No more business as usual: agile and effective responses to emerging pathogen threats require open data and open analytics

usegalaxy.org[1,2,3], usegalaxy.eu[4], usegalaxy.org.au[5], usegalaxy.be[6] and hyphy.org[7] development teams

[1] The Pennsylvania State University, USA
[2] Johns Hopkins University, USA
[3] Cleveland Clinic, USA
[4] University of Freiburg, Germany
[5] University of Melbourne, Australia
[6] Ghent University, Belgium
[7] Temple University, USA

Authors and contributions:

Dannon Baker[1], Marius Van Den Beek[2], Daniel Blankenberg[3], Dave Bouvier[4], John Chilton[5], Nate Coraor[6], Frederik Coppens[7], Simon Gladman[8], Björn Grüning[9], Delphine Larivière[10], Andrew Lonie[11], Sergei Kosakovsky Pond[12], Wolfgang Maier[13], Anton Nekrutenko[14], James Taylor[15], Steven Weaver[16]

Study design:	12, 14
Analysis:	10, 12, 14, 15
Writing:	12, 13, 14
System design:	6, 8, 9
Tool development:	2, 4, 13, 16
Infrastructure design:	1, 3, 5
Funding:	7, 9, 11,12, 14, 15

Correspondence: Anton Nekrutenko (anton@nekrut.org) and Sergei Kosakovsky Pond (spond@temple.edu)

Global cooperation, necessary for tackling public health emergencies such as the Wuhan pneumonia virus (COVID-19) outbreak, requires unimpeded access to data, analysis tools, and computational infrastructure. The current state of much of COVID-19 research shows regrettable lack of data sharing and considerable analytical obfuscation. In this study, we use all COVID-19 genomic data available in the public domain so far to (1) underscore the importance of access to raw data and to (2) demonstrate that existing community efforts in curation and deployment of biomedical software can reliably support rapid, reproducible research during global crises.

The initial publications describing genomic features of COVID-19 [1–4] used Illumina and Oxford nanopore data to elucidate the sequence composition of patient specimens (although only Wu et al. [3] explicitly provided the accession numbers for their raw short read sequencing data). However, their approaches to processing, assembly, and analysis of raw data differed widely (Table 1) and ranged from transparent [3] to entirely opaque [4]. Such lack of analytical transparency sets a dangerous precedent. Infectious disease outbreaks often occur in locations where infrastructure necessary for data analysis may be inaccessible or unbiased interpretation of results may be politically untenable. As a consequence, there is a global need to ensure access to free, open, and robust analytical approaches that can be used by anyone in the world to analyze, interpret, and share data. Can existing tools and computational resources support such a global need? Here we show that they can: we analyzed all available raw COVID-19 data to demonstrate that analyses described in [1–4] can be reproduced on public infrastructure using open source tools by any researcher with an Internet connection.

We exclusively used free software tools publicly available from the BioConda package distribution system [5], deployed through the worldwide network of open Galaxy platforms [6] and executed using public high throughput computational infrastructure (XSEDE in the US, de.NBI and ELIXIR in EU). We also used an open source Jupyter environment [7] for exploratory analysis of data. All analyses performed here are fully documented and accessible at https://github.com/galaxyproject/SARS-CoV-2/ and https://doi.org/10.5281/zenodo.3678710.

We divided our analysis into the following stages: (1) read pre-processing, (2) genome assembly, (3) timing the most recent common ancestor (MRCA), (4) analysis of genomic variation within individual samples, and (5) recombination and selection analyses.

We pre-processed six currently available (as of Feb 19, 2020) sequencing read datasets for COVID-19 (Table S1) by removing adapter contamination and reads derived from human transcripts and combined the resulting datasets. This was done to enrich COVID-19-specific reads that constitute only a fraction of the original data. These were used as inputs for SPAdes assembler [8] and Unicycler [9]—an assembly pipeline based on SPAdes that includes a number of pre-processing and polishing steps. Both approaches were able to reconstruct a full length COVID-19 genome with Unicycler producing a cleaner assembly graph. Its largest contig (29,781bp) had 100% identity to the published assembly NC_045512.

