SARS-CoV-2 Samples from Same Early COVID-19 Patients Were Sequenced Repeatedly with Errors Distorting Phylogenetic Trees

When the bat RaTG13 coronavirus is used as the outgroup for a SARS-CoV-2 phylogenetic study, the resulting phylogenetic tree is rooted near a virus strain isolated from the first patient in the United States [1]. The patient had traveled to and from Wuhan, but the strain has not yet been found in the city. It was postulated that there had been a lack of sequencing efforts in early days of the outbreak, and that newer strains gained competitive advantage during later days. The hypothesis has aroused substantial public interests and debates. Hence it is worthwhile to understand the availability and quality of viral genomic sequences from early patients.

In this exercise, I examined SARS-CoV-2 genomic sequences from China National Center for Bioinformation (CNCB, https://bigd.big.ac.cn/ncov/network?lang=en) [2]. It incorporates data from many sources, including pointers to GISAID, and offers convenient data download and a user-friendly visualization of haplotype trees as shown in Fig 1.

The CNCB site returned 24 complete SARS-CoV-2 genomic sequences when limited to samples collected on or before Jan 1, 2020 in Wuhan, with data released by March 8, 2020. There were only 41 confirmed SARS-CoV-2 patients publicly reported by Jan 1, 2020 [3]. Hence age and gender information is mostly sufficient for matching patient identity, unless rejected by published orthogonal information. GISAID was used to retrieve the patient age, gender, sequencing platform and assembly method for each sequence. The combined information is tabulated in Fig 2. Excluded from the study is the grey row for an incomplete genome, and the *italic* rows for suspected duplicated submissions with identical end indel variants. A 'U' in the 'AgeGender' column means Unknown. The 41M1 and 41M2 indicates two separate 41-year-old male patients, differentiated by admission dates and whether they work in the Huanan Seafood Wholesale Market where the outbreak had started [4, 5].

The SARS-CoV-2 sample isolated from a 49-year-old female (49F) patient was sequenced at least 5 times, with three other individuals (52F, 61M and 32M) each contributing to at least 2 sequences. The grouping by 49F, 52F and 61M is confirmed by independent news articles [6, 7]. While repeated sequencing facilitates the following error assessment, such duplicated efforts shrunk the small data size even further for any meaningful SARS-CoV-2 early evolutionary or epidemiology study. The impact on a haplotype tree is shown in Fig 3.

Sequencing and/or assembly errors were evaluated conservatively with these rules, with end indels ignored:

1. When there are only two genomic sequences of a sample, their divergence is the total error count of the two sequences.
2. Because of the high quality and wide acceptance of the SARS-CoV-2 reference Wuhan-Hu-1, a genomic sequence is considered error-free if it is a perfect match to the reference.
3. Sequences that are perfectly matched to each other and not duplicated submissions are considered error-free.
4. Once a first error-free sequence of a strain is established, other sequences of the same sample are measured by divergence from the first sequence.
5. Sequences that cannot be evaluated by the above rules were abstained.

As shown in Fig 4, in the 16 genomic sequences evaluated, at least 17 errors were counted in 7 (44%) sequences. Except for the rule 1, the errors can be located and compared to variants in other strains. All located errors are unique to the affected sequences. The sequence with the most (6) errors was flagged by CNCB's QC pipeline as containing densely clustered mutations. The farthest tip in Fig. 4 is also flagged but abstained from evaluation.

In summary, samples from at least 4 early COVID-19 patients in Wuhan were sequenced repeatedly, with 1-6 errors in estimated 44% of the resulting genomic sequences. The errors can cause apparent false branching of phylogenetic trees. The finding supports the necessity of sequencing more early strains with higher quality in order to trace the evolution of SARS-CoV-2.

Acknowledgement

I gratefully acknowledge the Authors, the Originating and Submitting Laboratories for their sequences, metadata and tools shared through CNCB and GISAID, on which this research is based.

