SARS-CoV-2 (Fig. 4), in agreement with the observation that the N-terminal end of the S protein is one of the most divergent regions among betacoronaviruses[3].

All analyzed viral sequences are available from the NCBI GenBank[7] (https://www.ncbi.nlm.nih.gov/nuccore), GISAID[8] (https://www.gisaid.org/) and Nextstrain[9] (https://nextstrain.org/sars-cov-2) public repositories, with the exception of the pangolin viral metagenomic dataset[18], which is available at NCBI BioProject PRJNA573298). Additional datasets used to create public data hubs hosted in the browser are listed in Supplementary Table 1.

Notably, the UCSC SARS-CoV-2 Genome Browser has recently been developed in parallel to the work described here[19], highlighting the need for comprehensive omic visualization resources as well as community interest and contribution. We hope that the WashU Virus Genome Browser will enable rapid sharing of processed data, facilitate collaboration and accelerate research on existing and novel pathogenic viruses. Moreover, the portable nature of the underlying technology enables us to swiftly spin up viral browser instances in response to other emerging zoonotic viruses. Our browser portal can be accessed at https://virusgateway.wustl.edu; documentation is available at https://virusgateway.readthedocs.io/; and general feedback, suggestions and bug reports may be sent to https://github.com/twlab/virusbrowser/issues.

Jennifer A. Flynn[1,3], Deepak Purushotham[1,3], Mayank N. K. Choudhary[1,3], Xiaoyu Zhuo[1,3], Changxu Fan[1,3], Gavriel Matt[1,3], Daofeng Li[1,4] ✉ and Ting Wang[1,2,4] ✉

[1]The Edison Family Center for Genome Sciences & Systems Biology, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [2]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. [3]These authors contributed equally: Jennifer A. Flynn, Deepak Purushotham, Mayank N. K. Choudhary, Xiaoyu Zhuo, Changxu Fan, Gavriel Matt. [4]These authors jointly supervised this work: Daofeng Li, Ting Wang.

✉*e-mail: dli23@wustl.edu; twang@wustl.edu*

Published online: 9 September 2020
https://doi.org/10.1038/s41588-020-0697-z

References

1. de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
2. Cui, J., Li, F. & Shi, Z. L. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
3. Zhou, P. et al. *Nature* **579**, 270–273 (2020).
4. Kim, D. et al. *Cell* **181**, 914–921.e10 (2020).
5. Blanco-Melo, D. et al. *Cell* **181**, 1036–1045.e9 (2020).
6. Bojkova, D. et al. *Nature* **583**, 469–472 (2020).
7. NCBI Resource Coordinators. *Nucleic Acids Res.* **46**, D8–D13 (2018).
8. Shu, Y. & McCauley, J. *Eur. Surveill.* **22**, 30494 (2017).
9. Hadfield, J. et al. *Bioinformatics* **34**, 4121–4123 (2018).
10. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. *Nucleic Acids Res.* **47**, W158–W165 (2019).
11. Zhou, X. et al. *Nat. Biotechnol.* **33**, 345–346 (2015).
12. Zhou, X. et al. *Nat. Methods* **10**, 375–376 (2013).
13. Zhou, X. et al. *Nat. Methods* **8**, 989–990 (2011).
14. van der Made, C. I. et al. *JAMA* **324**, 1–11 (2020).
15. Kent, W. J. et al. *Genome Res.* **12**, 996–1006 (2002).
16. Vita, R. et al. *Nucleic Acids Res.* **47**, D339–D343 (2019).
17. Korber, B. et al. *Cell* **182**, 812–827 e819 (2020).
18. Liu, P., Chen, W. & Chen, J. P. *Viruses* **11**, 979 (2019).
19. Fernandes, J.D. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.04.075945 (2020).