Next we estimated the date of the most recent common ancestor (MRCA) of COVID-19. For this we used simple root-to-tip regression [10] (more complex and powerful phylodynamics methods could certainly be used, but for this data with very low levels of sequence divergence, simpler and faster methods suffice). Using a set of sequences from all COVID-19 sequences available as of Feb 16, 2020 we obtained an MRCA date of Oct 24, 2019, which is close to other existing estimates [11].

The vast majority of COVID-19 genomic data available at the time of writing are partially or fully assembled genomes. There is no public access to sequence reads that were used to produce these assemblies: as of Feb 19, 2020—more than two months since the beginning of the outbreak—there are only six raw datasets (Table S1). This state of affairs should be unacceptable, since raw read data can be used to uncover viral diversity within individual samples—the type of information gets removed by the assembly process, and to evaluate robustness and reliability of the assembly and variant calling process. To demonstrate that such diversity exists, we mapped Illumina reads against COVID-19 reference (NC_045512) and identified sequence variants with frequencies above 5% while taking into account quality of alternative bases and strand bias. Five percent was selected as a conservative threshold that can be reliably resolved from Illumina data [12]. Using this threshold, thirty nine single nucleotide variants (SNVs) were identified in total across all samples (Fig. 1, Supplemental Table 1). SRR10971381, the sample with most variants (and most reads), contained a cluster of three substitutions within the first two codons of the *M* gene with alternate allele frequencies ranging from 5.2% to 7.8%. There was no evidence of linkage among these variants.

The most prominent sequence variant was observed in sample SRR10903401. It is an A-to-C substitution with alternate allele frequency of 38% that causes a Lys^{921}Gln amino acid replacement within the spike glycoprotein S (product of gene *S*). S is a homotrimeric protein containing S1 and S2 subdomains mediating receptor recognition and membrane fusion, respectively [13]. S2 subdomains contain two heptad repeat (repeats of units containing seven amino acids) regions: HR1 and HR2. The Lys^{921}Gln substitution we observed is located in HR1 and forms a salt bridge with Gln^{1188} within HR2. This is one in a series of salt bridges involved in the formation of the HR1/HR2 hairpin structures [14]. This site invariably contains Lys in all human SARS-related coronaviruses (S protein residue 903) as well as in many other coronaviruses (Fig 2). However, other, more distantly related coronaviruses including transmissible gastroenteritis coronavirus (TGEV), the porcine respiratory coronavirus (PRCV), the canine coronavirus (CCV), the feline peritonitis virus (FIPV), and the porcine epidemic diarrhea virus (PDEV) all contain Gln at the corresponding position ([14] and Fig. 2). The Lys^{921}Gln change would prevent the formation of the salt bridge with Gln^{1188} and may have structural and functional implications for the spike protein. This potentially adaptive change was not observed in the other two samples and lack of raw read data prevented us from identifying it in other geographically and temporally distributed samples.

To detect potential genome rearrangement events that might have led to the emergence of COVID-19 we performed analysis of recombination using a genetic algorithm approach [15]. Wu et al. [3] identified two potential recombination breakpoints within the COVID-19 S-gene with some segments having higher similarity to Bat ZC45 and ZXC21 coronaviruses (accessions MG772933 and MG772934, respectively), while others were more similar to SARS Tor2 and SZ3 isolates (accessions AY274119 and AY304486). Our attempt at reproducing this analysis did identify a set of potential breakpoints similar to the ones reported by Wu et al. [3], but lacking robust statistical support (Fig. 3).

Finally, we performed a branch-level test for positive selection on a codon-alignment of the *S* gene from COVID-19, SARS-Tor2 as well as Bat ZC45, ZXC21, and Rp3 coronaviruses, specifically to identify if there was any evidence of diversifying selection along the ancestral branch leading to COVID-19 isolates. We found statistically significant evidence of positive diversifying selection (~7% of *S*-gene sites) along the branch leading to COVID-19 (Fig. 4).

