bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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[5,6], usegalaxy.be[7,8] and hyphy.org[9] 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] Queensland Cyber Infrastructure Foundation, Australia
[7] Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium
[8] VIB Center for Plant Systems Biology, Belgium
[9] 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], Ignacio Eguinoa[8], Simon Gladman[9], Björn Grüning[10], Delphine Larivière[11], Andrew Lonie[12], Sergei Kosakovsky Pond[13], Wolfgang Maier[14], Anton Nekrutenko[15], James Taylor[16], Steven Weaver[17]

Study design:	13, 15
Analysis:	11, 13, 15, 16
Writing:	13, 15
System design:	6, 8, 9, 10
Tool development:	2, 4, 14, 17
Infrastructure design:	1, 3, 5
Funding:	7, 10, 12, 13, 15, 16

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The current state of much of the Wuhan pneumonia virus (COVID-19) research shows a regrettable lack of data sharing and considerable analytical obfuscation. This impedes global research cooperation, which is essential for tackling public health emergencies, and requires unimpeded access to data, analysis tools, and computational infrastructure. Here we show that community efforts in developing open analytical software tools over the past ten years, combined with national investments into scientific computational infrastructure, can overcome these deficiencies and provide an accessible platform for tackling global health emergencies in an open and transparent manner. Specifically, 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 health crises. All our analyses are fully documented at https://github.com/galaxyproject/SARS-CoV-2.

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, VSC in Belgium, de.NBI and ELIXIR in EU, NeCTAR Research Cloud in Australia). 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.3685264 (note that these are being continuously updated).

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 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 should be unacceptable, since raw read data can be used to uncover viral diversity within individual samples and to evaluate robustness and reliability of the assembly. 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). 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, 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the salt bridge with Gln^{1188} and may have structural and functional implications for the spike protein structure and, consequently, COVID-19 virulence. 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 open source publicly available 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 highly cited publication. As a result, individual researchers are naturally averse to sharing primary data prior to manuscript acceptance. The second issue—underutilization of existing, community supported tools and analysis frameworks—may be due to the lack of sustained efforts to educate the biomedical community about best practices in (genomic) data analysis. Such efforts exist (e.g., [16]) but have difficulties reaching a wide audience because prominent scientific publication outlets are reluctant to accept data analysis tutorials or reviews. Yet the only way to improve accessibility and reproducibility of biomedical research is through dissemination of best analysis practices.

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 [17]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [18], 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 irreproducible.

Our short study demonstrates that viral genome analyses can be performed using open worldwide scientific infrastructure, relying entirely on community-curated and -supported 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[19], Nextflow[20], or Snakemake[21]. Whatever the execution environment or workflow engine, using community supported, versioned, open source tools makes data analyses robust and transparent. This increases the quality, efficiency and, ultimately, impact of biomedical research.

As we were submitting this manuscript, 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 (Feb 23) 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. In an age of digital connectedness, open, highly accessible, globally shared data and analysis platforms have the potential to transform the way biomedical research is done, opening the way to 'global research markets', where competition arises from deriving understanding rather than access to samples and data. Other disciplines have embraced the benefits of global data generation and sharing, astronomy and high energy physics being two highly successful examples. We have the opportunity to mirror their successes in infrastructure funding by demonstrating that biological research can embrace the same global perspective on common infrastructure investment and data sharing.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. Usegalaxy.org.au is supported by Bioplatforms Australia and the Australian Research Data Commons through funding from the Australian Government National Collaborative Research Infrastructure Strategy. Hyphy.org development team is supported by NIH grant R01GM093939. Usegalaxy.be is supported by the Research Foundation-Flanders (FWO) grant I002919N and the Flemish Supercomputer Center (VSC).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Methods used for the analysis of primary COVID-19 data. ? = uncertain (e.g., Holshue et al. [4] identify FreeBayes as an assembly tool).

Analysis stage	Publication			
	[3]	[1]	[2]	[4]
Tools	Bowtie2 MegaHit Trinity MAFFT PhyML MEGA RDP4	BWA Geneous MegaHit MAFFT Clustal RAxML	BWA SPAdes CNCBio MAFFT RAxML	minimap Sequencher FreeBayes?
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	?	?