Acknowledgements

J.A.F. is supported in part by the Siteman Cancer Center Precision Medicine Pathway (T32CA113275). X.Z. is supported in part by NIH 5R25DA027995. T.W. is supported by NIH grants R01HG007175, U24ES026699, U01CA200060, U01HG009391 and U41HG010972, and by American Cancer Society Research Scholar grant RSG-14-049-01-DMC.

Author contributions

Conceptualization, T.W.; web development, D.L. and D.P.; SNV track development, J.F. and C.F.; dynamic track development, M.N.K.C. and D.L.; immune-epitope analysis, M.N.K.C.; data download, metadata generation and annotation, G.M., D.L. and C.F.; data-hub preparation, J.F., C.F., M.N.K.C. and G.M.; sequence alignment and tree generation, C.F. and X.Z.; manuscript preparation, J.F., C.F., M.N.K.C., G.M., D.L. and T.W.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0697-z.

Check for updates

The UCSC SARS-CoV-2 Genome Browser

The UCSC SARS-CoV-2 Genome Browser (https://genome.ucsc.edu/covid19.html) is an adaptation of our popular genome-browser visualization tool for this virus, containing many annotation tracks and new features, including conservation with similar viruses, immune epitopes, RT–PCR and sequencing primers and CRISPR guides. We invite all investigators to contribute to this resource to accelerate research and development activities globally.

Jason D. Fernandes, Angie S. Hinrichs, Hiram Clawson, Jairo Navarro Gonzalez, Brian T. Lee, Luis R. Nassar, Brian J. Raney, Kate R. Rosenbloom, Santrupti Nerli, Arjun A. Rao, Daniel Schmelter, Alastair Fyfe, Nathan Maulding, Ann S. Zweig, Todd M. Lowe, Manuel Ares Jr, Russ Corbet-Detig, W. James Kent, David Haussler and Maximilian Haeussler

The University of California, Santa Cruz (UCSC) Human Genome Browser[1], a web-based, interactive viewer for human and other vertebrate genome sequences featuring research data, clinical molecular data, annotations and sequence alignments, has been used for almost 20 years by hundreds of thousands of biomedical researchers and cited in more than 37,500 scientific articles. To address the current COVID-19 epidemic, we have built a similar browser for the SARS-CoV-2 reference genome (NC_045512v2, wuhCor1). Here, we provide an overview of this tool for the international community racing to understand the details of the virus, its evolution, its mechanisms of action in human cells, and its immunological and molecular vulnerabilities.

A brief introduction to the Genome Browser

The UCSC SARS-CoV-2 browser, like alternative genome browsers[2–4], displays the reference nucleotide sequence of the viral genome and provides an intuitive

