The Ensembl COVID-19 resource: Ongoing integration of public SARS-CoV-2 data

Nishadi H. De Silva, Jyothish Bhai, Marc Chakiachvili, Bruno Contreras-Moreira, Carla Cummins, Adam Frankish, Astrid Gall, Thiago Genez, Kevin L. Howe, Sarah E. Hunt, Fergal J. Martin, Benjamin Moore, Denye Ogeh, Anne Parker, Andrew Parton, Magali Ruffier, Manoj Pandian Sakthivel, Dan Sheppard, John Tate, Anja Thormann, David Thybert, Stephen J. Trevanion, Andrea Winterbottom, Daniel R. Zerbino, Robert D. Finn, Paul Flicek, Andrew D. Yates*

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

*To whom correspondence should be addressed. Tel: +44(0)1223 492538 Email: ayates@ebi.ac.uk

ABSTRACT

The Ensembl COVID-19 browser (covid-19.ensembl.org) was launched in May 2020 in response to the ongoing pandemic. It is Ensembl's contribution to the global efforts to develop treatments, diagnostics and vaccines for COVID-19, and it supports research into the genomic epidemiology and evolution of the SARS-CoV-2 virus. This freely available resource incorporates a new Ensembl gene set, multiple sets of variants, and alignments of annotation from several resources against the reference assembly for SARS-CoV-2. It represents the first virus to be encompassed within the Ensembl platform. Additional data are being continually integrated via our new rapid release protocols alongside tools such as the Ensembl Variant Effect Predictor. Here we describe the data and infrastructure behind the resource and discuss future work.

INTRODUCTION

Over the past couple of decades, multiple coronaviruses have crossed the host species barrier to cause serious outbreaks within human populations. Examples include the SARS epidemic caused by severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003 and the Middle East respiratory syndrome coronavirus (MERS-CoV) outbreak in 2012. Both belong to the *betacoronavirus* genus and are believed to have originated in bats with an intermediary animal host before transmission to humans [Wu A 2020].

Similarly, the SARS-CoV-2 virus responsible for the current COVID-19 pandemic is also a *betacoronavirus* with a 29,903-nucleotide positive-strand RNA genome encoding ~30 known and hypothetical mature proteins. The first open reading frame (ORF), representing approximately 67% of the entire genome, encodes 16 non-structural proteins (nsps). The remaining ORFs encode accessory proteins and four major structural proteins: spike surface glycoprotein (S), small envelope protein (E), matrix protein (M) and nucleocapsid protein (N) [Wu A 2020].

Genomic sequencing has played a crucial role in understanding the mechanisms, spread and evolution of this virus. In support, established genomic resources, such as Ensembl, have adapted quickly to enable our new and existing user communities to engage with the rapidly emerging SARS-CoV-2 data landscape within a familiar browser and computational platform.

Ensembl [Howe 2020, Howe 2019] is a system for generating, integrating and disseminating genomic data for vertebrate and non-vertebrate species. The Ensembl COVID-19 browser (https://covid-19.ensembl.org) supports our first viral genome and was launched in May 2020 in response to the current pandemic. It contains a new Ensembl gene set, multiple sets of variants and alignments to many external data resources to provide one consolidated view of publicly available SARS-CoV-2 genomic data. All of these aspects are detailed in the sections to follow.

Ensembl was originally built to capture data from the Human Genome Project. In the twenty years since its launch, we have steadily integrated a large variety of genomes. Complex genomes such as wheat and the increasing volumes of microbial genomes have consistently pushed us to develop and streamline our processes, schemas and tools. In recent years, our involvement with large-scale sequencing projects including the Darwin Tree of Life[1] has identified developments required to cater to the needs of the genomic community. At the process level, rapid and incremental releases of emerging data are now critical to disseminate information to the relevant expert communities and to reduce duplication of effort.

Ensembl's rapid release concepts address this need [Howe 2020]. This new model was launched in 2020 as Ensembl Rapid Release (https://rapid.ensembl.org) and enables us to quickly annotate and release genomes for the benefit of domain experts and incrementally add facets of analysis as they become available in two week cycles. Our COVID-19 browser was implemented with these ideas and workflows at its core, and prepares Ensembl to react quickly in the face of future outbreaks and challenges.