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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:

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. Batut B, Hiltemann S, Bagnacani A, Baker D, Bhardwaj V, Blank C, Bretaudeau A, Brillet-Guéguen L, Čech M, Chilton J, Clements D, Doppelt-Azeroual O, Erxleben A, Freeberg MA, Gladman S, Hoogstrate Y, Hotz H-R, Houwaart T, Jagtap P, Larivière D, Le Corguillé G, Manke T, Mareuil F, Ramírez F, Ryan D, Sigloch FC, Soranzo N, Wolff J, Videm P, Wolfien M, Wubuli A, Yusuf D, Galaxy Training Network, Taylor J, Backofen R, Nekrutenko A, Grüning B. Community-Driven Data Analysis Training for Biology. Cell Syst. 2018 Jun 27;6(6):752–758.e1. PMCID: PMC6296361

17. 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

18. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv.org [Internet]. Cornell University Library; 2012 Jul 17;q-bio.GN. Available from: http://arxiv.org/abs/1207.3907v2

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

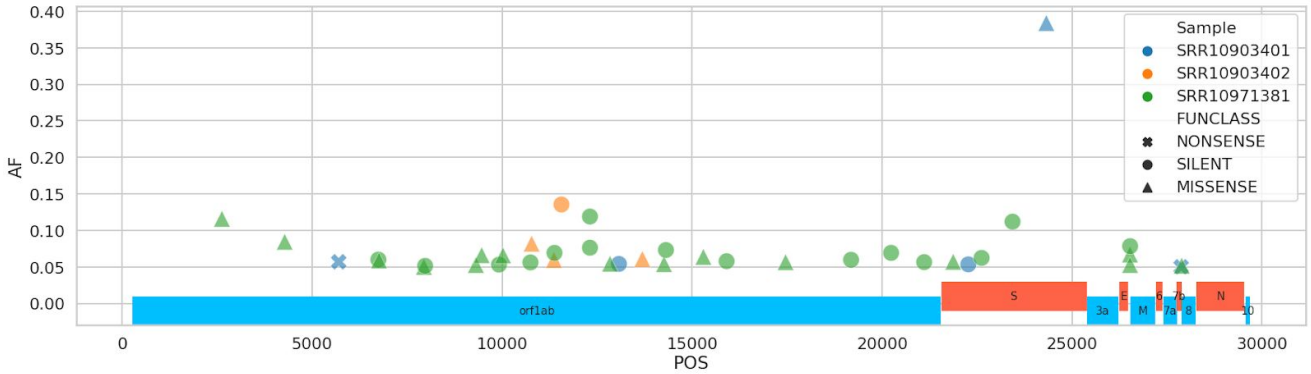
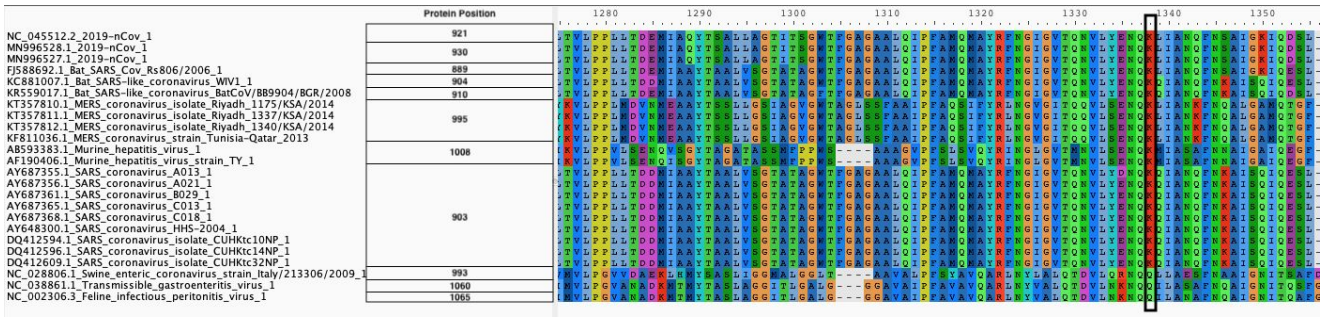


Figure 1. Distribution of nucleotide changes across COVID-19 genome. AF = minor allele frequency, POS = position

A.