The goal of our study was to (1) raise awareness of the lack of primary data necessary to effectively respond to global emergencies such as the COVID-19 outbreak and (2) demonstrate that all analyses can be performed transparently with already existing tools and computational infrastructure. The first problem — reluctance to share primary data — has its roots in the fact that the ultimate reward for research efforts is a publication. As a result individual researchers are naturally averse to sharing primary data prior to manuscript acceptance. One may hope that the need to fight global pandemics will downplay individualistic tendencies, yet this is likely wishful thinking. The second problem—the use of disparate tools in an irreproducible fashion — relates to the fact that data analysis remains a “second-class” citizen in infectious disease research and in life sciences in general.

We want to particularly emphasize the issue of irreproducibility. All researchers involved in any given outbreak research should have access to a set of community-curated tools in the same way as they have access to COVID-19 RT-PCR primers [16]. Moreover they should have access to computational infrastructure that can execute these tools and apply them to potentially large NGS datasets. This is essential as precious time is spent on “reinventing the wheel”. Instead, in an ideal world, after reading any of the original COVID-19 manuscripts any researcher should be able to apply the same analytical procedures to their own data. To illustrate these points we assessed the reproducibility of the four initial manuscripts describing the COVID-19 genome (Table 2). All manuscripts reported versions of the software used but none listed parameters used. This effectively prevents quality control and replication because outcomes of complex procedures such as genome assembly, phylogenetic reconstruction, and recombination analysis are notoriously parameter-dependent. One of the manuscripts [4] explicitly lists FreeBayes, a variant discovery tool, as software used for short read assembly—something that FreeBayes is not capable of doing. Finally, only [3] provided access to the raw data rendering the other three manuscripts unverifiable and completely irreproducible.

Our short study demonstrates that viral genome analyses can be performed using open worldwide scientific infrastructure and relying entirely on community-curated open-source software. While we used Galaxy as the platform to execute all analyses described here, the individual software components can be obtained directly from BioConda and run independently. They can be combined into workflows using systems like the CWL[17], Nextflow[18], or Snakemake[19]. Whatever the execution environment or workflow engine, using community supported, versioned, open source tools makes data analyses robust and transparent. This increases the quality and impact of biomedical research.

As we were submitting this manuscript on February 21, 2020, several new sets of raw Illumina reads have been deposited to the Short Read Archive at the NCBI. We did not analyze them in this version of our manuscript. However, this reinforces the key point we are making in this study—anyone can use the open workflows described here to analyze the new data!

Acknowledgments

Usegalaxy.org efforts are funded by NIH Grants U41 HG006620 and NSF ABI Grant 1661497. Usegalaxy.eu is supported by the German Federal Ministry of Education and Research grants 031L0101C and de.NBI-epi. Galaxy and HyPhy integration is supported by NIH grant R01 AI134384. Hyphy.org development team is supported by NIH grant R01GM093939.

Table 1. Methods used for the analysis of primary 2019-nCoV data. ? = uncertain (e.g., Holshue et al. [4] identify FreeBayes as an assembly tool).

Analysis stage	**Publication**			
	[3]	[1]	[2]	[4]
Tools	Bowtie2 (NV)	BWA	BWA	minimap
Assembly	MegaHit 1.1.3, Trinity 2.5.1	Geneous MegaHit	SPAdes CNCBio	Sequencher FreeBayes?
Comparative Analysis	MAFFT 7.407 PhyML MEGA RDP4	MAFFT Clustal RAxML	MAFFT RAxML	
Versions specified	+	+	+	+
Parameters specified	-	-	-	-
Raw data	+	?	-	-

Table S1. Raw COVID-19 sequencing data available at the time of writing (Feb 20, 2020). BALF = bronchoalveolar lavage fluid, * indicates that data may not be reliable (for example, the link between SRR10903402 and [1] is inferred: neither the SRA record nor the manuscript establish this relationship).

