Novel 2019 coronavirus genome

SARS-CoV-2 coronavirus

Jan 2020
1 / 28
Jan 2020
Feb 2020

edward_holmes 6 Jan '20

> 10th January 2020
> This posting is communicated by Edward C. Holmes, University of Sydney on behalf of the consortium led by Professor Yong-Zhen Zhang, Fudan University, Shanghai

The Shanghai Public Health Clinical Center & School of Public Health, in collaboration with the Central Hospital of Wuhan, Huazhong University of Science and Technology, the Wuhan Center for Disease Control and Prevention, the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control, and the University of Sydney, Sydney, Australia is releasing a coronavirus genome from a case of a respiratory disease from the Wuhan outbreak. The sequence has also been deposited on GenBank (accession MN908947 30.3k) and will be released as soon as possible.

Update: This genome is now available on GenBank and an updated version has been posted 30.3k.

Disclaimer:
Please feel free to download, share, use, and analyze this data. We ask that you communicate with us if you wish to publish results that use these data in a journal. If you have any other questions –then please also contact us directly.

Professor Yong-Zhen Zhang,
Shanghai Public Health Clinical Center & School of Public Health,
Fudan University,
Shanghai, China.

email: zhangyongzhen@shphc.org.cn

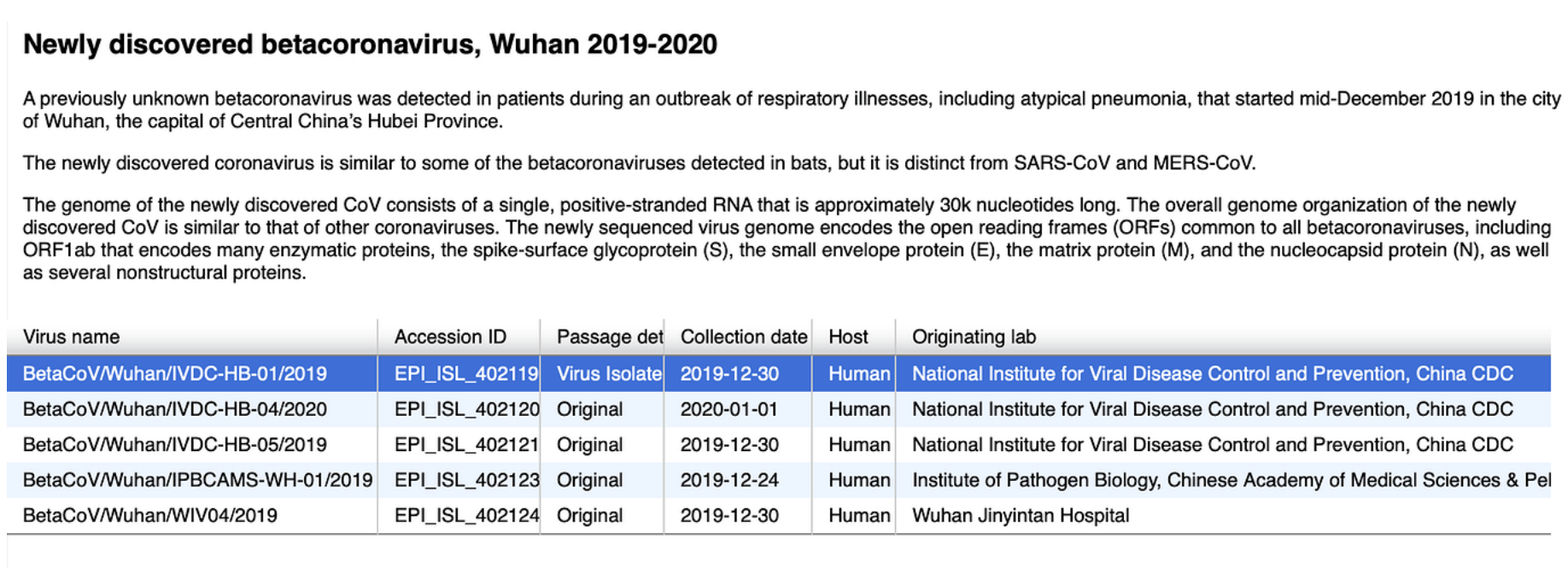
1

Initial assessment of the ability of published coronavirus primers sets to detect the Wuhan coro… 576
SARS-CoV-2 diversity in Uganda, December, 2020 23

created Jan '20 | last reply Feb '20 | 27 replies | 269k views | 9 users | 9 likes | 14 links | 7 6 5

arambaut ARTIC Network Jan '20

Five new genomes have been deposited in the GISAID platform:
https://gisaid.org/CoV2020 3.6k



Newly discovered betacoronavirus, Wuhan 2019-2020

A previously unknown betacoronavirus was detected in patients during an outbreak of respiratory illnesses, including atypical pneumonia, that started mid-December 2019 in the city of Wuhan, the capital of Central China's Hubei Province.