B.

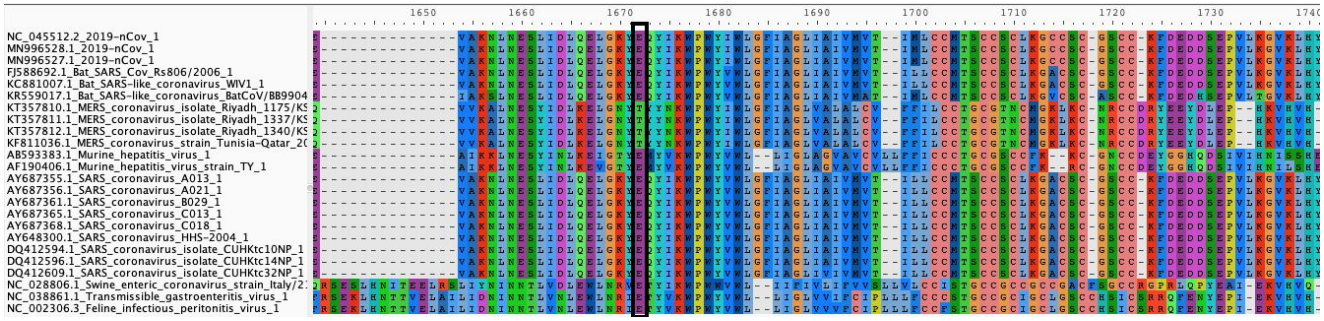


Figure 2. Amino acid alignment of spike glycoprotein regions HR1 (A) and HR2 (B). The site of the Lys[921]Gln substitution observed by us in a COVID-19 isolate is highlighted with a black rectangle in panel A. Its corresponding salt bridge partner is highlighted with black rectangle in panel B.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

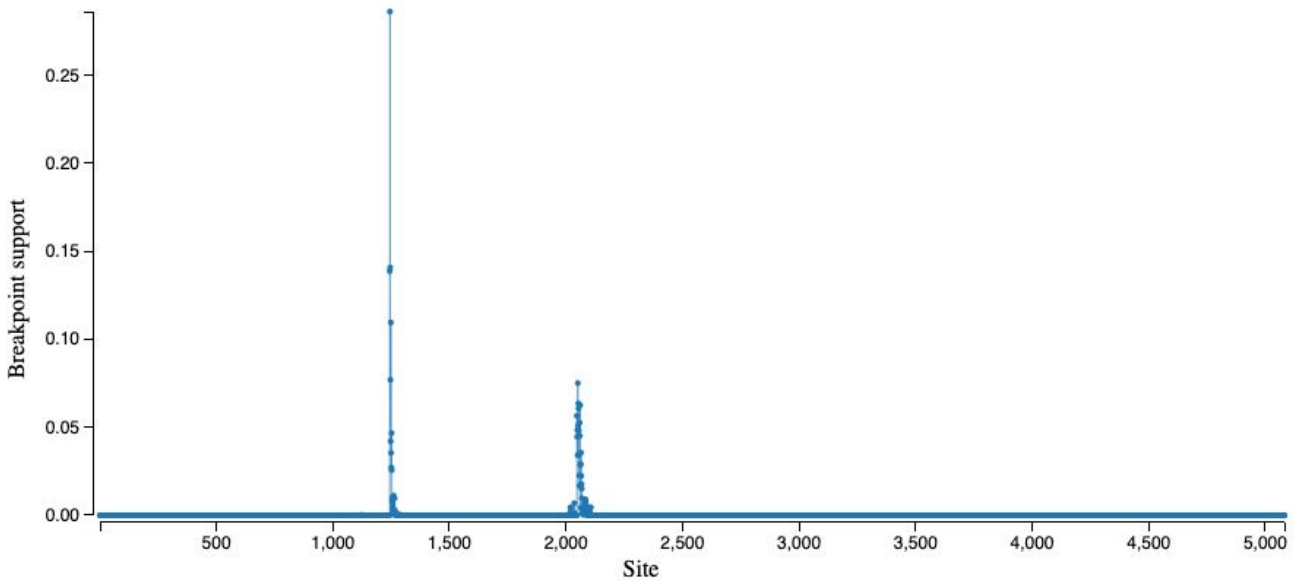


Figure 3. Location of potential recombination breakpoints along the *S*-gene (GARD analysis).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.21.959973; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

