PLOS PATHOGENS

OPINION

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

Dannon Baker[1], Marius van den Beek[2], Daniel Blankenberg[3], Dave Bouvier[2], John Chilton[2], Nate Coraor[2], Frederik Coppens[4,5], Ignacio Eguinoa[4,5], Simon Gladman[6,7], Björn Grüning[8], Nicholas Keener[2], Delphine Larivière[2], Andrew Lonie[6], Sergei Kosakovsky Pond[9]*, Wolfgang Maier[8], Anton Nekrutenko[2]*, James Taylor[1†], Steven Weaver[9]

1 Johns Hopkins University, Baltimore, Maryland, United States of America, **2** The Pennsylvania State University, University Park, Pennsylvania, United States of America, **3** Cleveland Clinic, Cleveland, Ohio, United States of America, **4** VIB Center for Plant Systems Biology, Ghent, Belgium, **5** Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, **6** University of Melbourne, Melbourne, Australia, **7** Queensland Cyber Infrastructure Foundation, St. Lucia, Australia, **8** University of Freiburg, Freiburg im Breisgau, Germany, **9** Temple University, Philadelphia, Pennsylvania, United States of America

† Deceased.
* aun1@psu.edu (AN); spond@temple.edu (SKP)

Citation: Baker D, van den Beek M, Blankenberg D, Bouvier D, Chilton J, Coraor N, et al. (2020) No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics. PLoS Pathog 16(8): e1008643. https://doi.org/10.1371/journal.ppat.1008643

Editor: Carolyn B. Coyne, University of Pittsburgh, UNITED STATES

Published: August 13, 2020

Funding: This work is funded by NIH Grant U41 HG006620 and NSF ABI Grant 1661497 to AN and JT. Usegalaxy.eu is supported by the German Federal Ministry of Education and Research grants 031L0101C and de.NBI-epi to BG. Galaxy and HyPhy integration is supported by NIH Grant R01 AI134384 to AN. 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 to SKP. Usegalaxy.be is supported by the Research Foundation-Flanders (FWO) grant

Abstract

The current state of much of the Wuhan pneumonia virus (severe acute respiratory syndrome coronavirus 2 [SARS-CoV-2]) 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 10 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 SARS-CoV-2 genomic data available in the public domain so far to (1) underscore the importance of access to raw data and (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 SARS-CoV-2 [1–4] used Illumina and Oxford nanopore data to elucidate the sequence composition of patient specimens (although only Wu and colleagues [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

I002919N and the Flemish Supercomputer Center (VSC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Dannon Baker, Dan Blankenberg, Nate Coroar, John Chilton, James Taylor, and Anton Nekrutenko are founders of and hold equity in GalaxyWorks, LLC. The results of the study discussed in this publication could affect the value of GalaxyWorks, LLC.

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 SARS-CoV-2 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], which were deployed through the worldwide network of open Galaxy platforms [6] and executed using public high-throughput computational infrastructure (XSEDE in the United States, VSC in Belgium, de.NBI, and ELIXIR in the European Union, 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 preprocessing, (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 preprocessed currently available (as of February 19, 2020) sequencing read data sets for SARS-CoV-2 (S1 Table) by removing adapter contamination and reads derived from human transcripts and combined the resulting data sets. This was done to SARS-CoV-2–specific reads, which 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 preprocessing and polishing steps. Both approaches were able to reconstruct a full-length SARS-CoV-2 genome, with Unicycler producing a cleaner assembly graph. Its largest contig (29,781 bp) had 100% identity to the published assembly NC_045512.

Next, we estimated the date of the MRCA of SARS-CoV-2. For this, we used simple root-to-tip regression [10] (more complex and powerful phylodynamics methods could certainly be used, but for these data with very low levels of sequence divergence, simpler and faster methods suffice). Using a set of sequences (we removed identical genomes by retaining only one representative from a set of identical sequences) from all SARS-CoV-2 sequences available as of February 16, 2020, we obtained an MRCA date of October 24, 2019, which is close to other existing estimates [11].

Table 1. Methods used for the analysis of primary SARS-CoV-2 data.

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	+	?	–	–

? = uncertain (e.g., Holshue and colleagues [4] identify FreeBayes as an assembly tool).

Abbreviation: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2

https://doi.org/10.1371/journal.ppat.1008643.t001

The vast majority of SARS-CoV-2 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 February 19, 2020—more than 2 months since the beginning of the outbreak—there were only six raw data sets (S1 Table).

