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 COVID-19 pandemic has seen unprecedented use of SARS-CoV-2 genome sequencing for epidemiological tracking and identification of emerging variants. Understanding the potential impact of these variants on the infectivity of the virus and the efficacy of emerging therapeutics and vaccines has become a cornerstone of the fight against the disease. To support the maximal use of genomic information for SARS-CoV-2 research, we launched the Ensembl COVID-19 browser, incorporating a new Ensembl gene set, multiple variant sets (including novel variation calls), and annotation from several relevant resources integrated into the reference SARS-CoV-2 assembly. This work included key adaptations of existing Ensembl genome annotation methods to model ribosomal slippage, stringent filters to elucidate the highest confidence variants and utilisation of our comparative genomics pipelines on viruses for the first time. Since May 2020, the content has been regularly updated and tools such as the Ensembl Variant Effect Predictor have been integrated. The Ensembl COVID-19 browser is freely available at https://covid-19.ensembl.org.

INTRODUCTION

Over the past twenty years, multiple zoonotic respiratory diseases caused by coronaviruses have been identified. 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[1].

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).

Genomic sequencing has played a crucial role in understanding the mechanisms, spread and evolution of this virus. In the UK alone, at the time of writing, close to 5% of all reported infections each week were being sequenced (COG-UK, January 2021: https://www.cogconsortium.uk/wp-content/uploads/2021/02/COG-UK-geo-coverage_2021-02-01_summary.pdf) and this trend is likely to grow. Established genomic resources, such as Ensembl, have been able to leverage these data and bring them to new and existing user communities supporting research leveraging the rapidly emerging SARS-CoV-2 data landscape.

Ensembl[2,3] was launched to capture data from the Human Genome Project and has since developed into a large scale system for generating, integrating and disseminating genomic information. The COVID-19 pandemic presented new challenges related to presenting SARS-CoV-2 annotation and data within Ensembl. Meeting these, we launched the Ensembl COVID-19 browser (https://covid-19.ensembl.org) in May 2020 using concepts and workflows that enable rapid update cycles to react quickly in the face of new data and potential future outbreaks.

NEW ENSEMBL COVID RESOURCE

Reference assembly and a new gene annotation

The SARS-CoV-2 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[4]. It is widely used as the standard reference and has been incorporated into other resources such as the UCSC SARS-CoV-2 genome browser[5]. This assembly was imported from the European Nucleotide

65 Archive (ENA) into an Ensembl database schema with minor modifications to software regularly used
66 to integrate assemblies from the ENA into Ensembl.
67
68 To enable the correct annotation of SARS-CoV-2, the Ensembl gene annotation methods[6] were
69 adapted to reflect the biology of the virus. To identify protein coding genes, we aligned SARS-CoV-2
70 proteins from RefSeq[7] to the genome using Exonerate[8]. A challenge for annotation is that the first and
71 largest ORF can result in either non-structural proteins nsp1-11 (ORF1a) or in nsp1-nsp10 and
72 nsp12-nsp16 (ORF1ab) via an internal programmed translational frameshift[9]. Exonerate handles this
73 ribosomal slippage by inserting a gap in the alignment and thus allowing the annotation of the full
74 ORF1ab locus. Our modified annotation methodology then removes the artificial gap to represent the
75 slippage frameshift as an RNA edit and ensures a biologically accurate representation of the locus
76 and product.
77
78 Our annotation approach was tested on 90 additional SARS-CoV-2 assemblies retrieved from the
79 ENA. We assessed alignment coverage and percentage identity of the resultant gene translations to
80 verify accuracy and consistency. In all cases, full length alignments were observed and average
81 amino acid percentage identity across all genes in most assemblies were 99.9% or 100% (one
82 assembly had 99.81% identity). These results demonstrate that our annotation approach is able to
83 scale consistently to larger volumes of viral data.
84
85 In addition to generating a fully integrated Ensembl gene annotation, we also imported the gene set
86 submitted to the ENA with the reference sequence by the Shanghai Public Health Clinical Centre. As
87 shown in figure 1, both the submitted (blue) and the Ensembl gene annotations (red) can be viewed
88 simultaneously on the browser. The submitted gene annotation is displayed as a separate annotation
89 track, accessed under the 'Genes and transcripts' heading after clicking on 'Configure this page' in
90 the left-hand menu. The major difference between the annotations is that the submitted annotation
91 does not include the short form ORF1a or ORF7b.
92