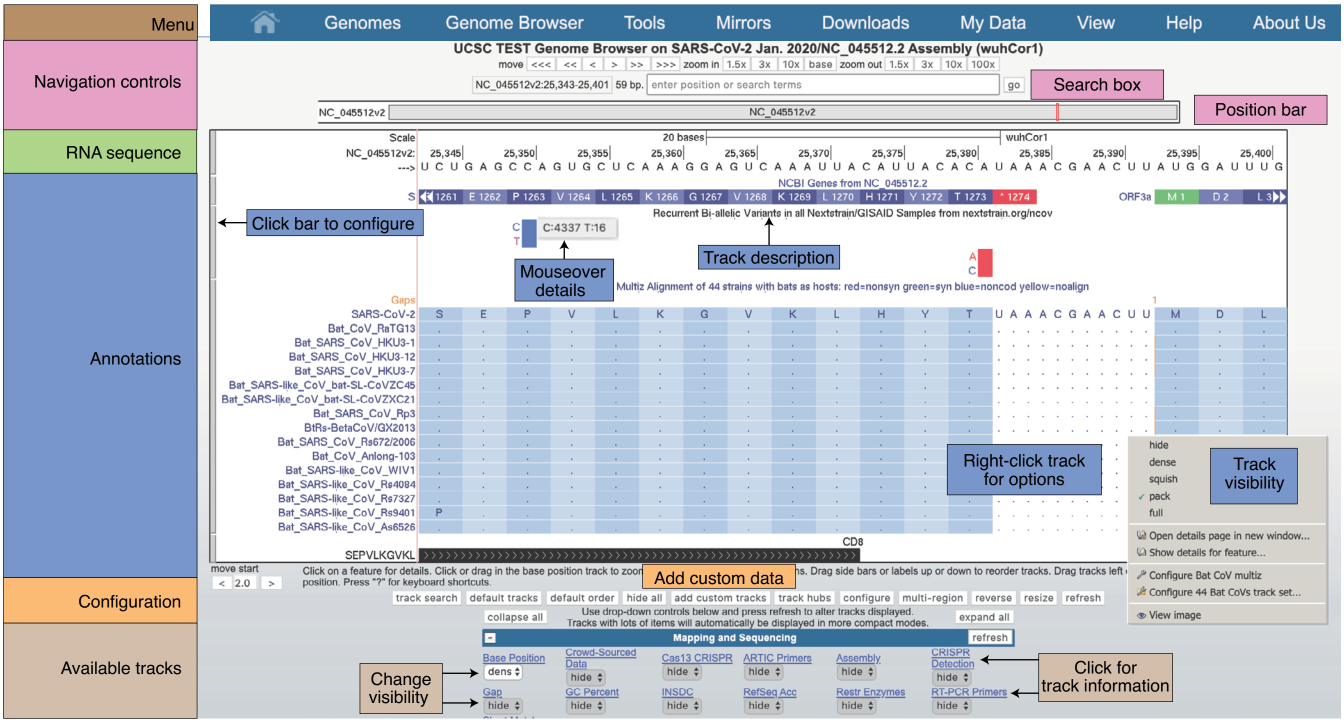

Fig. 1 | An overview of the UCSC Genome Browser user-interface structure. Navigation controls at the top allow users to move left and right and to zoom. The position bar shows a highlighted red box illustrating the current portion of the genome being viewed. The search box allows users to search for particular features or to move to exact genomic coordinates. The RNA sequence is shown only when the view is sufficiently zoomed in. Annotations are shown for data tracks that have been set to visible. Here, the NCBI Genes track shows the annotation of the end of the S protein and the start of *orf3a*, as well as the amino acid translation of their codons. Below that, a track shows recurrent SARS-CoV-2 variants that have been observed worldwide, as reported by Nextstrain. Bar graphs show the frequency of each allele, and mouseover shows the counts of each allele. The next track (Bat CoV multiz) shows a multiple alignment of 44 bat coronaviruses aligned to the reference. Overall, these viruses align well to this region of SARS-CoV-2 (dots indicate identical amino acids) although one non-synonymous substitution (p.Ser1261Pro) is observed in one virus, Rs9401. The final track shows a CD8$^+$ epitope from IEDB. Tracks can be configured with a right-click or alternatively by clicking on their name near the bottom of the page. Only 12 of the 50 currently available track-configuration buttons are shown. Custom data tracks generated by users can be added directly via the 'add custom tracks' button. Additional options can be set via the menu bar at the top (for example, the view menu allows for additional changes to the browser window). (A live interactive session for this figure can be accessed at http://genome.ucsc.edu/s/SARS_CoV2/Figure1.).