#	**Dataset**	**Reads**	**Source**	**Technology**	**Reference**	**Locality**
1	SRR10903401	476.6K	BALF RNA	Miseq, PE	* [1]	Wuhan
2	SRR10903402	676.7K	BALF RNA	Miseq, PE	* [1]	?
3	SRR10971381	28.3M	BALF RNA	Miseq, PE	[3]	?
4	SRR10948550	425.7K	* RNA	Minion	?	?
5	SRR10948474	505.5K	* RNA	Minion	?	?
6	SRR10902284	261.9K	* RNA	Minion	?	?

References

1. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature [Internet]. 2020 Feb 3; Available from: http://dx.doi.org/10.1038/s41586-020-2012-7 PMID: 32015507

2. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet [Internet]. 2020 Jan 30; Available from: http://dx.doi.org/10.1016/S0140-6736(20)30251-8 PMID: 32007145

3. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. A new coronavirus associated with human respiratory disease in China. Nature [Internet]. 2020 Feb 3; Available from: http://dx.doi.org/10.1038/s41586-020-2008-3 PMID: 32015508

4. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A, Diaz G, Cohn A, Fox L, Patel A, Gerber SI, Kim L, Tong S, Lu X, Lindstrom S, Pallansch MA, Weldon WC, Biggs HM, Uyeki TM, Pillai SK, Washington State 2019-nCoV Case Investigation Team. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med [Internet]. 2020 Jan 31; Available from: http://dx.doi.org/10.1056/NEJMoa2001191 PMID: 32004427

5. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J, Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018 Jul;15(7):475–476. PMID: 29967506

6. Goecks J, Nekrutenko A, Taylor J, Team G. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. Department of Biology and Department of Mathematics and Computer Science, Emory University, 1510 Clifton Road NE, Atlanta, GA 30322, USA. jeremy.goecks@emory.edu: BioMed Central Ltd; 2010 Jan 1;11(8):R86.

7. Grüning BA, Rasche E, Rebolledo-Jaramillo B, Eberhard C, Houwaart T, Chilton J, Coraor N, Backofen R, Taylor J, Nekrutenko A. Jupyter and Galaxy: Easing entry barriers into complex data analyses for biomedical researchers. PLoS Comput Biol. 2017 May;13(5):e1005425. PMCID: PMC5444614

8. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012 May 1;19(5):455–477.

9. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017 Jun;13(6):e1005595. PMCID: PMC5481147

10. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. Timing the ancestor of the HIV-1 pandemic strains. Science. 2000 Jun 9;288(5472):1789–1796. PMID: 10846155

11. arambaut, Pinned A, Duchene S, Duplessis L, Volz E, Unpinned A, Globally AP. Phylodynamic Analysis | 93 genomes | 15 Feb 2020 [Internet]. Virological. 2020 [cited 2020 Feb 17]. Available from: http://virological.org/t/phylodynamic-analysis-93-genomes-15-feb-2020/356

12. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. Departments of Pathology, Genome Sciences, and Biochemistry, University of Washington School of Medicine, Seattle, WA 98195, USA.; 2012 Sep 4;109(36):14508–14513.

13. Walls AC, Tortorici MA, Bosch B-J, Frenz B, Rottier PJM, DiMaio F, Rey FA, Veesler D. Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. Nature. 2016 Mar 3;531(7592):114–117. PMCID: PMC5018210

14. Duquerroy S, Vigouroux A, Rottier PJM, Rey FA, Bosch BJ. Central ions and lateral asparagine/glutamine zippers stabilize the post-fusion hairpin conformation of the SARS coronavirus spike glycoprotein. Virology. Elsevier; 2005 May 10;335(2):276–285. PMID: 15840526

15. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. 2006 Oct;23(10):1891–1901. PMID: 16818476

16. CDC. 2019 Novel Coronavirus (2019-nCoV) [Internet]. Centers for Disease Control and Prevention. 2020 [cited 2020 Feb 19]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html