The newly discovered coronavirus is similar to some of the betacoronaviruses detected in bats, but it is distinct from SARS-CoV and MERS-CoV.

The genome of the newly discovered CoV consists of a single, positive-stranded RNA that is approximately 30k nucleotides long. The overall genome organization of the newly discovered CoV is similar to that of other coronaviruses. The newly sequenced virus genome encodes the open reading frames (ORFs) common to all betacoronaviruses, including ORF1ab that encodes many enzymatic proteins, the spike-surface glycoprotein (S), the small envelope protein (E), the matrix protein (M), and the nucleocapsid protein (N), as well as several nonstructural proteins.

Virus name	Accession ID	Passage det	Collection date	Host	Originating lab
BetaCoV/Wuhan/IVDC-HB-01/2019	EPI_ISL_402119	Virus Isolate	2019-12-30	Human	National Institute for Viral Disease Control and Prevention, China CDC
BetaCoV/Wuhan/IVDC-HB-04/2020	EPI_ISL_402120	Original	2020-01-01	Human	National Institute for Viral Disease Control and Prevention, China CDC
BetaCoV/Wuhan/IVDC-HB-05/2019	EPI_ISL_402121	Original	2019-12-30	Human	National Institute for Viral Disease Control and Prevention, China CDC
BetaCoV/Wuhan/IPBCAMS-WH-01/2019	EPI_ISL_402123	Original	2019-12-24	Human	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Pel
BetaCoV/Wuhan/WIV04/2019	EPI_ISL_402124	Original	2019-12-30	Human	Wuhan Jinyintan Hospital

trvrb ARTIC Network Jan '20

A preliminary phylogenetic analysis of these 6 genomes is available at https://nextstrain.org/groups/blab/sars-like-cov 2.4k. The pipeline to generate this analysis is openly available at https://github.com/blab/sars-like-cov 949.

arambaut ARTIC Network Jan '20

Thanks @trvrb. A word of caution about interpreting this tree. I am almost certain that the divergent sequence `IVDC-HB-05/2019` is divergent because of sequencing and assembly artefacts. I strongly suggest not making any epidemic inferences from the 6 genomes available at the moment.

I have contacted the authors of this sequence but have not had a reply yet.

1

arambaut ARTIC Network Jan '20

The SNPs in `IVDC-HB-05/2019` are majority non-synonymous and non-sense:

`IVDC-HB-04/2020` is also suspect - it has 5 non-synonymous mutations and 3 synonymous:

arambaut ARTIC Network Jan '20

`IVDC-HB-01/2019` has been cell cultured with one round of passaging. This should be considered the most reliable. It may could have cell adaptations but it is identical to `WIV04/2019` which is direct sequenced so if independent, suggests there are no cell adaptations.

The first genome `WH-Human_1` has one SNP difference from all the others which may mean it is real. However it is not known if this genome is from a sample from one of the same patients as the other 5.

richard.neher 1 Jan '20

> `IVDC-HB-04/2020` is also suspect - it has 5 non-synonymous mutations and 3 synonymous

The nextstrain tool-tip is misleading here. The reference used has over-lapping annotations ORF1a and ORF1ab. There is a total of 3 mutations inferred for this branch. C1023T, C1025T, A18460G
The first two change the aa sequence of ORF1a in coding 253 and 245 (and these are the same as the mutations listed in ORF1ab).

I was mistaken. This is wrong:

> The last mutation is synonymous also in ORF1ab after the slippage site.
> So: 2 adjacent non-synonymous, 1 synonymous.

Correction:
The last mutation at A18460G is also non-synonymous. All three mutations are non-synonymous

cupton 1 Jan '20

HB-04 has a bunch of indels.

