01-05') and retained 1,853,355 genomes for downstream analysis. Note that the meta information (especially the collection date of viruses) of sequences could be changed or removed after several weeks (e.g., the 31 sequences collected on 2020-01-01 in Canada are currently not available). We mark these sequences as deposition irregularity.

Sequence alignment, SNP calling and annotation. We aligned these 1,853,355 genome sequences to the reference sequence (Wuhan-Hu-1[2], GenBank: NC_045512, GISAID: EPI_ISL_402125) using MAFFT (--auto --keeplength) [3]. We used snp-sites (-v) [4] to identify SNPs and BCFtools (merge --force-samples -O v) [5] to merge the vcf files. Interestingly, although the reference genome size is only 29,903 nt, we found 67,650 SNP sites from these 1,853,355 genome sequences, indicating multiple substitutions at the same sites. We annotated the 67,650 SNP sites by ANNOVAR [6].

Analysis of site frequencies in UK. In this study, we track the variant frequency at each variable site (e.g., C→T). If we set the cutoff by ignoring variants that fail to reach 0.3 in frequency, there are usually fewer than 100 variants to keep track of. To observe the changes in site frequencies in a geographic region, we group the data into bins each covering a 10-day period. Only variants reaching the frequency cutoff of 0.3 at their peaks are retained. For the retained variants, we grouped them by pairwise Pearson coefficients with the cutoff of 0.9 (i.e., if the Pearson coefficient across the time span between two variants is equal or greater than 0.9, they would belong in the same group). With this cutoff, there are 5 major waves (W0 to W4) in UK (see Table S1, Fig 1A).

Confirmative re-analysis of the DG group variants. The sequences from GISAID database are mixed in quality and the meta information may be changed or even be removed as mentioned above. To update the meta information of the sequences, we re-downloaded the 56,508 genome sequences as of 2020-03-31 to see the early evolution of the 4 DG group variants (download date: 2021-08-16). Same as previous section, we removed sequences without complete date and remained 49,992 genome sequences of SARS-CoV-2 in human host. We then aligned these sequences to reference sequence (Wuhan-Hu-1) and extracted the four bases of the 4 DG group sites (241, 3037, 14408, and 23403) for each sequence (see Table S7). Because the number of sequences was small in some geographic regions, we separated the sequences by a 10-day (instead of 1-week) gap and then calculated the frequencies of the D614G 4-base haplotypes.

Outgroup sequences for determining the ancestral state of site variants. We downloaded 16 SARS-CoV-2 sequences from bats and pangolins. For comparison, we also downloaded 1 S lineage sequence, 1 L lineage and 12 earliest DG1111 sequences (i.e., TTTT in reference positions 241, 3037, 14408, 23403). We then aligned these 30 sequences (by MAFFT) with reference genome (Wuhan-Hu-1) and used Mega-X [7] to construct the maximum likelihood

phylogenetic tree (Fig. S1). To determine the ancestral state of 4 DG group sites, we extracted the bases on the four sites (Fig. S2). The phylogenetic tree and the haplotype of the four sites suggests that the ancestral state of DG group sites in human host can be DG0000 (CCCA) or DG0100 (CTCA). The analyses are based on DG0000 being the ancestral state while the alternative yields a stronger conclusion (see the main text).

References and Notes

1 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46, doi:10.1002/gch2.1018 (2017).

2 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

3 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

4 Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056, doi:10.1099/mgen.0.000056 (2016).

5 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

6 Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* **10**, 1556-1566, doi:10.1038/nprot.2015.105 (2015).