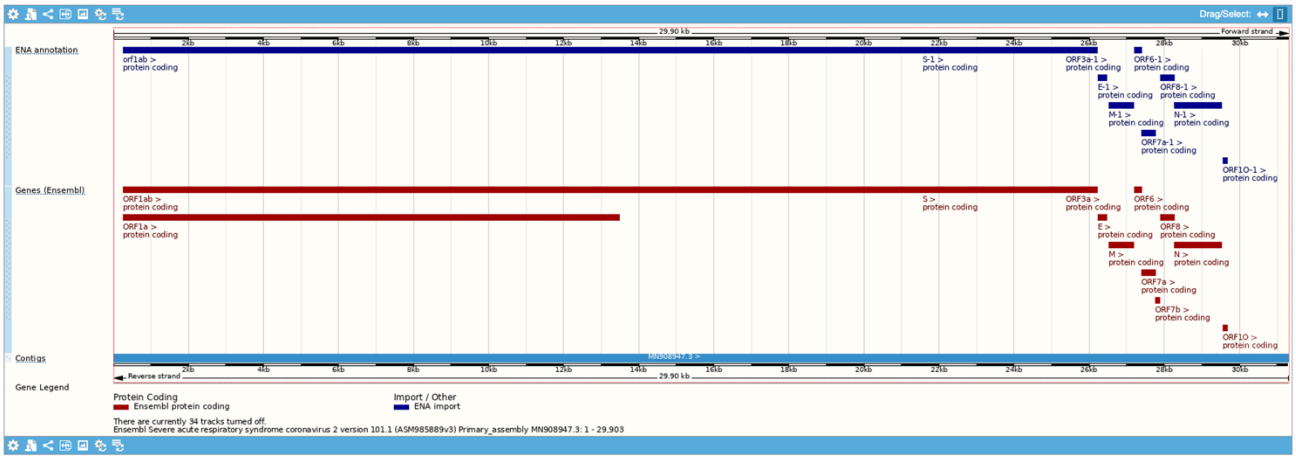


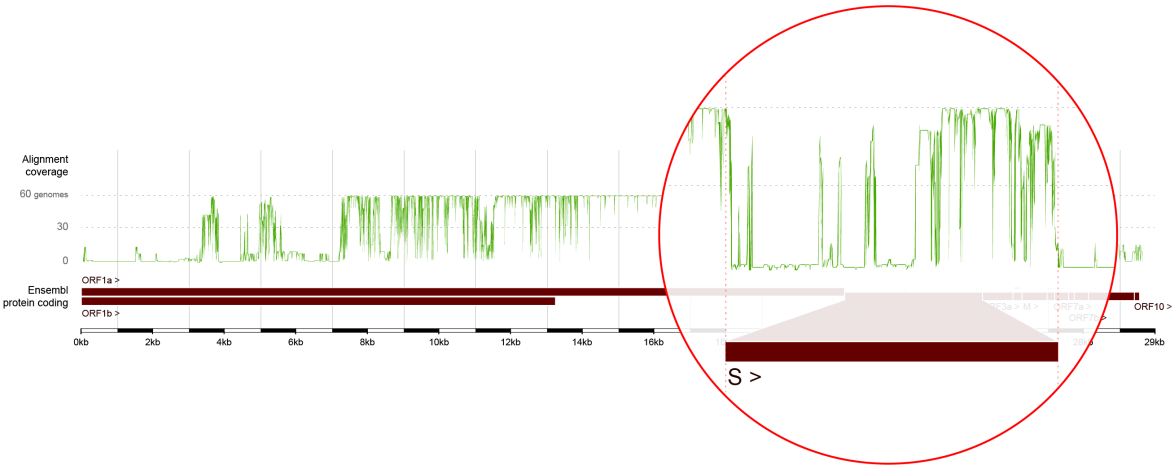
Figure 1: A comparison of the Ensembl gene set and the gene set submitted to the ENA by the Shanghai Public Health Clinical Centre for the SARS-CoV-2 reference assembly

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

We used Cactus[11] to align SARS-CoV-2 to 60 publicly available virus genomes from the *Orthocoronavirinae* subfamily. The results showed 78% of the SARS-CoV-2 genome aligned with at least one other genome and 35% of the genome aligned with the complete set of *Orthocoronavirinae* genomes. The multiple sequence alignment gives evolutionary context for each region of the genome and is a powerful method to explore functionality. For example, comparative genomics information such as this can be used for analyses such as a recent comparison of the gene sets of 44 complete *Sarbecovirus* genomes suggesting both a potentially novel alternate frame gene ORF3c and that ORF10, ORF9c, ORF3b and ORF3d are unlikely to be protein coding[10].