17. Common Workflow Language [Internet]. [cited 2020 Feb 21]. Available from: https://www.commonwl.org/

18. Nextflow - A DSL for parallel and scalable computational pipelines [Internet]. [cited 2020 Feb 21]. Available from: https://www.nextflow.io/

19. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012 Oct 1;28(19):2520–2522. PMID: 22908215

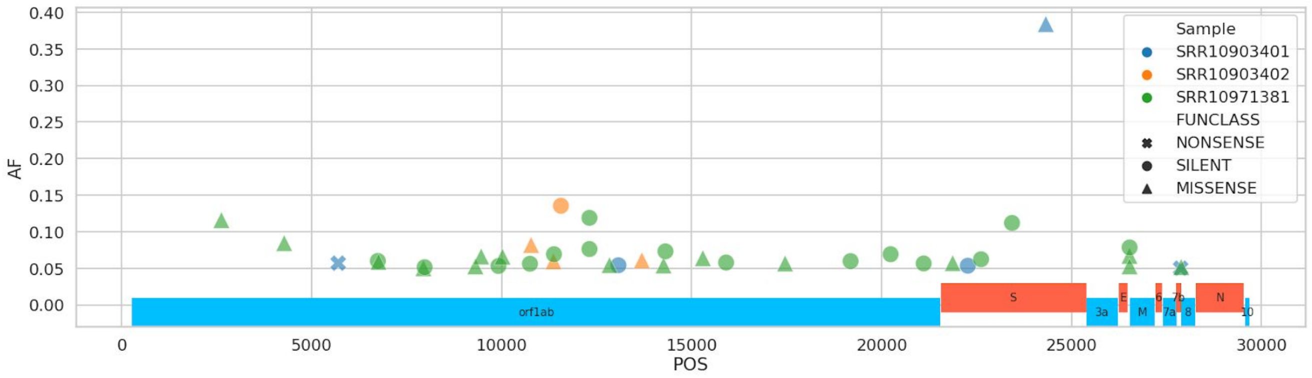

Figure 1. Distribution of nucleotide changes across 2019_nCoV genome. AF = minor allele frequency, POS = position