way to visualize annotations or data on specific parts of the genome (Fig. 1). The genome sequence is shown from left to right, 5′ to 3′, as an image with the label NC_045512v2, which is the National Center for Biotechnology Information (NCBI)/International Nucleotide Sequence Database Collaboration (INSD) accession ID for the reference sequence. This reference sequence is the RNA genome isolated from one of the first cases in Wuhan, China, and is known as 'Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome'[5]. This sequence is widely used as a standard reference and, because of its early identification, is used as the root genome in phylogenetic trees produced by Nextstrain[6], COVID-19 Genomics UK (COG-UK)[7] and the China National Center for Bioinformation project[8]. Above this image are the navigation controls. Buttons in the navigation controls allow users to move left and right along the genomic sequence and zoom in and out by various factors. Clicking the zoom factor 'base' zooms in to show the nucleotide bases in the selected region. The position bar shows the current region of the genome being viewed via a red box, and the search box allows users to manually enter a specific position to view. Coordinates in the position bar should be entered with the prefix NC_045512v2 (the NCBI RefSeq accession of this genome) followed by the stop and start nucleotide numbers (for example, NC_045512v2:25,341-25,401). Users can also directly enter a nucleotide sequence or the name of a gene (corresponding to the NCBI/INSDC annotation) to go directly to that region of the genome.

Genome annotations are shown underneath the position bar. At the top of the annotations, the genome sequence is shown (when zoomed in) as an RNA sequence. At the time of writing, there are 50 tracks with different types of molecular information, 11 of which are shown by default. The small gray buttons on the left of the image are the track-configuration buttons, which allow users to configure how track information is displayed. Alternatively, users can right-click to configure the track.

Under the genome sequence in the image, gene and protein annotations are shown by default as blue rectangles, with the direction of transcription indicated by small arrows. If activated, additional tracks displaying alignments and variants are shown underneath.

Under the annotations display, various buttons allow users to reset the currently shown tracks back to the defaults. Two buttons, 'add custom tracks' and 'track hubs'[9], allow experienced users to add their own data. The 'configure' button allows users to make viewing adjustments (for example, increase the font size). The 'reverse' button reverse-complements the display, to show the antisense sequence from 3′ to 5′,

Table 1 | Important tools of the UCSC Genome Browser, accessible through the menu bar

My Data → My Sessions	The current view (session) can be saved to a link that can be shared with others. Other users can explore the session interactively and make adjustments without affecting the original session.
View → PDF/PS	The current view can be downloaded to a PDF for publications and presentations.
View → DNA	The genome sequence currently visible within the browser window can be downloaded. Although the viral sequence is viewable as RNA (with U instead of T), the downloaded sequence uses the DNA nucleotides (T instead of U).
Tools → BLAT	BLAT[45], the BLAST-like alignment tool, allows users to quickly find the coordinates of a short nucleotide or protein sequence encoded within the genome.
Tools → Table Browser	The table browser is a simple interface to download the data stored within the annotation tracks as spreadsheet tables or common genomics formats (for example, BED or GFF). Users can also perform basic analysis, such as intersecting tracks and displaying them as new tracks. API or downloads are described in the main text.

and the 'resize' button fits the image to the current screen size.

Below the buttons, the track list shows all available tracks (the first 12 are shown in Fig. 1). To provide a compact display, most tracks are hidden by default. For example, in Fig. 1, only 4 of the more than 50 currently available tracks are visible. Hovering the mouse over the title of a track for several seconds reveals a longer description of the track data. Clicking the title of the track shows a full description and configuration page for the track data. On this page, or in the track list, tracks can be set to one of the four visibility modes: 'dense', 'squish', 'pack' and 'full' (Supplementary Fig. 1). Depending on the track, different viewing modes may emphasize features of the data not immediately apparent in other modes (for example, in Supplementary Fig. 1, individual open reading frames (ORFs) are hidden in dense mode, and names are hidden in squish mode). In general, setting a track to pack mode provides a good starting point to explore the data. At the top of the page is the menu bar, which contains several tools helpful for working with the genome sequence (Table 1).

Genomic organization of SARS-CoV-2

The reference SARS-CoV-2 genome (NCBI RefSeq NC_045512.2) is a single strand of 29,903 RNA nucleotides, yet, like many viral genomes, it encodes substantial molecular complexity, generating ~10 canonical RNA transcripts, ~14 ORFs and ~29 proteins. This complex organization has several features that are atypical of standard genomic analyses.