The alignment coverage (see figure 2) represents the number of genomes aligned to a given reference genomic position and is distributed heterogeneously across the SARS-CoV-2 genome. An immediate observation is that the central region of the genome (starting from ~7.1kb and ending at 21.3kb), including a significant segment of the 3' part of ORF1a, is highly shared across the *Orthocoronavirinae* subfamily. This indicates that the non-structural proteins encoded by this region (nsp3 - nsp16) likely originate from the *Orthocoronavirinae* ancestral genome. Conversely, both ends of the SARS-CoV-2 genome have very low alignment coverage and are only shared with closely related viruses. As a further demonstration of the utility of the alignment coverage, we focused in on the genomic region encoding for the SARS-CoV-2 spike protein (figure 2). The spike protein has two subunits: S1 which binds to the host cell receptor angiotensin-converting enzyme 2 (ACE2) and S2,

120 which is involved in membrane fusion. The region of the S ORF encoding for the S2 subunit of the
121 spike protein clearly displays a high alignment coverage while the region encoding for the S1 subunit
122 has large portions that are shared only by one other related genome. This demonstrates the dramatic
123 difference in conservations between the S1 and S2 subunits.
124



125
126
127 **Figure 2:** Alignment coverage across the SARS-CoV-2 reference genome based on a multiple sequence alignment with 60
128 other *Orthocoronavirinae* genomes. The green plot of alignment coverage shows that the central region of the genome is highly
129 shared across the subfamily, while the ends are generally shared only with closely related viruses. The region encoding for the
130 spike protein S has been highlighted within the red circle showing the difference between the low alignment coverage of the
131 upstream S1 subunit and the high coverage of the downstream S2 subunit.
132
133
134 Additionally, we applied our gene tree method[12] to group the protein coding genes into families and to
135 predict orthologous and paralogous relationships between genes. These results will be incorporated
136 into the COVID-19 resource in Q2 2021.

137 **Genetic variation data**

138 Analysis of genetic variants of viral genomes is important for understanding the spread of infection
139 across different geographic regions. We display 6,134 sequence variants for SARS-CoV-2 and show
140 their regional frequency distributions alongside predicted molecular consequences calculated by the
141 Ensembl Variant Effect Predictor (VEP)[13]. The variants on our site are derived from overlapping
142 sample sets produced by two groups and a small collection of variants of special interest.
143
144 One set comes from the Nextstrain project which creates phylogenetic trees for tracking pathogen
145 evolution based on virus subsamples[14]. We converted their SARS-CoV-2 data release from 08-04-

146 2020 to VCF for integration into our system and display frequency distributions by country and
147 Nextstrain-inferred clade.
148
149 The second variant set comes from the ENA team, who developed a LoFreq-based[15] pipeline to call
150 variants from SARS-CoV-2 sequence data sets submitted to their archives. LoFreq reports the
151 proportion of each variant seen in a sample from an individual. For simplicity, we represent only the
152 alleles seen in each sample and not the proportions estimated. Variants were called for each host
153 sample individually and, to provide a more accurate estimation of the frequency of each allele across
154 the entire sample set, it is assumed that sites at which a variant was not called in a sample match the
155 reference genome used in the Ensembl COVID-19 browser. We currently display ENA's variant data
156 from 17-08-2020 and have applied strict filters to reduce the proportion of lower confidence sites.
157 Specifically, we have not included variants from sequence data sets with more than 40 calls and we
158 have removed variants where no sample has a frequency of 20% or more for the non-reference allele
159 and variants where all samples show strand bias.
160
161 Some sites are annotated as a further guide to quality. For example, variants seen in more than one
162 sample in either set have an evidence status of 'Multiple observations' and variants at sites
163 recommended for masking by De Maio *et al*[16] have a flag of ‘Suspect reference location’. Variants can
164 be displayed as three separate tracks in the genome browser: those from ENA, those from Nextstrain
165 and those observed in more than one sample in either project as shown in figure 3.
166
167 We also display a set of variants which were reported as a tracking priority by the COVID-19
168 Genomics UK Consortium (COG-UK, https://www.cogconsortium.uk/) in December 2020. This
169 includes 17 variants from the rapidly spreading B.1.1.7 strain (https://virological.org/t/preliminary-
170 genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-
171 spike-mutations/563) and four variants from the mink associated strain. The D614G, A222V and
172 N439K mutations associated with an effect on transmissibility, a fast growing lineage and increased
173 binding affinity to the ACE2 receptor[17–19], respectively, have also been included. We extracted the
174 gene and protein change information from the reports and used the Ensembl VEP to map these
175 descriptions to genomic coordinates and create a VCF file, which was then loaded into the Ensembl

176 database with associated phenotype information. These variants can be viewed as a 'COG-UK
177 priority mutations' track alongside the gene annotations.