NEW ENSEMBL COVID RESOURCE

Reference assembly and a new gene annotation

The sequence represented in Ensembl (INSDC accession GCA_009858895.3, MN908947.3) is the viral RNA genome isolated from one of the first cases in Wuhan, China [Wu F 2020]. It is widely used as the standard reference and has been incorporated into other resources such as the UCSC SARS-CoV-2 genome browser [Fernandes 2020]. We imported this assembly from the European Nucleotide Archive (ENA) into an Ensembl core database schema to facilitate display and downstream analysis.

[1] https://www.darwintreeoflife.org/

We customised our gene annotation methods [Aken 2016] to handle the biology of the SARS-CoV-2 genome. To identify protein coding genes, we aligned SARS-CoV-2 proteins downloaded from RefSeq [O'Leary 2016] to the genome using Exonerate [Slater & Birney 2005]. A challenge for annotation is that the first and largest ORF can result in either non-structural proteins nsp1-11 (ORF1a) or in nsp1-nsp10 and nsp12-nsp16 (ORF1ab) when an internal programmed translational frameshift occurs [Chen 2020]. Exonerate handles this ribosomal slippage by inserting a gap in the alignment and thus allowing the annotation of the full ORF1ab locus. The modified annotation methodology then removes the artificial gap and instead models the slippage frameshift as an RNA edit. This creates a more biologically accurate representation of the locus and product.

Our emerging viral annotation pipeline has been used to annotate over 90 additional SARS-CoV-2 assemblies from ENA. These additional annotation sets, while not integrated into our browser, were used to verify the accuracy of the pipeline. To assess accuracy and completeness, we calculated the alignment coverage and percentage identity of the resultant gene translations. We observed full length alignments, with a lowest average percent identity of 99.81 (most being 99.9 or 100).

In addition to generating a fully integrated Ensembl gene annotation, we also imported the gene set submitted to the ENA with the reference sequence by the Shanghai Public Health Clinical Center. As can be seen in figure 1, both the submitted gene annotation (blue) and the Ensembl gene set (red) can be viewed on the browser. The submitted gene annotation is available as a separate track on our browser that can be accessed under the ‘Genes and transcripts’ heading after clicking on ‘Configure this page’ in the left hand menu. The major difference to our annotation is that the submitted annotation does not include the short form ORF1a separately and does not annotate ORF7b.