Use Base-By-Base 304 to view the alignment (visual sumary shows all diffs)

1 Reply

trvrb ARTIC Network 1 Jan '20

Thanks for the feedback Andrew and Richard. I've updated https://nextstrain.org/groups/blab/sars-like-cov 414 to split ORF1a and ORF1b. This makes it clearer how nucleotide mutations map to amino acid substitutions.

Ignoring the divergent BetaCoV/Wuhan/IVDC-HB-05/2019 sequence and masking the initial 11 bases of the alignment, we have the following 5 strains and their mutations relative to the base of the outbreak clade:

- WIV04/2019 - no mutations
- IVDC-HB-01/2019 - no mutations
- IPBCAMS-WH-01/2019 - 3 nucleotide mutations / 2 AA changes
- IVDC-HB-04/2020 - 3 nucleotide mutations / 3 AA changes (includes C1023T and C1025T which are suspect being so close together)
- WH-Human_1 - 2 nucleotide mutations / 1 AA change

This alignment was stripped to map to reference https://github.com/blab/sars-like-cov/blob/master/config/sars-like-cov_reference.gb 344 and so lacks indels.

arambaut ARTIC Network → cupton Jan '20

The single basepair gaps are in homopolymeric runs suggesting a sequencing platform that maybe has problems with those. There are 2 larger deletions which I assume are missing read coverage.

1 Reply

Kristian_Andersen 1 Jan '20

I get slightly different stats on the # mutations - HB-04 has some indels that need corrections. Keeping HB-01 as the reference (should maybe be WH-01 though, as that's the oldest sequence):

IVDC-HB-01/2019: [ref]
IPBCAMS-WH-01/2019: 3 mutations (2 non-syn / 1 syn)
WIV04/2019: 0 mutations
Hu-1/2019: 1 mutation (1 non-syn)
IVDC-HB-04/2020: 2 mutations (2 non-syn) (however, I don't believe these, so I think this should also be 0 mutations)

I agree with Trevor that the mutations in HB-04 are suspect - right next to each other, non-synonymous, close to a poly-T stretch, and this sequence also needed some manual editing for indels. I think these are probably not correct and that sequence would then also be identical.

As for IVDC-HB-05, I agree with everybody that this sequence is definitely wrong (clustering of mutations, wacky ts/tv ratio, etc). If I do my very best to eliminate sequencing errors that I have commonly observed over the years, then I get a maximum of 7 mutations in this sequence, 4 of which are non-synonymous. These 7 can't be excluded as likely errors (unlike the other 46 mutations in this sequence), but I think they still represent a (substantial) over-estimation.

1

cupton → arambaut Jan '20

Or poor assembly.
Think the coverage is so low that regions are missing?

arambaut ARTIC Network Jan '20

I think it would be unlikely that you would get zero coverage just for 1 basepair. The fact that they are in homopolymeric runs suggests systematic run-length errors. This is probably not Illumina data.

kihohong 2 Jan '20

Several sequences including Thailand cases has been added to GISAID today.
Although GISAID announced their whole genome analysis result, I am wondering if you would update your analysis, since your analysis provide much more information including diversities.
Sincerely.

trvrb ARTIC Network Jan '20

https://nextstrain.org/ncov 652 has been updated with all genomes currently in GISAID.

1

trvrb ARTIC Network Jan '20

The Zhejiang Provincial Center for Disease Control and Prevention has shared two new genomes via gisaid.org 83. We've updated https://nextstrain.org/ncov 218 to include them in our analysis bringing total up to 15 highly related samples.

Kristian_Andersen 1 Jan '20

Four more genomes were released, bringing the total to 19. Note that a couple of these look suspicious:

EPI_ISL_403928: A lot of mutations - can't be trusted at this stage
EPI_ISL_403931: Mutations in the 5' end that are wrong

Still not a lot of diversity.

Based on this dataset I count 17 SNPs that appear to be real and 35 that do not (this is not including indels in 402120). All SNPs are private - none of them transmitted.

1

trvrb ARTIC Network Jan '20

Thanks @Kristian_Andersen. We've updated https://nextstrain.org/ncov 282 accordingly, but have also left out Wuhan/IPBCAMS-WH-05/2020 / EPI_ISL_403928 due to appearance of spurious mutations.

However, I'd note that if there's been multiple spillover events from the animal reservoir, I would really expect to be seeing clusters of genetically distinct human cases. So, a divergent sequence by itself is an expected thing. What threw me here however, was strange clustering of mutations in Wuhan/IPBCAMS-WH-05/2020.

1 Reply 1