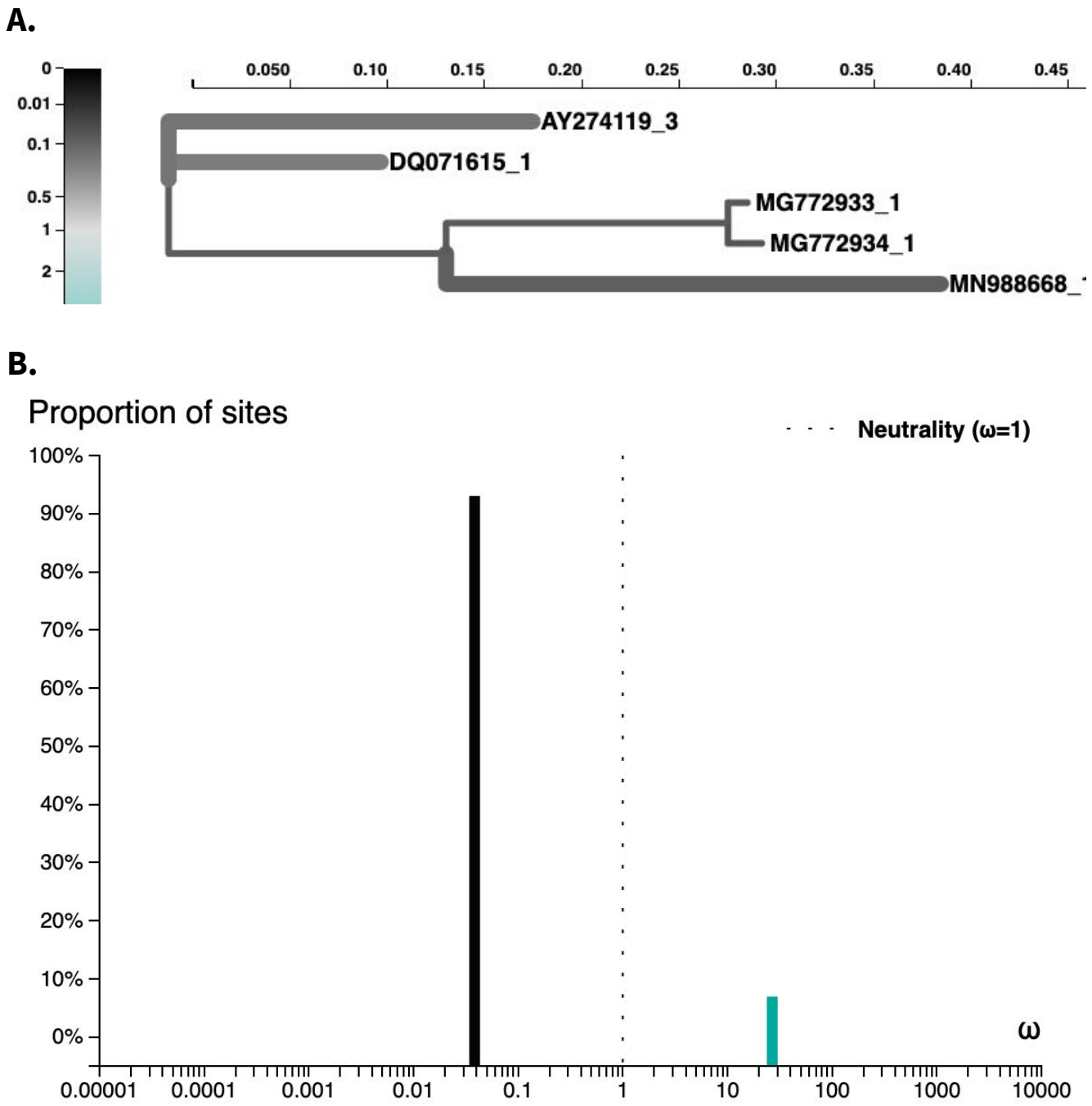


Figure 4. Analysis of branch-specific positive diversifying selection (aBSREL) along the branch leading to COVID-19 (MN988688).