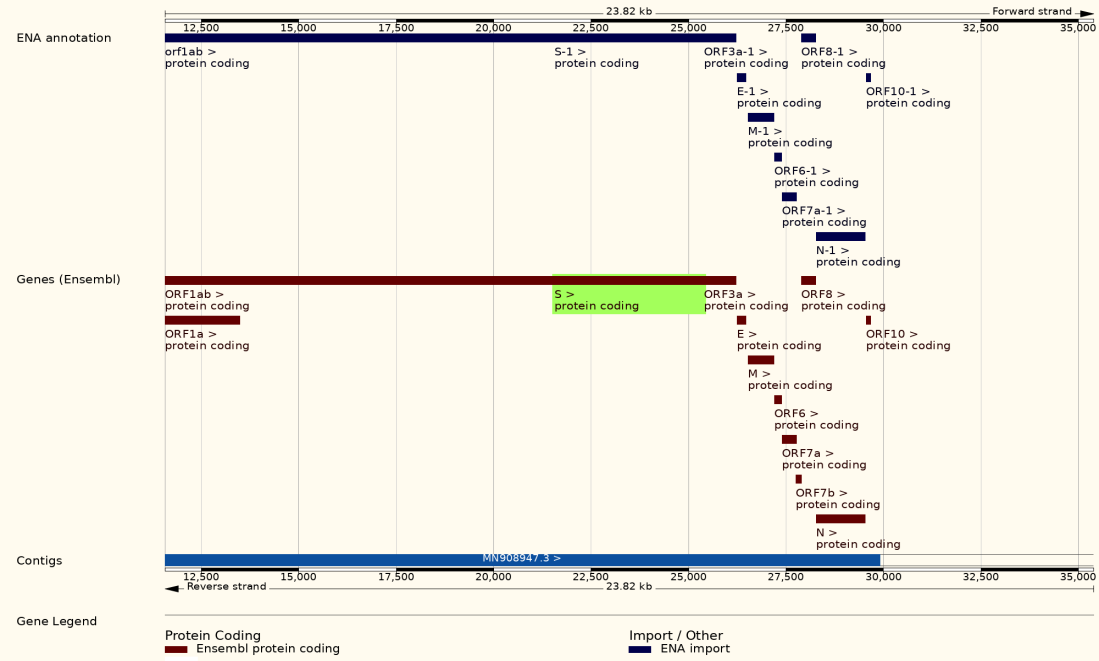

Figure 1: A comparison of the Ensembl gene set and the gene set submitted to the ENA by the Shanghai Public Health Clinical Center for the SARS-CoV-2 reference assembly

Comparison of SARS-CoV-2 with 60 other *orthocoronavirinae* genomes

Comparative genomics places data in an evolutionary context and is thus a powerful toolkit to functionally explore a genome. For instance, a recent comparison of the gene sets available for 44 complete *Sarbecovirus* genomes shows that, while many of the genes align closely, there is a potentially novel alternate frame gene ORF3c and that ORF10, ORF9c, 3b, and 3d are unlikely to be protein coding [Jungreis *et al* 2020].

In a similar approach, we used Cactus [Armstrong 2020] to align SARS-CoV-2 and 60 publicly available virus genomes from the *Orthocoronavirinae* subfamily. This multiple genome alignment showed that 78% of the SARS-CoV-2 genome aligned with at least one other genome and 35% of the genome aligned with the full set of *Orthocoronavirinae* genomes.

The alignment coverage, which represents the number of genomes aligned to a given position within a particular genome, distributes heterogeneously across SARS-CoV-2 genomes. The central region of the SARS-CoV-2 genome (starting from ~7.1Kb and ending at 21.3kb), representing a significant segment of the 3' part of ORF1a, is highly shared across the *Orthocoronavirinae* subfamily. This indicates that the non-structural proteins nsp3 - nsp16 encoded by this region, and involved in the replicative process of the viral nucleic acid, originate directly from the *Orthocoronavirinae* ancestral genome. Inversely, both ends of the SARS-CoV-2 genome have very low alignment coverage and are only shared with closely related viruses.

We also looked at the alignments of the SARS-CoV-2 genomic region coding for the spike protein. The region of the ORF coding for the S2 subunit of the spike protein, responsible for the fusion of the viral and host cell membranes and entry into the host cell [Huang 2020], displays a high alignment coverage while the ORF coding for the S1 subunit of the spike protein has large regions that are shared only by one other related genome; confirming that the S2 subunit is more conserved than S1. We are in the process of analysing the rest of the alignments further.

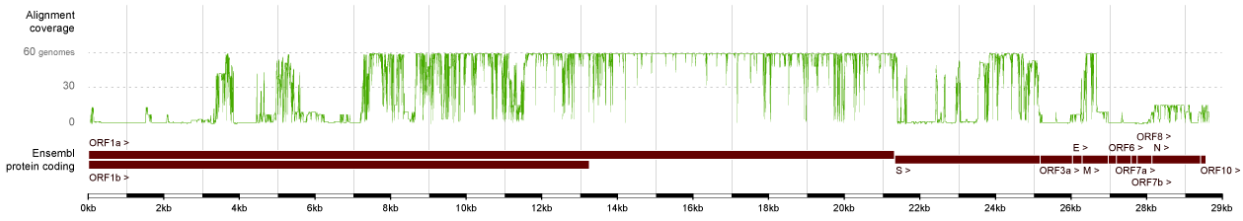

Figure 2: Distribution of the alignment coverage across the SARS-CoV-2 genome

Additionally, we applied our gene tree method [Herrero *et al* 2016] to group the protein coding genes into families and to predict orthologous and paralogous relationships between genes. These results are currently being examined and will be incorporated into the COVID-19 resource.

Genetic variation data

Analysis of the genetic variants of viral genomes is important for understanding the spread of infection across different geographic regions. We display 6,123 sequence variants for SARS-CoV-2 and show their regional frequency distributions alongside predicted molecular consequences as calculated by the Ensembl Variant Effect Predictor (VEP) [McLaren 2016]. The variants on our site are derived from overlapping sample sets by two groups who used different analysis methods.