The SARS-CoV-2 transcriptome

Like other coronaviruses, SARS-CoV-2 is a single-stranded positive-sense RNA, which shares many features with the messenger RNAs of most human genes, including 5′ and 3′ untranslated regions as well as a 3′-poly(A) tail[10]. In infected cells, the first gene (*orf1a*/*orf1ab*) is directly translated and cleaved into proteins that form the replication–transcription complexes (RTCs)[11]. These RTCs use the same positive-strand RNA genome as a template to generate negative-strand RNAs. During negative-strand transcription, RTCs occasionally encounter a body transcription-regulatory sequence (TRS-B). TRS-B sequences work in combination with a single leader transcription-regulatory sequence (TRS-L) at the 5′ end of the genome. When RTCs encounter a TRS-B sequence, they can 'jump' to the TRS-L sequence via long range RNA–RNA interactions, thereby generating a negative-strand RNA with a large portion of the genome omitted[12] (Fig. 2a). These negative-strand RNAs then serve as templates for the transcription of positive-strand subgenomic mRNAs. Positive-strand subgenomic RNAs created through this mechanism have downstream AUGs in an optimal start-codon context, thus allowing ribosomes to initiate translation at locations normally not available in the full-length genomic RNA. These subgenomic RNAs serve as mRNAs for viral proteins encoded downstream in the genome.

In the SARS-CoV-2 browser, we have included a transcriptome track, which annotates each predicted TRS site in the reference genome according to the presence of the motif ACGAAC, the reported core TRS for SARS-CoV[13]. The track also includes the canonical mRNAs produced from the full-length positive-strand RNA (Fig. 2b). Another track, Kim transcripts, shows subgenomic mRNAs that have been experimentally validated by transcriptomic sequencing[14]. In addition, the Kim-transcripts track includes annotations of several recently reported experimentally observed subgenomic RNAs, some of which have non-canonical junctions[14]. As the SARS-CoV-2 transcriptome continues to be elucidated, we will update and add appropriate tracks, including recent reports of variant ORFs[15].

The SARS-CoV-2 proteome

SARS-CoV-2 proteome consists of two polyproteins, four structural proteins and possibly nine accessory proteins[5,16,17]. The two polyproteins are processed into 16 non-structural proteins, thus requiring consideration of as many as 29 proteins in analyses (Fig. 2c).

The polyproteins pp1a and pp1ab are products of *orf1a* and *orf1ab*, respectively, both of which are produced by translation of the full-length genomic RNA. To generate pp1a, the ribosome initiates translation at the AUG codon at nucleotides 266–268. Translation continues in the canonical manner until a UAA stop codon is encountered at nucleotides 13481–13483, thus producing the 4,405–amino acid pp1a polyprotein. This polyprotein contains two viral proteases (nsp3/PL-PRO) and (nsp5/3CL-PRO), which cleave the pp1a polyprotein into 11 mature non-structural proteins (nsp1–11)[18,19], which are shown in the UniProt/mature proteins and NCBI Proteins tracks (Fig. 2d).

The pp1ab polyprotein is generated from translation initiation at the same start codon as pp1a; however, the viral RNA contains a 'slippery' sequence and structured RNA element that occasionally cause the translation machinery to slip near nucleotide C13468 and therefore read this nucleotide twice: once as the final nucleotide in the AAC codon for amino acid p.Asn4401 and once again as the first nucleotide in the CGG codon for amino acid p.Arg4402 (refs. [20,21] and Fig. 3a). The result is a 'programmed' −1 frameshift, with the pp1a 'stop' codon no longer in frame, thus lengthening pp1ab by 2,695 entirely different amino acids. These additional amino acids encode essential non-structural proteins (Figs. 2b and 3); thus, pp1ab encodes nsp1–10 as well as nsp12–16. To properly display this frameshift in the genome browser, we use specialized colored codons that lead to faithful translation of each ORF (Fig. 3a) in the track NCBI Genes.