References:

1. Yu, Wen-Bin, Tang, Guang-Da, Zhang, Li,Corlett, Richard T..(2020).Decoding evolution and transmissions of novel pneumonia coronavirus using the whole genomic data.[ChinaXiv:202002.00033]
2. Zhao WM, Song SH, Chen ML, et al. The 2019 novel coronavirus resource. Yi Chuan. 2020;42(2):212–221. doi:10.16288/j.yczz.20-030 [PMID: 32102777]
3. Huang, C., Wang, Y., Li, X., et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020; 395: 497–506.
4. Wu, F., Zhao, S., Yu, B. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020). https://doi.org/10.1038/s41586-020-2008-3
5. Li-Li Ren, Ye-Ming Wang, *et al.* Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chinese Medical Journal (English)*: February 11, 2020. doi: 10.1097/CM9.0000000000000722
6. Caixin.com 2020-02-27. Tracking gene sequencing of the novel coronavirus: when did the alarm go off? (财新网: 新冠病毒基因测序溯源：警报是何时拉响的 https://new.qq.com/omn/TWF20200/TWF2020022701654200.html)
7. Southern Weekly 2020-03-05. "Restructuring" Jinyintan: Secrets in the Eye of Strom (南方周末：“重组”金银潭：疫情暴风眼的秘密 http://infzm.com/contents/178385)