One set comes from the Nextstrain project which creates phylogenetic trees for tracking pathogen evolution based on virus subsamples [Hadfield 2018]. We converted their 2020-04-08 SARS-CoV-2 data release to VCF for integration into our system and display frequency distributions by country and Nextstrain-inferred clade.

In parallel, the ENA team developed a LoFreq-based [Wilm 2012] pipeline to call variants from SARS-CoV-2 records submitted to their archives. LoFreq reports the proportion of each variant seen in a sample from an individual. However, for simplicity we represent only the alleles seen in each sample and not the proportions estimated. We currently display an early version of the ENA based variant data set (2020-08-17 version) and have applied strict filters to reduce the proportion of lower confidence sites. Specifically, we have removed sequencing runs with more than 40 variant calls; variants where no sample has a frequency of 20% or more for the non-reference allele; and variants where all samples show strand bias. Variants were called for each sample individually and, to provide a more accurate estimation of the frequency of each allele across the entire sample set, it is assumed that sites at which a variant was not called in a sample match the reference.

Some sites are annotated as a further guide to quality. For example, variants seen in more than one sample in either set have an evidence status of 'Multiple observations' and variants at sites recommended for masking by De Maio *et al* (version 2020-07-29) have a flag of ‘Suspect reference location’. These sites for masking are described further below. Variants can be displayed as three separate tracks in the genome browser: those from ENA, those from Nextstrain and those observed in more than one sample in either project as shown in figure 3.

Integration of data from other resources

As with all genomes that go through Ensembl’s production processes, alignments to data in several external repositories were integrated into the SARS-CoV-2 genome.

We added alignments to Rfam’s covariance model using their COVID-19 release 14.2 (http://rfam.xfam.org/covid-19), cross references to RefSeq peptides, UniProt and INSDC proteins, functional annotation from the Gene Ontology Consortium, and annotation of protein domains from InterProScan as well as annotations from other protein annotation resources (including Superfamily, SMART and Gene3D). These are accessible via our region views and the gene and transcript tabs. We also created a genome browser track projecting the protein-domain annotations onto the genome,

facilitating a genome-oriented view of the gene products (including the non-structural cleavage products of ORF1a/ORF1ab).

We display community annotation of sites and regions using results co-ordinated by the UCSC genome browser. Additions to this annotation resource are open to all and done via a publicly available spreadsheet hosted by UCSC[2] which we periodically integrate into our browser. This is done via code[3] that uses Git workflows to convert the annotations into BigBed files that can be visualised on a variety of genome browsers.

We also integrate Oxford Nanopore sequencing primers (version 3) made available by the ARTIC network[4] to assist in sequencing the virus. Though mainly focused on the Oxford Nanopore MinION sequencer, some aspects of the protocol may be generalised to other sequencing platforms. The complete list of primers integrated onto our site are available on GitHub[5]

Furthermore, we provide tracks to visualise problematic and caution sites, which result from common systematic errors associated with lab-specific practices and have been observed in submitted sequences [De Maio *et al* 2020]. Inclusion of these can adversely influence phylogenetic and evolutionary inference. De Maio *et al* (2020) document recommendations for these sites. Visualising them on our browser alongside the locations of primers and other community derived notes helps determine how best to proceed with each of these sites.