This should be unacceptable because raw read data can be used to uncover viral diversity within individual samples and evaluate robustness and reliability of the assembly. To demonstrate that such diversity exists, we mapped Illumina reads against SARS-CoV-2 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, 39 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, which mediate 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 severe acute respiratory syndrome (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 the salt bridge with Gln^{1188} and may have structural and functional implications for the spike protein structure and, consequently, SARS-CoV-2 virulence. This potentially adaptive change was not observed in the other two samples, and a 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 SARS-CoV-2, we performed analysis of recombination using a genetic algorithm approach [15]. Wu and colleagues [3] identified two potential recombination breakpoints within the SARS-CoV-2 *S* gene with some segments having higher similarity to Bat ZC45 and ZXC21

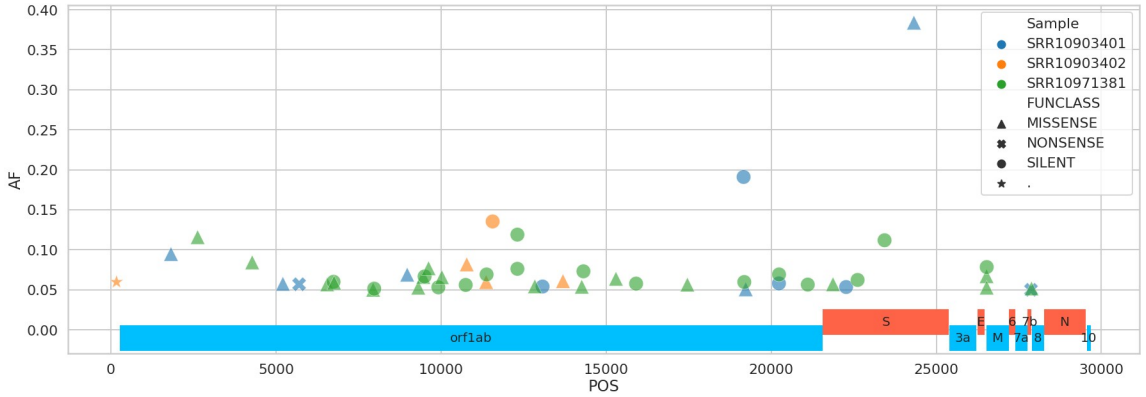

Fig 1. Distribution of nucleotide changes across SARS-CoV-2 genome. AF, minor allele frequency; POS, position; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

https://doi.org/10.1371/journal.ppat.1008643.g001

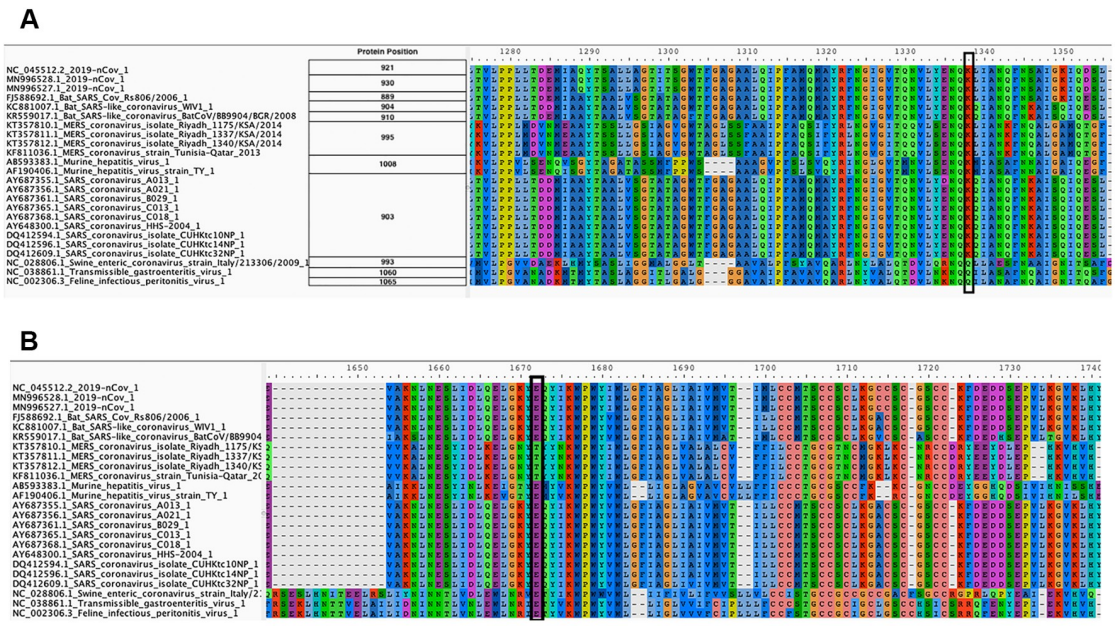

Fig 2. Amino acid alignment of spike glycoprotein regions HR1 (A) and HR2 (B). The site of the Lys^{921}Gln substitution observed by us in a SARS-CoV-2 isolate is highlighted with a black rectangle in panel A. Its corresponding salt bridge partner is highlighted with a black rectangle in panel B. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

https://doi.org/10.1371/journal.ppat.1008643.g002

coronaviruses (accessions MG772933 and MG772934, respectively), whereas 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 and colleagues [3], but they lacked robust statistical support (Fig 3).