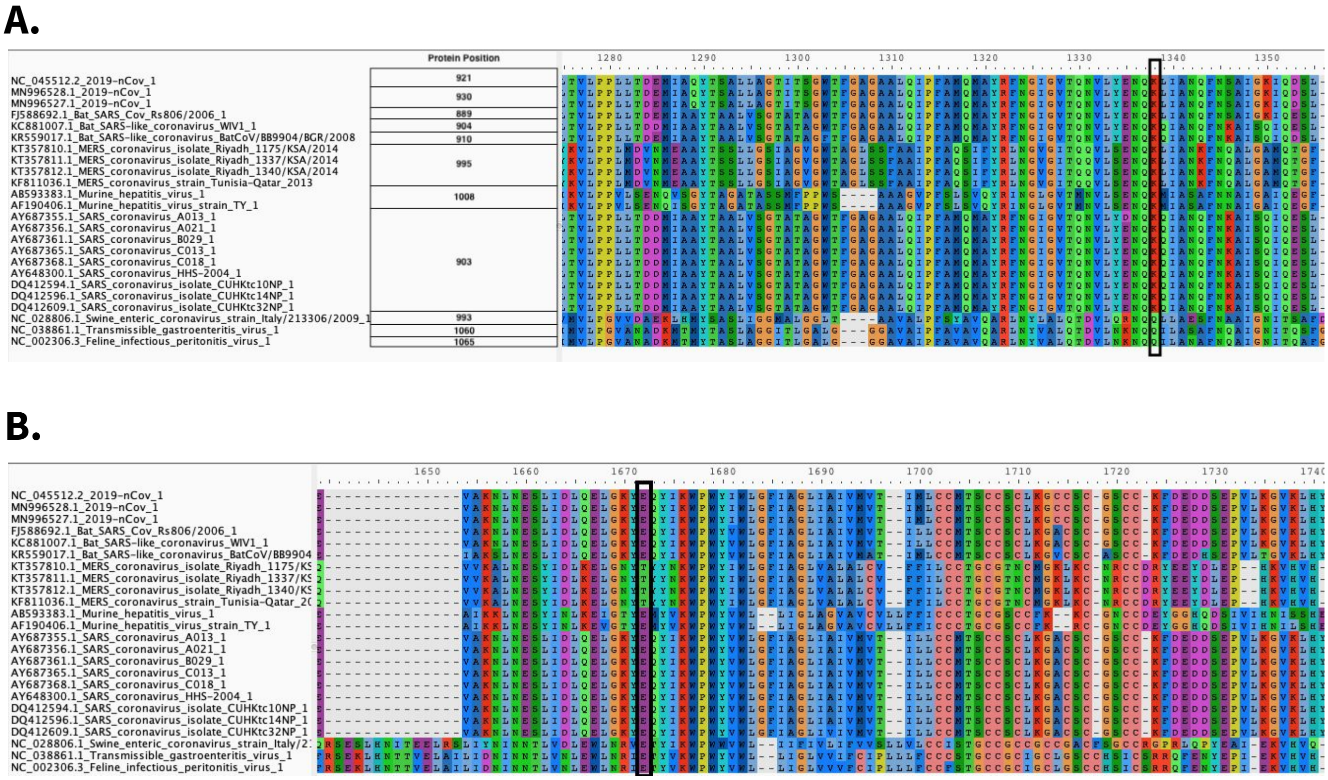
A.
Protein Position
1280
1290
1300
1310
1320
1330
1340
1350
NC_045512.2_2019-nCov_1
MN996528.1_2019-nCov_1
MN996527.1_2019-nCov_1
FJ588692.1_Bat_SARS_Cov_Rs806/2006_1
KC881007.1_Bat_SARS-like_coronavirus_WIV1_1
KR559017.1_Bat_SARS-like_coronavirus_BatCoV/BB9904/BGR/2008
KT357810.1_MERS_coronavirus_isolate_Riyadh_1175/KSA/2014
KT357811.1_MERS_coronavirus_isolate_Riyadh_1337/KSA/2014
KT357812.1_MERS_coronavirus_isolate_Riyadh_1340/KSA/2014
KF811036.1_MERS_coronavirus_strain_Tunisia-Qatar_2013
AB593383.1_Murine_hepatitis_virus_1
AF190406.1_Murine_hepatitis_virus_strain_TY_1
AY687355.1_SARS_coronavirus_A013_1
AY687356.1_SARS_coronavirus_A021_1
AY687361.1_SARS_coronavirus_B029_1
AY687365.1_SARS_coronavirus_C013_1
AY687368.1_SARS_coronavirus_C018_1
AY648300.1_SARS_coronavirus_HHS-2004_1
DQ412594.1_SARS_coronavirus_isolate_CUHKtc10NP_1
DQ412596.1_SARS_coronavirus_isolate_CUHKtc14NP_1
DQ412609.1_SARS_coronavirus_isolate_CUHKtc32NP_1
NC_028806.1_Swine_enteric_coronavirus_strain_Italy/213306/2009_1
NC_038861.1_Transmissible_gastroenteritis_virus_1
NC_002306.3_Feline_infectious_peritonitis_virus_1
921
930
889
904
910
995
1008
903
993
1060
1065
B.
1650
1660
1670
1680
1690
1700
1710
1720
1730
1740
NC_045512.2_2019-nCov_1
MN996528.1_2019-nCov_1
MN996527.1_2019-nCov_1
FJ588692.1_Bat_SARS_Cov_Rs806/2006_1
KC881007.1_Bat_SARS-like_coronavirus_WIV1_1
KR559017.1_Bat_SARS-like_coronavirus_BatCoV/BB9904
KT357810.1_MERS_coronavirus_isolate_Riyadh_1175/KS
KT357811.1_MERS_coronavirus_isolate_Riyadh_1337/KS
KT357812.1_MERS_coronavirus_isolate_Riyadh_1340/KS
KF811036.1_MERS_coronavirus_strain_Tunisia-Qatar_2(
AB593383.1_Murine_hepatitis_virus_1
AF190406.1_Murine_hepatitis_virus_strain_TY_1
AY687355.1_SARS_coronavirus_A013_1
AY687356.1_SARS_coronavirus_A021_1
AY687361.1_SARS_coronavirus_B029_1
AY687365.1_SARS_coronavirus_C013_1
AY687368.1_SARS_coronavirus_C018_1
AY648300.1_SARS_coronavirus_HHS-2004_1
DQ412594.1_SARS_coronavirus_isolate_CUHKtc10NP_1
DQ412596.1_SARS_coronavirus_isolate_CUHKtc14NP_1
DQ412609.1_SARS_coronavirus_isolate_CUHKtc32NP_1
NC_028806.1_Swine_enteric_coronavirus_strain_Italy/2:
NC_038861.1_Transmissible_gastroenteritis_virus_1
NC_002306.3_Feline_infectious_peritonitis_virus_1