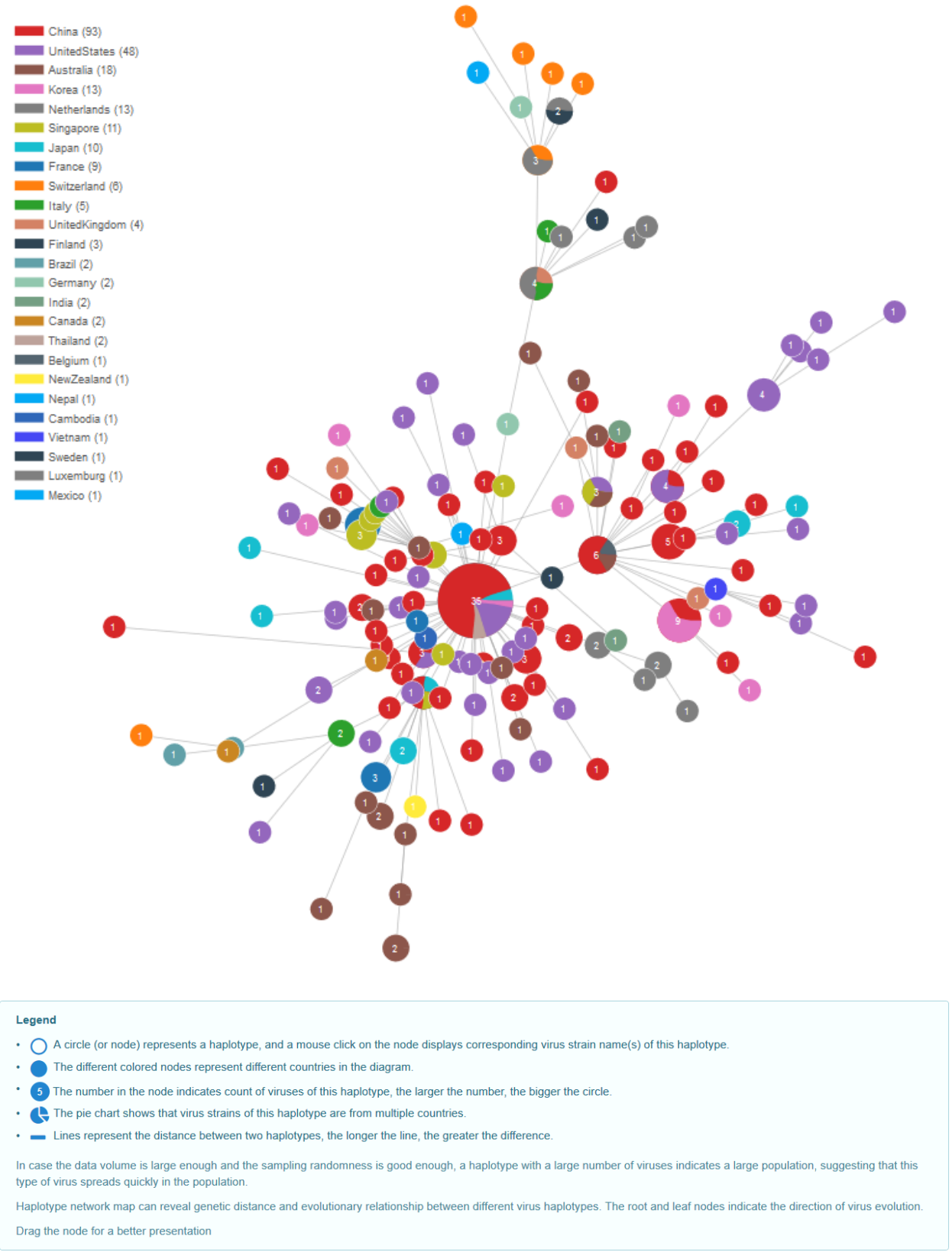

Figure1. The full haplotype tree on CNCB SARS-CoV-2 web page.

AgeGender	Virus Name 病毒株名	Serial Number 序列号	Related IDs 相关ID	Sequence Completeness	Sample Date	Sample Provider 样本提供单位	Submission Time	Submitter 数据递交单位
21F	BetaCoV/Wuhan/WH-03/2019	CNA0007334	LR757996, EPI_ISL_406800, NMDC60013002-03	Complete	2020-01-01	General Hospital of Central Theater Command of People's Liberation Army of China	2020-01-30	BGI PathoGenesis Pharmaceutical Technology Co., Ltd; China CDC; Shandong First Medical University & Shandong Academy of Medical Sciences; Hubei Provincial Center for Disease Control and Prevention
32M	WIV02	GWHABKK00000000	EPI_ISL_402127, MN996527	Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-01-18	CAS Key Laboratory of Special Pathogens and Biosafety and Center for Emerging Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences
32M	BetaCoV/Wuhan/IVDC-HB-05/2019	EPI_ISL_402121		Complete	2019-12-30	National Institute for Viral Disease Control and Prevention, China CDC	2020-01-10	National Institute for Viral Disease Control and Prevention, China CDC
32M	*BetaCoV/Wuhan/WH19005/2019*	*NMDC60013002-10*		*Complete*	*2019-12-30*		*2020-01-28*	*China CDC; Shandong First Medical University & Shandong Academy of Medical Sciences; Hubei Provincial Center for Disease Control and Prevention; BGI PathoGenesis Pharmaceutical Technology Co., Ltd*
40M	WIV06	GWHABKN00000000	EPI_ISL_402129, MN996530	Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-01-18	CAS Key Laboratory of Special Pathogens and Biosafety and Center for Emerging Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences
41M1	BetaCoV/Wuhan/IPBCAMS-WH-03/2019	GWHABKH00000000	EPI_ISL_403930, MT019531	Complete	2019-12-30	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	2020-01-21	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
41M2	Wuhan-Hu-1	MN908947	NC_045512, EPI_ISL_402125	Complete	2019-12-30		2020-01-17	Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China
43M	BetaCoV/Wuhan/WH-02/2019	CNA0007333	LR757997, EPI_ISL_406799, NMDC60013002-02	Partial/scaffold level	2019-12-31	General Hospital of Central Theater Command of People's Liberation Army of China	2020-01-30	BGI PathoGenesis Pharmaceutical Technology Co., Ltd; China CDC; Shandong First Medical University & Shandong Academy of Medical Sciences; Hubei Provincial Center for Disease Control and Prevention
44M	BetaCoV/Wuhan/WH-01/2019	CNA0007332	LR757998, EPI_ISL_406798, NMDC60013002-01	Complete	2019-12-26	General Hospital of Central Theater Command of People's Liberation Army of China	2020-01-30	BGI PathoGenesis Pharmaceutical Technology Co., Ltd; China CDC; Shandong First Medical University & Shandong Academy of Medical Sciences; Hubei Provincial Center for Disease Control and Prevention
49F	BetaCoV/Wuhan/HBCDC-HB-03/2019	EPI_ISL_412899		Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-02-28	Hubei Provincial Center for Disease Control and Prevention
49F	WIV04	GWHABKL00000000	EPI_ISL_402124, MN996528	Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-01-11	CAS Key Laboratory of Special Pathogens and Biosafety and Center for Emerging Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences