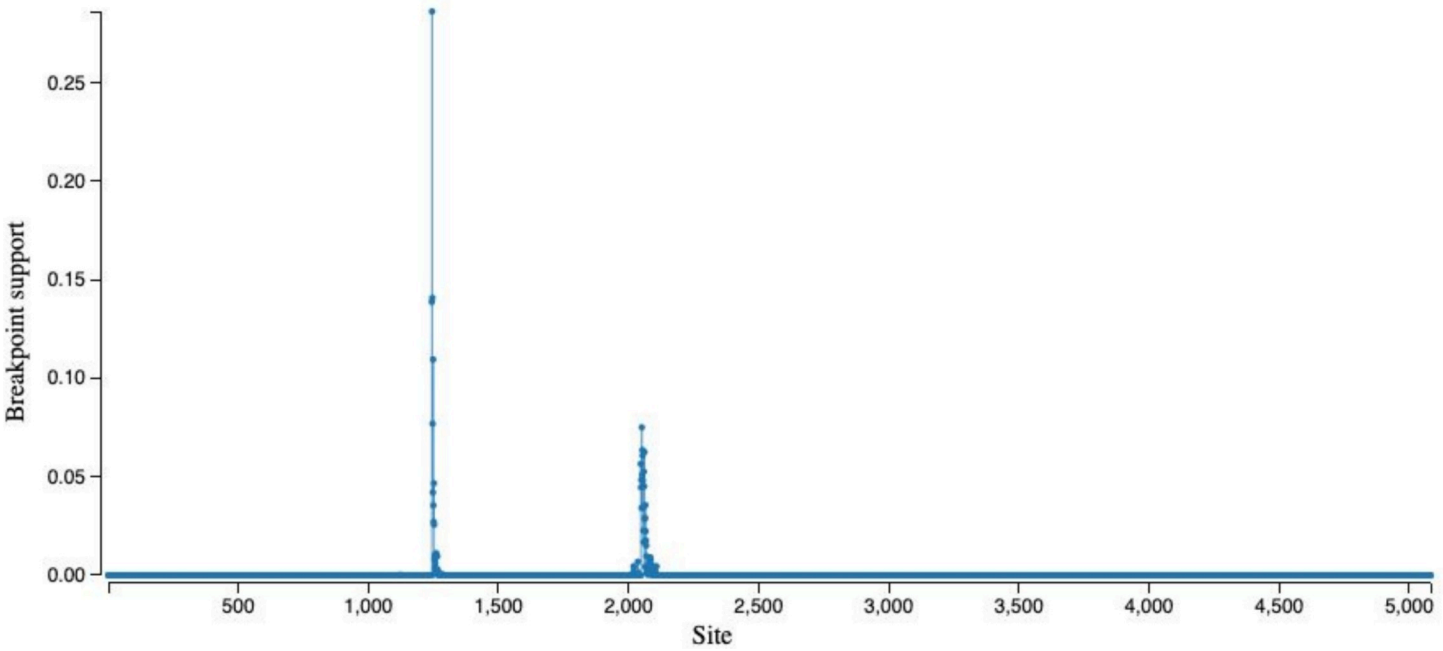

Fig 3. Location of potential recombination breakpoints along the *S* gene (GARD analysis).

https://doi.org/10.1371/journal.ppat.1008643.g003

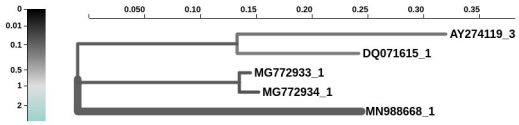

Fig 4. Analysis of branch-specific positive diversifying selection (aBSREL) along the branch leading to SARS-CoV-2 (MN988688). SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

https://doi.org/10.1371/journal.ppat.1008643.g004

Finally, we performed a branch-level test for positive selection on a codon alignment of the *S* gene from SARS-CoV-2, SARS-Tor2, as well as Bat ZC45, ZXC21, and Rp3 coronaviruses, specifically to identify whether there was any evidence of diversifying selection along the ancestral branch leading to SARS-CoV-2 isolates. We found statistically significant evidence of positive diversifying selection (approximately 7% of *S* gene sites) along the branch leading to SARS-CoV-2 (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 coronavirus disease 2019 (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 appropriate 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 appropriate 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 SARS-CoV-2 real-time PCR (RT-PCR) primers [17]. Moreover, they should have access to computational infrastructure that can execute these tools and apply them to potentially large next-generation sequencing (NGS) data sets. This is essential because precious time is spent on "reinventing the wheel." Instead, in an ideal world, after reading any of the original SARS-CoV-2 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 SARS-CoV-2 genome (S1 Table). 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 that relies entirely on community-curated and -supported open-source software. Although 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 Common Workflow Language (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.