Figure 2. Amino acid alignment of spike glycoprotein regions HR1 (A) and HR2 (B). The site of the Lys^{921}Gln substitution observed by us in a COVID-19 isolate is highlighted with a black rectangle in panel A. Its corresponding salt bridge partner is highlighted with black rectangle in panel B.

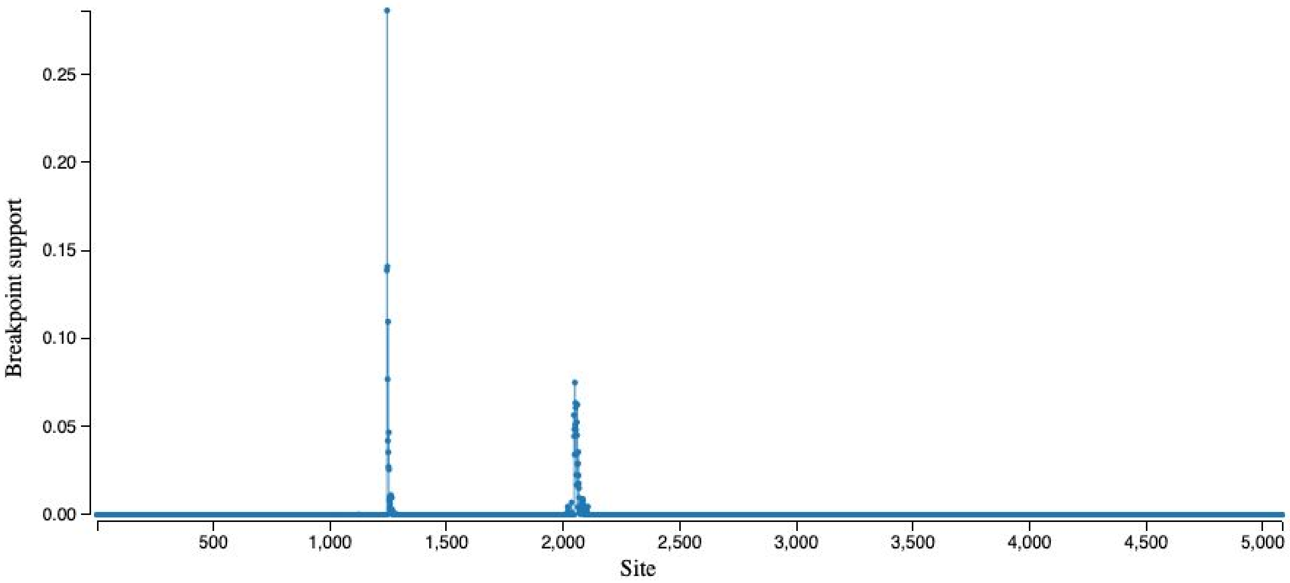

Figure 3. Location of potential recombination breakpoints along the *S*-gene (GARD analysis).