178 **Integration of data from other resources**

179 To enrich the SARS-CoV-2 genome annotation we aligned and integrated data from several external
180 repositories in a similar manner to other genomes available in Ensembl.

181 Specifically, we aligned Rfam[20] covariance models using their COVID-19 release 14.2
182 (http://rfam.xfam.org/covid-19) to highlight conserved non-coding RNA structures which are
183 responsible for various stages of the viral life cycle. These include the frame shifting stimulation
184 element and the pseudoknot necessary for the genome replication of SARS-CoV-2[21]. We also provide
185 cross references to proteins from RefSeq, UniProt[22] and the International Nucleotide Sequence
186 Database Collaboration (INSDC); functional annotation from the Gene Ontology Consortium; and
187 annotation of protein domains using InterProScan. These additional annotations are accessible via
188 our region views and the gene and transcript tabs. We also created a genome browser track
189 projecting the protein-domain annotations onto the genome to facilitate a genome-oriented view of the
190 gene products including the non-structural cleavage products of ORF1a/ORF1ab.
191
192 The browser also displays community annotation of sites and regions using results co-ordinated by
193 the UCSC genome browser. Additions to this annotation resource are open to all and done via a
194 publicly available spreadsheet hosted by UCSC (http://bit.ly/cov2annots), the data from which is
195 integrated periodically into the Ensembl browser. This is achieved via specialised code that uses Git
196 workflows to convert the annotations into BigBed files that can be visualised on a variety of genome
197 browsers (available freely at https://github.com/Ensembl/sarscov2-annotation).

198 We have integrated Oxford Nanopore sequencing primers (version 3) made available by the ARTIC
199 network (https://artic.network/ncov-2019) to assist in sequencing the virus. Though mainly focused on
200 the Oxford Nanopore MinION sequencer, some aspects of the protocol may be generalised to other
201 sequencing platforms. The complete list of primers included is available on GitHub
202 (https://github.com/artic-network/artic-ncov2019/blob/master/primer_schemes/nCoV-2019/V3/nCoV-
203 2019.tsv).

Finally, we provide tracks to visualise problematic and caution sites, which result from common systematic errors associated with laboratory protocols and have been observed in submitted sequences[16]. Inclusion of these can adversely influence phylogenetic and evolutionary inference. Visualising these in the browser alongside the locations of primers and other community derived annotations helps determine how best to proceed with analyses of each these sites.

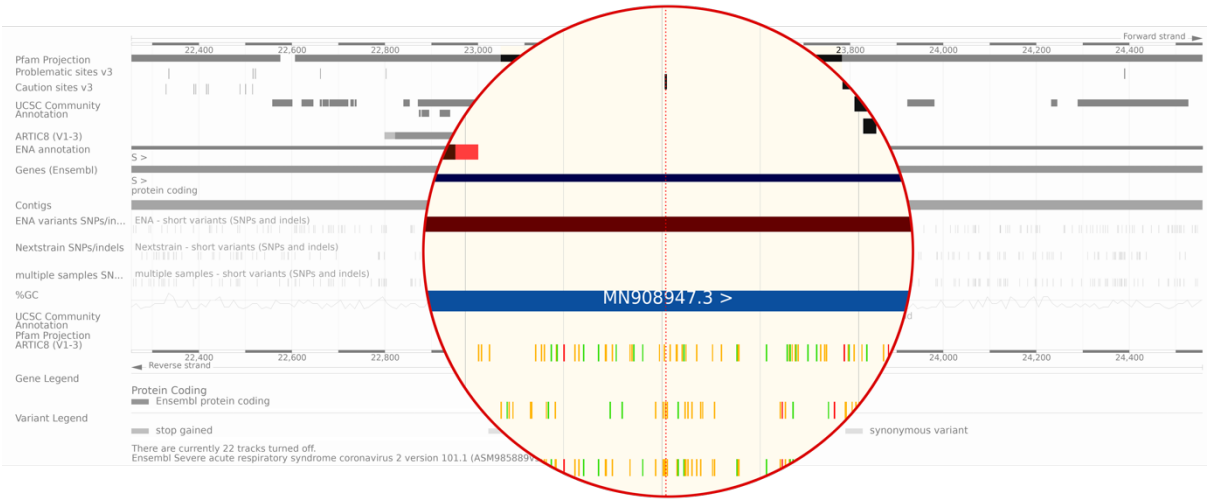


Figure 3: The browser with several tracks turned on and highlighting a substitution flagged up early on in the UCSC community annotation at position 23403 (D614G) in the S spike glycoprotein gene. Due to the prompt nature of community driven annotation, this data was available on our browser as soon 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 spike protein is predominantly observed in Europe[23]; patterns that can also be seen in the variation data we host.