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" in which 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.

Changes during review process

The review and publication process of the original version of this manuscript took approximately 5 months. During this time the number of available data sets increased significantly (see https://covid19.galaxyproject.org for the latest results). As of July 31, 2020, the number of available raw read data sets for SARS-CoV-2 was 28,973—a significant increase over the number described in the beginning of this manuscript. However, it is still a fraction of the data: there were ~75,000 assembled genomes in GISAID on the day the proofs were submitted. Until the raw data are released, we will have no ability to validate available assemblies, and any talk about unobstructed analysis of SARS-CoV-2 has no merit. This is not the way serious science is done.

Methods

Our data and analyses are constantly updated. Exact methods and tool versions are available from https://covid19.galaxyproject.org. This repository contains six directories, corresponding to the analyses we have performed: (1) data preprocessing, (2) genome assembly, (3) estimation of MRCA timing, (4) analysis of intrahost variation, (5) analysis of substitutions within the *S* gene, and (6) analysis of recombination and selection. Every page begins with links to workflows and histories at four Galaxy instances in the US (http://usegalaxy.org), EU (http://usegalaxy.eu), Belgium (http://usegalaxy.be), and Australia (http://usegalaxy.org.au). Exact versions of tools used in the analysis are provided at the end of each section.

Supporting information

S1 Table. Raw SARS-CoV-2 sequencing data available at the time of writing (Feb 20, 2020). *Indicates that data may not be reliable (for example, the link between SRR10903402 and [1] is inferred: neither the SRA record nor the manuscript establishes this relationship). On February 21, 2020 new human data sets SRR11092056, SRR11092057, SRR11092058, and SRR11092064 were added to the SARS-CoV-2 archive [22]. Our analyses indicate that these data sets contain no useful SARS-CoV-2 data [23]. BALF, bronchoalveolar lavage fluid; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SRA, Sequence Read Archive.
(RTF)

Acknowledgments

The authors are grateful to the broader Galaxy community for their support and software development efforts. This paper is dedicated to one of its authors, Dr. James Peter Taylor, who

passed away unexpectedly while working on this manuscript on April 2, 2020. He was 40 years old.

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 [cited 2020 Feb 16]; 579:270–273. Available from: https://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 [cited 2020 Feb 16]; 395. Available from: https://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 [cited 2020 Feb 16]; 579:265–269. Available from: https://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 [cited 2020 Feb 16]. Available from: https://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. https://doi.org/10.1038/s41592-018-0046-7 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. 2010 Jan 1; 11(8):R86. https://doi.org/10.1186/gb-2010-11-8-r86 PMID: 20738864
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. https://doi.org/10.1371/journal.pcbi.1005425 PMID: 28542180
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. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599
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. https://doi.org/10.1371/journal.pcbi.1005595 PMID: 28594827
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. https://doi.org/10.1126/science.288.5472.1789 PMID: 10846155.
11. Rambaut A, 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. 2012 Sep 4; 109(36):14508–14513. https://doi.org/10.1073/pnas.1208715109 PMID: 22853953
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. https://doi.org/10.1038/nature16988 PMID: 26855426
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. https://doi.org/10.1016/j.virol.2005.02.022 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. https://doi.org/10.1093/molbev/msl051 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. https://doi.org/10.1016/j.cels.2018.05.012 PMID: 29953864

17. CDC. 2019 Novel Coronavirus (2019-nCoV) [Internet]. Atlanta: 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. https://doi.org/10.1016/j.ajpath.2020.07.001 PMID: 32628931

18. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv [Preprint]. Cornell University Library; 2012 Jul 17 [cited 2020 Feb 16];q-bio.GN. Available from: http://arxiv.org/abs/1207.3907v2.

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

20. Nextflow—A DSL for parallel and scalable computational pipelines [Internet]. Nextflow. [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. https://doi.org/10.1093/bioinformatics/bts480 PMID: 22908215.

22. SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences [Internet]. National Library of Medicine [cited 2020 Feb 25]. Available from: https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/.

23. galaxyproject. galaxyproject/SARS-CoV-2 [Internet]. GitHub. [cited 2020 Feb 25]. Available from: https://github.com/galaxyproject/SARS-CoV-2.