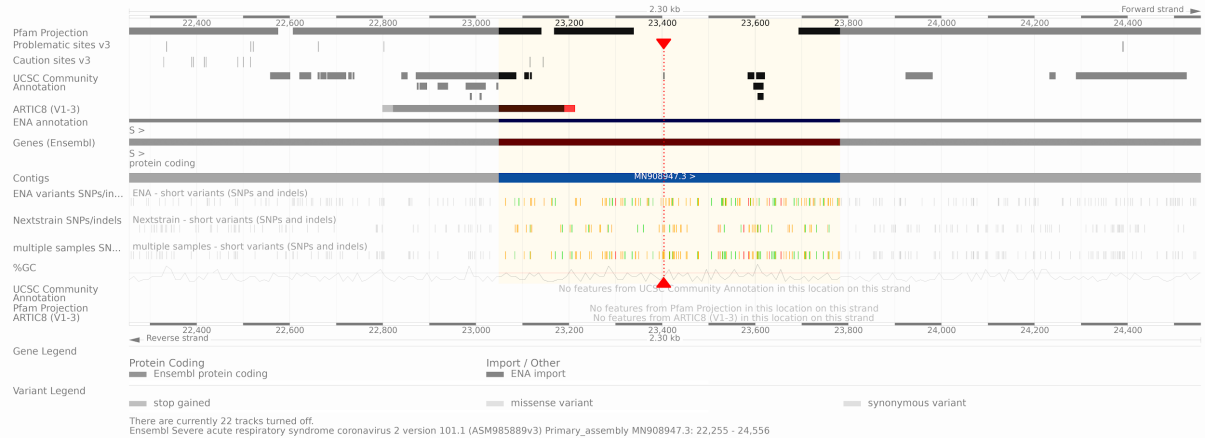

Figure 3: A screenshot of the browser with several tracks turned on and highlighting a substitution flagged up early on in the UCSC community annotation at position 23403 (D614G) in the S spike glycoprotein gene. Due to the prompt nature of community driven annotation, this data was available on our browser as soon as the annotation appeared in a preprint. It is labelled as a common missense mutation in SARS-CoV-2 with a notably high difference in resulting isoelectric point (D->G). Pachetti *et al* (2020) looked at 220 genomic sequences obtained from the GISAID database and characterised 8 novel recurrent mutations; the one at 23403 is one of them. Many studies now show that this particular missense mutation in the

[2] http://bit.ly/cov2annots

[3] https://github.com/Ensembl/sarscov2-annotation

[4] https://artic.network/ncov-2019

[5] https://github.com/artic-network/artic-ncov2019/blob/master/primer_schemes/nCoV-2019/V3/nCoV-2019.tsv

spike protein is predominantly observed in Europe [Isabel *et al* 2020]; patterns that can also be seen in the variation data we host.

A new rapid release browser

The Ensembl website code is highly configurable and we were able to exploit this to rapidly create a test site for internal review. We also took this opportunity to design a new landing page, which prioritises key views and data to help direct researchers into relevant sections of our site.

Although some of the processes used to build this site differ from those used for the Ensembl Rapid Release site, there is no doubt that our experiences in developing the Ensembl Rapid Release site proved very useful in bringing the COVID-19 site to public release in a number of weeks. We have chosen not to make some virus-specific modifications to our existing web codebase - such as showing a single nucleic acid strand and removing all mentions of exons - because we felt the data could be effectively visualised without these changes that would have required a time-consuming code overhaul. However, we have altered the vocabulary wherever possible and are reviewing feedback as we receive it.

The Ensembl COVID-19 resource is also integrated into the European COVID-19 Data Portal hosted by EMBL-EBI (https://www.covid19dataportal.org/). The portal enables searches across the multiple strands of effort on COVID-19 including virus and human sequences; relevant biochemical pathways, interactions, complexes, targets and compounds; protein and expression data; and literature. We have also reached out to our existing and new user communities using our blog and social media accounts, and the Ensembl helpdesk team have been triaging user queries.

DISCUSSION

The swift spread of COVID-19 since emerging in late 2019 has highlighted the necessity for data resources to be prepared for rapid adaptation to developing outbreaks. Our experience integrating thousands of genomes into the Ensembl infrastructure and supporting hundreds of thousands of active users enabled us to effectively develop and release the Ensembl COVID-19 resource. Indeed, our work on a new rapid release browser framework was easily repurposed when the pandemic demanded a quicker response. All of our pipelines and schemas - including gene annotation and integration of references to external data repositories - worked seamlessly even though Ensembl was not designed to support RNA genomes and had not previously been used for viruses. The launch of covid-19.ensembl.org in May 2020 harnessed and tested many of our plans for disseminating data in incremental steps to the community. Although we learned many lessons from this experience and we are not signalling the imminent arrival of a new viral Ensembl platform, we have shown that Ensembl pipelines and infrastructure can effectively cope with genomes beyond our typical scope. Wherever genomics can be used for scientific advances, including understanding the threat of emerging viruses, Ensembl is well positioned as a platform for research and discovery.