Integration and engagement

The Ensembl COVID-19 resource features a newly designed landing page, which prioritises key views and data to help direct researchers into relevant sections of the site. To support expeditious data release, we have not made potentially time-consuming 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 understood without these changes. However, we have altered the vocabulary wherever possible and are reviewing feedback as we receive it.

228 Our COVID-19 resource is also integrated into the European COVID-19 Data Portal hosted by EMBL-
229 EBI (https://www.covid19dataportal.org/). The portal enables searches across the multiple research
230 outputs on COVID-19 including viral and human sequences; relevant biochemical pathways,
231 interactions, complexes, targets and compounds; protein and expression data; and literature.
232
233 We have engaged our existing and new user communities using our blog and social media accounts
234 to announce the release and updates to the Ensembl COVID-19 resource. We also highlighted
235 the changes made to our gene annotation method to ensure the complete set of ORFs because these
236 have been overlooked by other annotation tools.
237

238 **DISCUSSION**

239 The swift spread of COVID-19 has highlighted the necessity for data resources to be prepared for
240 rapid adaptation to developing outbreaks. Our development and release of the Ensembl COVID-19
241 resource leveraged our experience integrating thousands of genomes into the Ensembl infrastructure
242 and supporting hundreds of thousands of users. The Ensembl COVID-19 browser provides a unique
243 view on SARS-CoV-2 using our gene annotation method and variation data processed to focus on the
244 highest confidence variants. Additionally, the Ensembl VEP and haplotype views enable the
245 consequences of the variants to be assessed within the context of specific strains and geographical
246 locations. The data is made accessible via the widely used Ensembl platform making it immediately
247 familiar to a large userbase who may be able to repurpose existing software and browser knowledge
248 to support their work during the pandemic and beyond.
249
250 When the COVID-19 pandemic hit, we had been working for several months to develop Ensembl
251 Rapid Release (https://rapid.ensembl.org) to distribute annotated genomes within days of their
252 annotation being completed. This experience proved useful in bringing the COVID-19 site to public
253 release quickly. We have also demonstrated the flexibility of the Ensembl infrastructure and its value
254 as a platform for research and discovery. Indeed, all of our pipelines and schemas worked seamlessly
255 even though Ensembl was not designed to support RNA genomes and had not previously been used
256 for viruses. The adapted gene annotation method, for instance, produced consistent annotation with
257 ribosomal slippage correctly modelled and can be reused in the future. Similarly, the gene tree and

258 alignment pipelines have been applied to the viral data with only minimal changes to parameters. We
259 will continue to regularly update the site as new data emerges to support research into understanding
260 the genomic evolution of this virus, identifying hotspots of genomic variation and enabling the rational
261 design of future therapeutics, vaccines and policies well beyond the end of the current pandemic.
262
263
264 **REFERENCES**

265 1. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV)
266 Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
267 2. Howe, K. L. *et al.* Ensembl Genomes 2020—enabling non-vertebrate genomic research.
268 *Nucleic Acids Res.* **48**, D689–D695 (2020).
269 3. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
270 4. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*
271 **579**, 265–269 (2020).
272 5. Fernandes, J. D. *et al.* The UCSC SARS-CoV-2 Genome Browser. *Nat. Genet.* **52**, 991–998
273 (2020).
274 6. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* **2016**, baw093 (2016).
275 7. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
276 taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
277 8. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence
278 comparison. *BMC Bioinformatics* **6**, (2005).
279 9. Chen, Y., Liu, Q. & Guo, D. Emerging coronaviruses: Genome structure, replication, and
280 pathogenesis. *J. Med. Virol.* **92**, 418–423 (2020).
281 10. Jungreis, I., Sealfon, R. & Kellis, M. SARS-CoV-2 gene content and COVID-19 mutation
282 impact by comparing 44 Sarbecovirus genomes. *bioRxiv (preprint)* (2020)
283 doi:10.1101/2020.06.02.130955.
284 11. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome
285 era. *Nature* **587**, 246–251 (2020).
286 12. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016**, bav096 (2016).
287 13. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