7 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

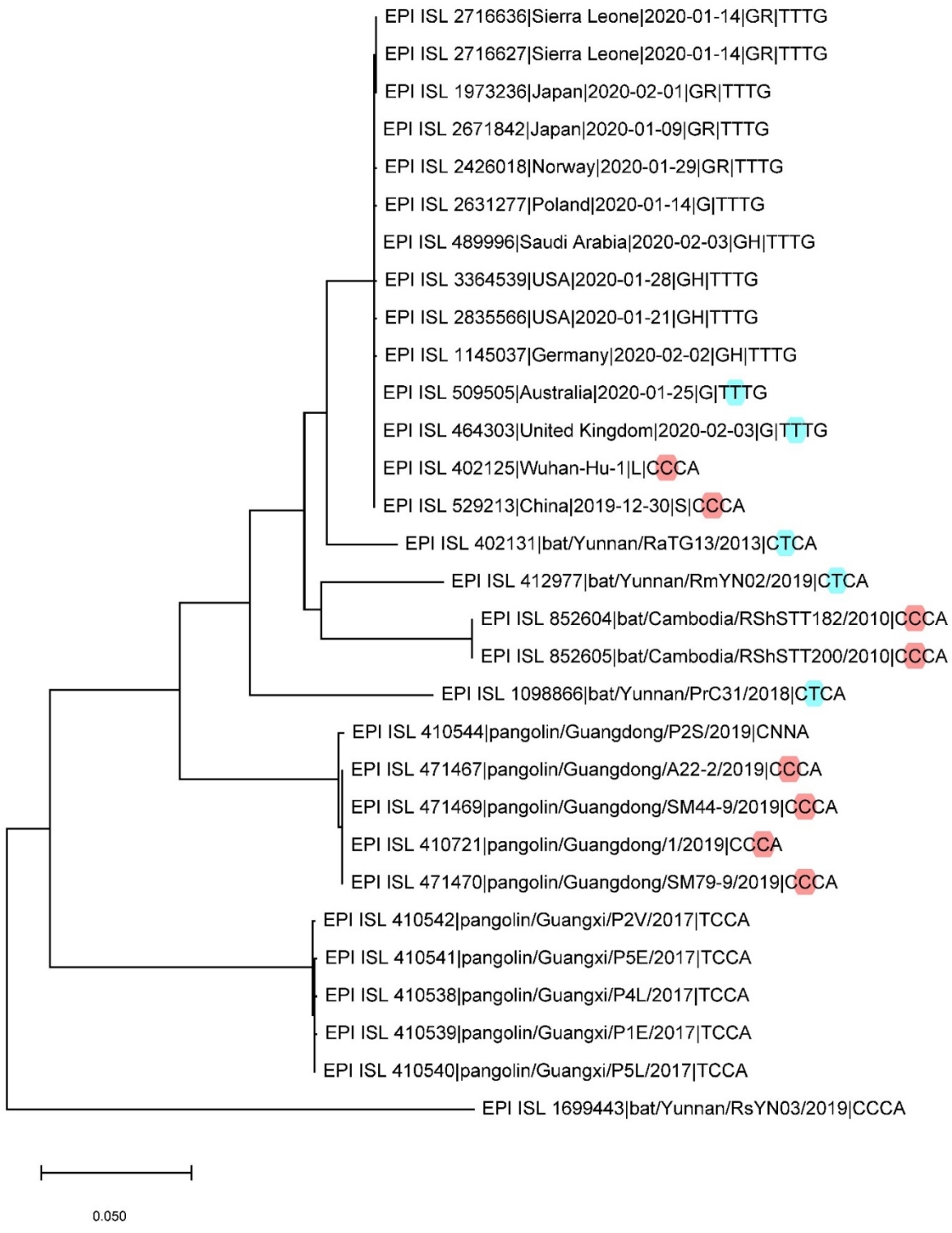

Fig. S1. The maximum likelihood phylogenetic tree of early SARS-CoV-2 sequences in human, bats, or pangolins.