The remaining proteins are produced through traditional ribosomal recognition of AUG sequences in subgenomic RNAs (Fig. 2). Of particular interest is the spike protein (S), which is the target of most immunology-based therapies. The S protein governs the entry of the

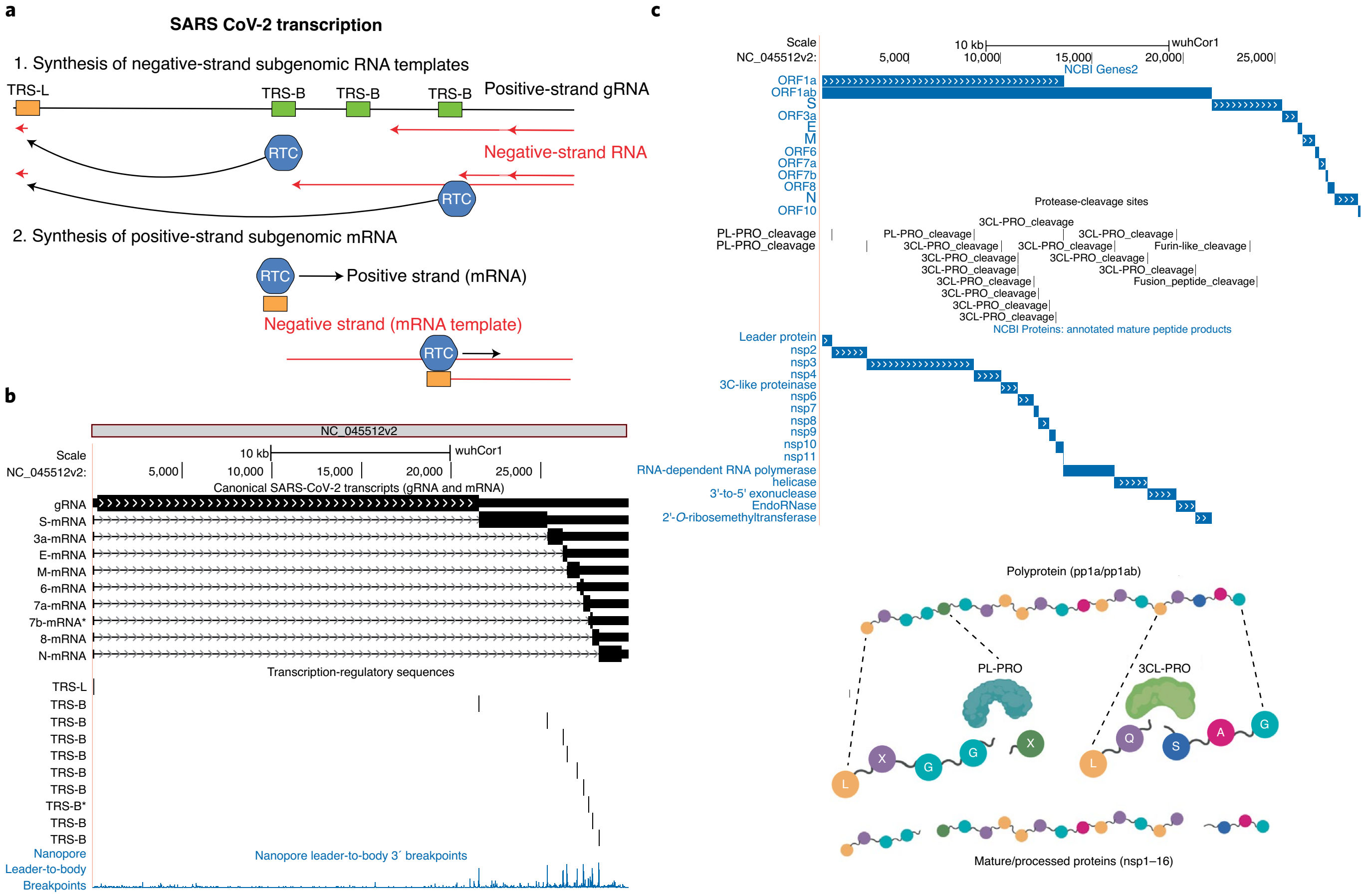

Fig. 2 | Molecular and genomic visualizations of SARS-CoV-2 transcription and protein cleavage in the browser. a, Discontinuous transcription. RTCs initiate generation of negative-strand RNA (red) from the positive strand (black). When RTCs encounter a TRS-B, they are able to jump to the TRS-L. The jumping process generates different negative-strand RNAs that lack regions between the various TRS-Bs and the TRS-L. These negative strands then serve as templates for positive transcription from the TRS-L to the viral mRNAs for different proteins. **b**, Genomic visualization of discontinuous transcription. Top, viral mRNAs are shown as genome annotations. Black bars, nucleotides present in an mRNA species; arrows, sequence skipped by discontinuous transcription; thick black bars, predicted coding sequence in these RNA species. Middle, the core TRS motif, ACGAAC, annotated on the genome, corresponds to transcript junctions. Bottom, experimental data representing breakpoints, fusions of TRS-B to TRS-L sequence identified by nanopore direct RNA sequencing[14]. High peaks indicate that the 5′ TRS-L sequence is found directly upstream of the annotated bases. Most of these breakpoints correlate with TRS-B motifs (live interactive session: http://genome.ucsc.edu/s/SARS_CoV2/Figure2b). **c**, Cleavage of polyprotein peptide sequences by the viral proteases. The viral polyproteins are cleaved by the PL-PRO protease at three locations that match the amino acid pattern LXGGX (where X is any amino acid) as indicated; 3CL-PRO cleaves many more sites, typically at the sequence LQSAG, as shown. **d**, Genomic visualization of cleavage. The