288 14. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–
289 4123 (2018).
290 15. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering
291 cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.*
292 **40**, 11189–11201 (2012).
293 16. Nicola, D. M. *et al.* Issues with SARS-CoV-2 sequencing data. https://virological.org/t/issues-
294 with-sars-cov-2-sequencing-data/473 (2020).
295 17. Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on
296 Transmissibility and Pathogenicity. *Cell* **184**, 64-75.e11 (2021).
297 18. Bartolini, B. *et al.* The newly introduced SARS-CoV-2 variant A222V is rapidly spreading in
298 Lazio region, Italy. *medRxiv (preprint)* doi:10.1101/2020.11.28.20237016.
299 19. Thomson, E. *et al.* The circulating SARS-CoV-2 spike variant N439K maintains fitness while
300 evading antibody-mediated immunity. (2020) doi:10.1101/2020.11.04.355842.
301 20. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families.
302 *Nucleic Acids Res.* **49**, D192–D200 (2021).
303 21. Williams, G. D., Chang, R.-Y. & Brian, D. A. A Phylogenetically Conserved Hairpin-Type 3′
304 Untranslated Region Pseudoknot Functions in Coronavirus RNA Replication. *J. Virol.* **73**, 8349
305 LP – 8355 (1999).
306 22. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*
307 **49**, D480–D489 (2021).
308 23. Isabel, S. *et al.* Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein
309 mutation now documented worldwide. *Sci. Rep.* **10**, 14031 (2020).
310

311 **ACKNOWLEDGEMENTS**

312 We would like to thank the following people at the EMBL-EBI for their contributions to our resource
313 and thoughtful discussions: Nick Goldman, Zamin Iqbal, Guy Cochrane, Rodrigo Lopez Sanchez,
314 Conor Walker, Nadim Rahman, Jeena Rajan, Alexy Sokolov, Peter Harrison, Youngmi Park, Nicola
315 Buso, Suran Jayathilaka, Anton Petrov, James Allen, Luca Da Rin Fioretto, Thomas Maurel and Vinay
316 Kaikala.
317

This work was supported by the Wellcome Trust [WT108749/Z/15/Z] and the European Molecular Biology Laboratory. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

AUTHOR CONTRIBUTIONS

N.H.D.S., R.D.F., P.F., A.F., K.L.H., S.E.H., F.J.M., M.R., A.D.Y. and D.R.Z. conceptualised the resource. N.H.D.S., S.E.H., F.J.M., M.R. and A.D.Y. contributed to the methodology. M.C., B.C., C.C., T.G., S.E.H., F.J.M., D.N.O., A Parker, A Parton, M.R., M.P.S., D.S., J.T. and A.T. developed the software. M.C., C.C., N.H.D.S., A.G., T.G., M.R., D.T. and A.D.Y. validated the data while C.C., T.G., K.L.H., S.E.H., M.R. and D.T. conducted formal analysis on the computed results. C.C., T.G., D.N.O. and D.T. helped with investigations of software and results. N.H.D.S. wrote the original draft of this manuscript and R.D.F., P.F., A.F., A.G., K.L.H., S.E.H., B.M., A Parker, M.R., D.T., S.J.T., A.D.Y. and D.R.Z. reviewed and edited it. A.W. created the visualisation for the resource landing page. N.H.D.S., R.D.F., P.F., K.L.H., S.E.H., M.R., S.J.T. and A.D.Y. supervised various aspects of the project. M.R. and A.D.Y. were involved in project administration and K.L.H., P.F., A.D.Y. and D.R.Z. acquired funds to support the project.

COMPETING INTERESTS

P.F. is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

DATA AVAILABILITY

The COVID-19 resource from Ensembl is available without restrictions at https://covid-19.ensembl.org. The reference genome assembly for SARS-CoV-2 with the accession GCA_009858895.3 was obtained from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/view/GCA_009858895.3).

CODE AVAILABILITY

The selection of our code to convert CSV files into BigBed files is at https://github.com/Ensembl/sarscov2-annotation. The code relevant to processing SARS-CoV-2 variants in Ensembl is at https://github.com/Ensembl/ensembl-variation, the gene annotation pipeline is available at https://github.com/Ensembl/ensembl-annotation and the code used for comparative analysis is at https://github.com/Ensembl/ensembl-compara