	2	3	1	2	3	8	2	8
	4	0	4	4	4	7	1	4
	4	3	0	8	0	8	4	4
	1	7	8	3	3	2	4	4
EPI ISL 402125|Wuhan-Hu-1|L|CCCA (Reference)	C	C	C	A	G	C	T	C
EPI ISL 2716636|Sierra Leone|2020-01-14|GR|TTTG	T	T	T	G	T	.	.	T
EPI ISL 2716627|Sierra Leone|2020-01-14|GR|TTTG	T	T	T	G	T	.	.	T
EPI ISL 2671842|Japan|2020-01-09|GR|TTTG	T	T	T	G	T	.	.	T
EPI ISL 1973236|Japan|2020-02-01|GR|TTTG	T	T	T	G	T	.	.	T
EPI ISL 2426018|Norway|2020-01-29|GR|TTTG	T	T	T	G	T	.	.	T
EPI ISL 2631277|Poland|2020-01-14|G|TTTG	T	T	T	G	T	.	.	T
EPI ISL 489996|Saudi Arabia|2020-02-03|GH|TTTG	T	T	T	G	T	.	.	T
EPI ISL 3364539|USA|2020-01-28|GH|TTTG	T	T	T	G	T	.	.	T
EPI ISL 2835566|USA|2020-01-21|GH|TTTG	T	T	T	G	T	.	.	T
EPI ISL 1145037|Germany|2020-02-02|GH|TTTG	T	T	T	G	T	.	.	T
EPI ISL 509505|Australia|2020-01-25|G|TTTG	T	T	T	G	T	.	.	T
EPI ISL 464303|United Kingdom|2020-02-03|G|TTTG	T	T	T	G	T	.	.	T
EPI ISL 402125|Wuhan-Hu-1|L|CCCA	.	.	.	.	.	.	.	.
EPI ISL 529213|China|2019-12-30|S|CCCA	.	.	.	.	.	T	C	T
EPI ISL 402131|bat/Yunnan/RaTG13/2013|CTCA	.	T	.	.	.	T	C	T
EPI ISL 412977|bat/Yunnan/RmYN02/2019|CTCA	.	T	.	.	.	T	C	T
EPI ISL 852604|bat/Cambodia/RShSTT182/2010|CCCA	.	.	.	.	.	T	C	T
EPI ISL 852605|bat/Cambodia/RShSTT200/2010|CCCA	.	.	.	.	.	T	C	T
EPI ISL 1098866|bat/Yunnan/PrC31/2018|CTCA	.	T	.	.	.	T	C	T
EPI ISL 410544|pangolin/Guangdong/P2S/2019|CNNA	.	N	N	.	.	T	C	T
EPI ISL 471467|pangolin/Guangdong/A22-2/2019|CCCA	.	.	.	.	.	T	C	T
EPI ISL 471469|pangolin/Guangdong/SM44-9/2019|CCCA	.	.	.	.	.	T	C	T
EPI ISL 471470|pangolin/Guangdong/SM79-9/2019|CCCA	.	.	.	.	.	T	C	T
EPI ISL 410538|pangolin/Guangxi/P4L/2017|TCCA	T	.	.	.	.	T	C	T
EPI ISL 410539|pangolin/Guangxi/P1E/2017|TCCA	T	.	.	.	.	T	C	T
EPI ISL 410540|pangolin/Guangxi/P5L/2017|TCCA	T	.	.	.	.	T	C	T
EPI ISL 410541|pangolin/Guangxi/P5E/2017|TCCA	T	.	.	.	.	T	C	T
EPI ISL 410542|pangolin/Guangxi/P2V/2017|TCCA	T	.	.	.	.	T	C	T
EPI ISL 410721|pangolin/Guangdong/1/2019|CCCA	.	.	.	.	.	T	C	T
EPI ISL 1699443|bat/Yunnan/RsYN03/2019|CCCA	.	.	.	.	.	.	A	.

Fig. S2. The DG 4-base haplotype of early SARS-CoV-2 sequences in human, bats, or pangolins. All the sequences were aligned to reference genome (Wuhan-Hu-1). N represents any base; dot represent the same base as the reference genome.