49F	BetaCoV/Wuhan/IPBCAMS-WH-02/2019	GWHABKG00000000	EPI_ISL_403931, MT019530	Complete	2019-12-30	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	2020-01-21	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
49F	BetaCoV/Wuhan/HBCDC-HB-01/2019	EPI_ISL_402132		Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-01-19	Hubei Provincial Center for Disease Control and Prevention
49F	BetaCoV/Wuhan/IVDC-HB-01/2019	EPI_ISL_402119		Complete	2019-12-30	National Institute for Viral Disease Control and Prevention, China CDC	2020-01-10	National Institute for Viral Disease Control and Prevention, China CDC
49F	*BetaCoV/Wuhan/WH19001/2019*	*NMDC60013002-08*		*Complete*	*2019-12-30*		*2020-01-28*	*China CDC; Shandong First Medical University & Shandong Academy of Medical Sciences; Hubei Provincial Center for Disease Control and Prevention; BGI PathoGenesis Pharmaceutical Technology Co., Ltd*
52F	WIV05	GWHABKM00000000	EPI_ISL_402128, MN996529	Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-01-18	CAS Key Laboratory of Special Pathogens and Biosafety and Center for Emerging Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences
52F	BetaCoV/Wuhan/IPBCAMS-WH-04/2019	GWHABKI00000000	EPI_ISL_403929, MT019532	Complete	2019-12-30	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	2020-01-21	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
56M	WIV07	GWHABKO00000000	EPI_ISL_402130, MN996531	Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-01-18	CAS Key Laboratory of Special Pathogens and Biosafety and Center for Emerging Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences
61M	BetaCoV/Wuhan/IPBCAMS-WH-05/2020	GWHABKJ00000000	EPI_ISL_403928, MT019533	Complete	2020-01-01	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	2020-01-21	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College; China National Center for Bioinformation
61M	BetaCoV/Wuhan/IVDC-HB-04/2020	EPI_ISL_402120		Complete	2020-01-01	National Institute for Viral Disease Control and Prevention, China CDC	2020-01-11	National Institute for Viral Disease Control and Prevention, China CDC
61M	*BetaCoV/Wuhan/WH19004/2020*	*NMDC60013002-09*		*Complete*	*2020-01-01*		*2020-01-28*	*China CDC; Shandong First Medical University & Shandong Academy of Medical Sciences; Hubei Provincial Center for Disease Control and Prevention; BGI PathoGenesis Pharmaceutical Technology Co., Ltd*
65M	BetaCoV/Wuhan/IPBCAMS-WH-01/2019	GWHABKF00000000	EPI_ISL_402123, MT019529	Complete	2019-12-23	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	2020-01-11	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College; Vision Medicals Co., Ltd
UM1	BetaCoV/Wuhan/HBCDC-HB-02/2019	EPI_ISL_412898		Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-02-28	Hubei Provincial Center for Disease Control and Prevention
UM2	BetaCoV/Wuhan/HBCDC-HB-04/2019	EPI_ISL_412900		Complete	2019-12-30	Wuhan Jinyintan Hospital	2020-02-28	Hubei Provincial Center for Disease Control and Prevention
UU1	BetaCoV/Wuhan/WH19008/2019	NMDC60013002-06		Complete	2019-12-30		2020-01-28	China CDC; Shandong First Medical University & Shandong Academy of Medical Sciences; Hubei Provincial Center for Disease Control and Prevention; BGI PathoGenesis Pharmaceutical Technology Co., Ltd