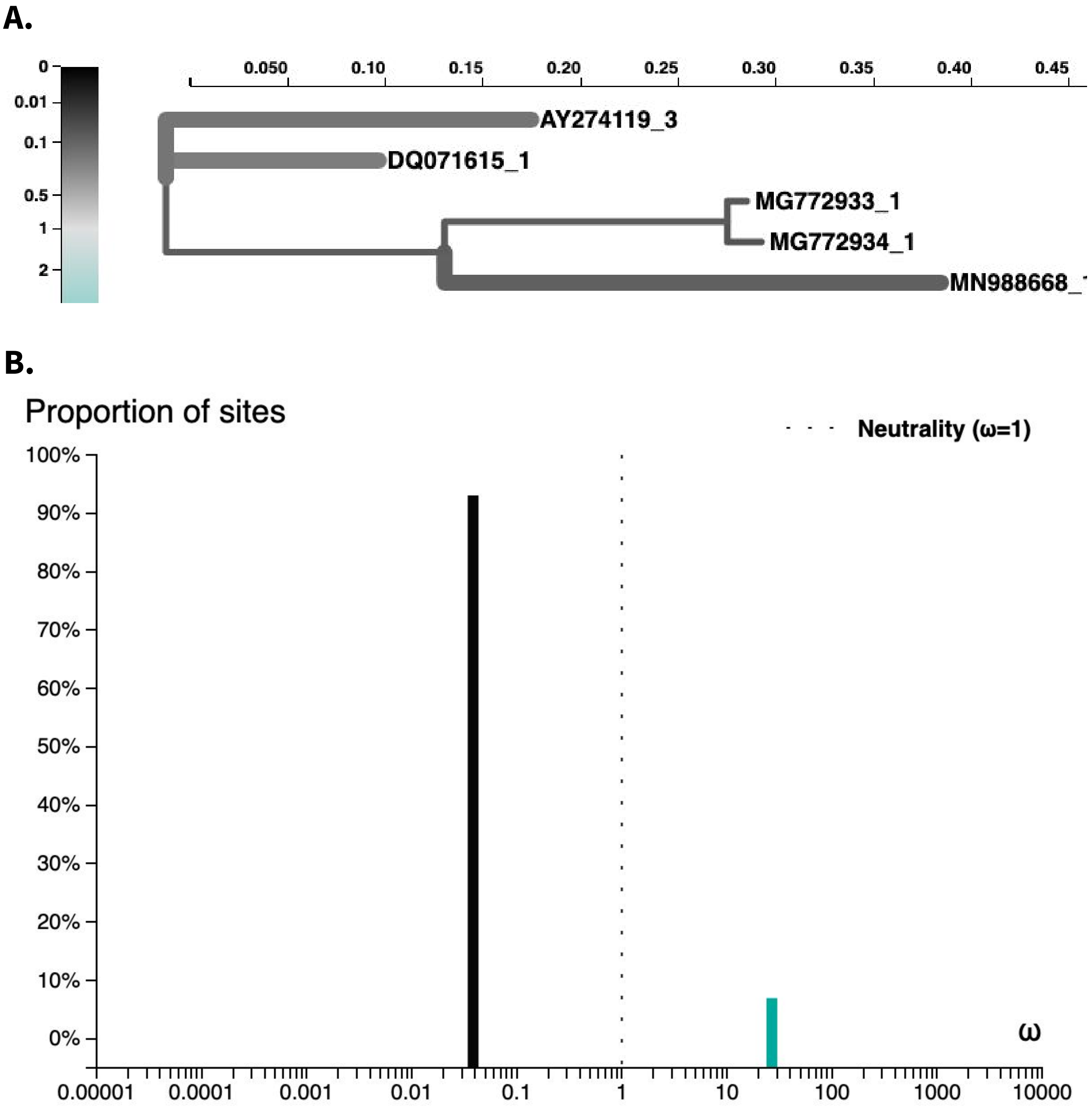

Figure 4. Analysis of branch-specific positive diversifying selection (aBSREL) along the branch leading to COVID-19 (MN988688).