AVAILABILITY

The COVID-19 resource from Ensembl is available without any restrictions at https://covid-19.ensembl.org

The selection of code to convert CSV files into BigBed files is at https://github.com/Ensembl/sarscov2-annotation

ACCESSION NUMBERS

The reference genome assembly for SARS-CoV-2 with the accession GCA_009858895.3 was obtained from the European Nucleotide Archive.

ACKNOWLEDGEMENTS

We would like to thank the following people at the EMBL-EBI for their contributions to our resource and thoughtful discussions: Nick Goldman, Zamin Iqbal, Guy Cochrane, Rodrigo Lopez Sanchez, Conor Walker, Nadim Rahman, Jeena Rajan, Alexy Sokolov, Peter Harrison, Youngmi Park, Nicola Buso, Suran Jayathilaka, Anton Petrov, James Allen, Luca Da Rin Fioretto, Thomas Maurel and Vinay Kaikala.

FUNDING

This work was supported by the Wellcome Trust [WT108749/Z/15/Z] and the European Molecular Biology Laboratory (EMBL).

CONFLICT OF INTEREST

No conflicts of interest.

REFERENCES

Aken BL, Ayling S, Barrell D, *et al.* (2016) The Ensembl gene annotation system. Database : the Journal of Biological Databases and Curation. 2016;2016. doi:10.1093/database/baw093.

Armstrong J, Hickey G, Diekhans M *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 587**,** 246–251 (2020). doi:10.1038/s41586-020-2871-y

Chen Y, Liu Q, Guo D. (2020) Emerging coronaviruses: Genome structure, replication, and pathogenesis . *J Med Virol*. 2020;92(4):418-423. doi:10.1002/jmv.25681

De Maio N *et al* (2020) Issues with SARS-CoV-2 sequencing data. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473

Fernandes JD, Hinrichs AS, Clawson H. *et al.* (2020) The UCSC SARS-CoV-2 Genome Browser. *Nat Genet* **52,** 991–998. Doi:10.1038/s41588-020-0700-8

Hadfield J *et al.* (2018) Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics*, vol. 34, no. 23, 2018, pp. 4121–4123., doi:10.1093/bioinformatics/bty407

Herrero J, Muffato M, Beal K, et al. (2016) Ensembl comparative genomics resources. *Database (Oxford)*. Published 2016 Feb 20. doi:10.1093/database/bav096

Howe KL, *et al.* (2019) Ensembl Genomes 2020—Enabling Non-Vertebrate Genomic Research. *Nucleic Acids Research*, vol. 48, no. D1, 2019, doi:10.1093/nar/gkz890.

Howe KL, *et al.* (2020) Ensembl 2021. *Nucleic Acids Research,* 2020, 1, doi:10.1093/nar/gkaa942

Huang Y, Yang C, Xu Xf. *et al.* (2020) Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* 41, 1141–1149 (2020). doi:10.1038/s41401-020-0485-4

Isabel S, Graña-Miraglia L, Gutierrez JM *et al.* (2020) Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *Sci Rep* 10, 14031 (2020). doi:10.1038/s41598-020-70827-z

Jungreis I, *et al*. (2020) SARS-CoV-2 Gene Content and COVID-19 Mutation Impact by Comparing 44 Sarbecovirus Genomes. 2020, doi:10.1101/2020.06.02.130955 (preprint)

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. (2016) The Ensembl Variant Effect Predictor. *Genome Biology* Jun 6;17(1):122. 2016, doi:10.1186/s13059-016-0974-4

O'Leary NA, Wright MW, Brister JR, et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745. doi:10.1093/nar/gkv1189

Pachetti M, Marini B, Benedetti F *et al.* (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med* **18,** 179 (2020). doi:10.1186/s12967-020-02344-6

Wilm A, Aw PP, Bertrand D, *et al.* (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res*. 2012;40(22):11189-11201. doi:10.1093/nar/gks918

Wu A, *et al.* (2020) Genome Composition and Divergence of the Novel Coronavirus (2019-NCoV) Originating in China. *Cell Host & Microbe*, vol. 27, no. 3, 2020, pp. 325–328., doi:10.1016/j.chom.2020.02.001

Wu F, Zhao S, Yu B. *et al.* (2020) A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269 (2020). https://doi.org/10.1038/s41586-020-2008-3

Slater GS, Birney E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31. Published 2005 Feb 15. doi:10.1186/1471-2105-6-31