Figure 2. Combined data table from CNCB and GISAID.

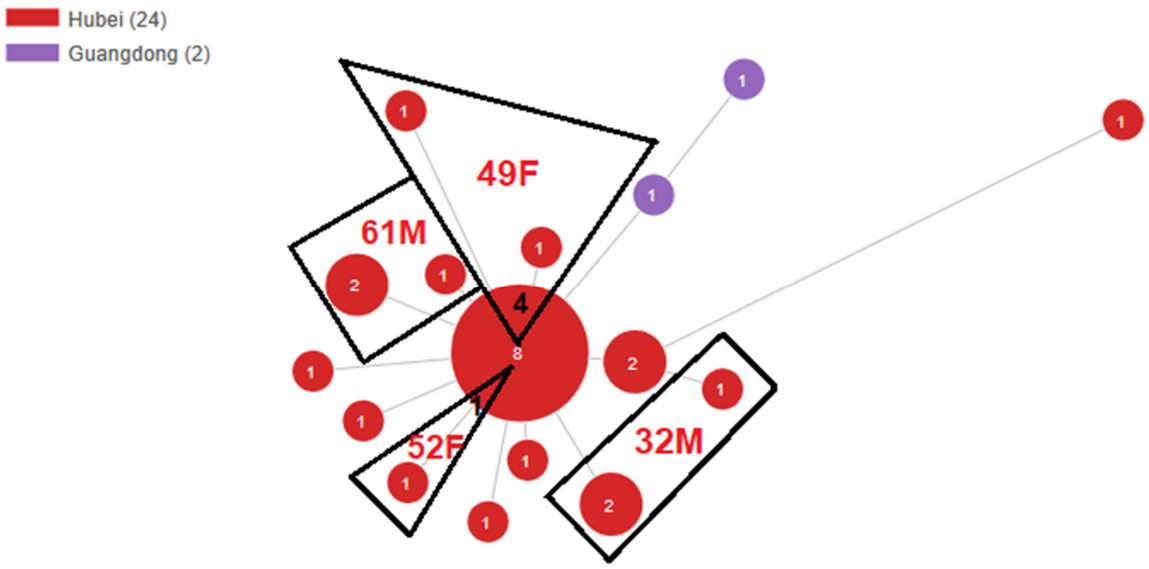

Figure 3. Grouping of early viral genomic sequences by patients.

Errors	Platform	Assembler	AgeGender	Virus Name 病毒株名
0	DNBSEQ	SPAdes v3.12.0	21F	BetaCoV/Wuhan/WH-03/2019
4	Illumina MiSeq, MGISEQ 2000	Geneious v11.0.3, MEGAHIT v1.2.9	32M	WIV02
			32M	BetaCoV/Wuhan/IVDC-HB-05/2019
			32M	*BetaCoV/Wuhan/WH19005/2019*
0	Illumina MiSeq, MGISEQ 2000	Geneious v11.0.3, MEGAHIT v1.2.9	40M	WIV06
	Illumina NextSeq		41M1	BetaCoV/Wuhan/IPBCAMS-WH-03/2019
0	Illumina MiniSeq	Megahit v1.1.3	41M2	Wuhan-Hu-1
	DNBSEQ	SPAdes v3.12.0	43M	BetaCoV/Wuhan/WH-02/2019
	DNBSEQ	SPAdes v3.12.0	44M	BetaCoV/Wuhan/WH-01/2019
0	Illumina Miseq	CLC Genomics Workbench 12 and Geneious 12.0.1	49F	BetaCoV/Wuhan/HBCDC-HB-03/2019
0			49F	WIV04
6	Illumina NextSeq		49F	BetaCoV/Wuhan/IPBCAMS-WH-02/2019
1	Illumina Miseq	CLC Genomics Workbench 12 and Geneious 12.0.1	49F	BetaCoV/Wuhan/HBCDC-HB-01/2019
0			49F	BetaCoV/Wuhan/IVDC-HB-01/2019
			49F	*BetaCoV/Wuhan/WH19001/2019*
2	MGISEQ 2000	Geneious v11.0.3, MEGAHIT v1.2.9	52F	WIV05
0	Illumina NextSeq		52F	BetaCoV/Wuhan/IPBCAMS-WH-04/2019
			56M	WIV07
4	Illumina NextSeq		61M	BetaCoV/Wuhan/IPBCAMS-WH-05/2020
			61M	BetaCoV/Wuhan/IVDC-HB-04/2020
			61M	*BetaCoV/Wuhan/WH19004/2020*
	Illumina NextSeq		65M	BetaCoV/Wuhan/IPBCAMS-WH-01/2019
0	Illumina Miseq	CLC Genomics Workbench 12 and Geneious 12.0.1	UM1	BetaCoV/Wuhan/HBCDC-HB-02/2019
	Illumina Miseq	CLC Genomics Workbench 12 and Geneious 12.0.1	UM2	BetaCoV/Wuhan/HBCDC-HB-04/2019
0			UU1	BetaCoV/Wuhan/WH19008/2019

Figure 4. Counting sequencing and/or assembly errors.