NBCI Genes track shows all annotated viral ORFs. Below, a track showing annotated cleavage sites, including two sites in the S protein (furin_like_cleavage and fusion_peptide_cleavage) that are recognized by host cellular proteases instead of the above two viral proteases. At the bottom, mature protein products that result from cleavage of the viral polyproteins by nsp3/PL-PRO and nsp5/3CL-PRO. The cleavage products allow the virus to enter cells (live interactive session: http://genome.ucsc.edu/s/SARS_CoV2/Figure2d).

virus into the cell and is cleaved by host proteases. Other therapeutic targets include the RNA-dependent RNA polymerase (RdRp) protein, which makes copies of the viral RNA genome (also known as Pol/nsp12), the virally encoded proteases 3CL-PRO and PL-PRO, the viral envelope protein E, the membrane protein M and the nucleocapsid protein N, which organizes the RNA genome of the virus in viral particles.

SARS-CoV-2 has at least five accessory proteins encoded by *orf3a*, *orf6*, *orf7a*, *orf7b* and *orf8*. Additional accessory ORFs not in the NCBI and UniProt SARS-CoV-2 annotations have been observed in other coronaviruses (*orf3b*, *orf9b* and *orf9c*), although whether these ORFs produce functional proteins in SARS-CoV-2 is unclear[22,23]. Additionally, *orf10*, present in most annotation sets, has little experimental support as a protein-coding gene[14,24]. A variety of variant ORFs have also been recently reported for non-canonical subgenomic mRNAs[15].

An overview of SARS-CoV-2 genome annotation tracks

We provide several standard annotation tracks based on molecular data generated by experimental and computational analyses of the SARS-CoV-2 genome. These annotations are sorted into groups as described below.

Mapping and sequencing. Tracks in this group are all based on short segments of local nucleotide composition. We have also added tracks specific to this viral genome: COVID-19 RT–PCR primers

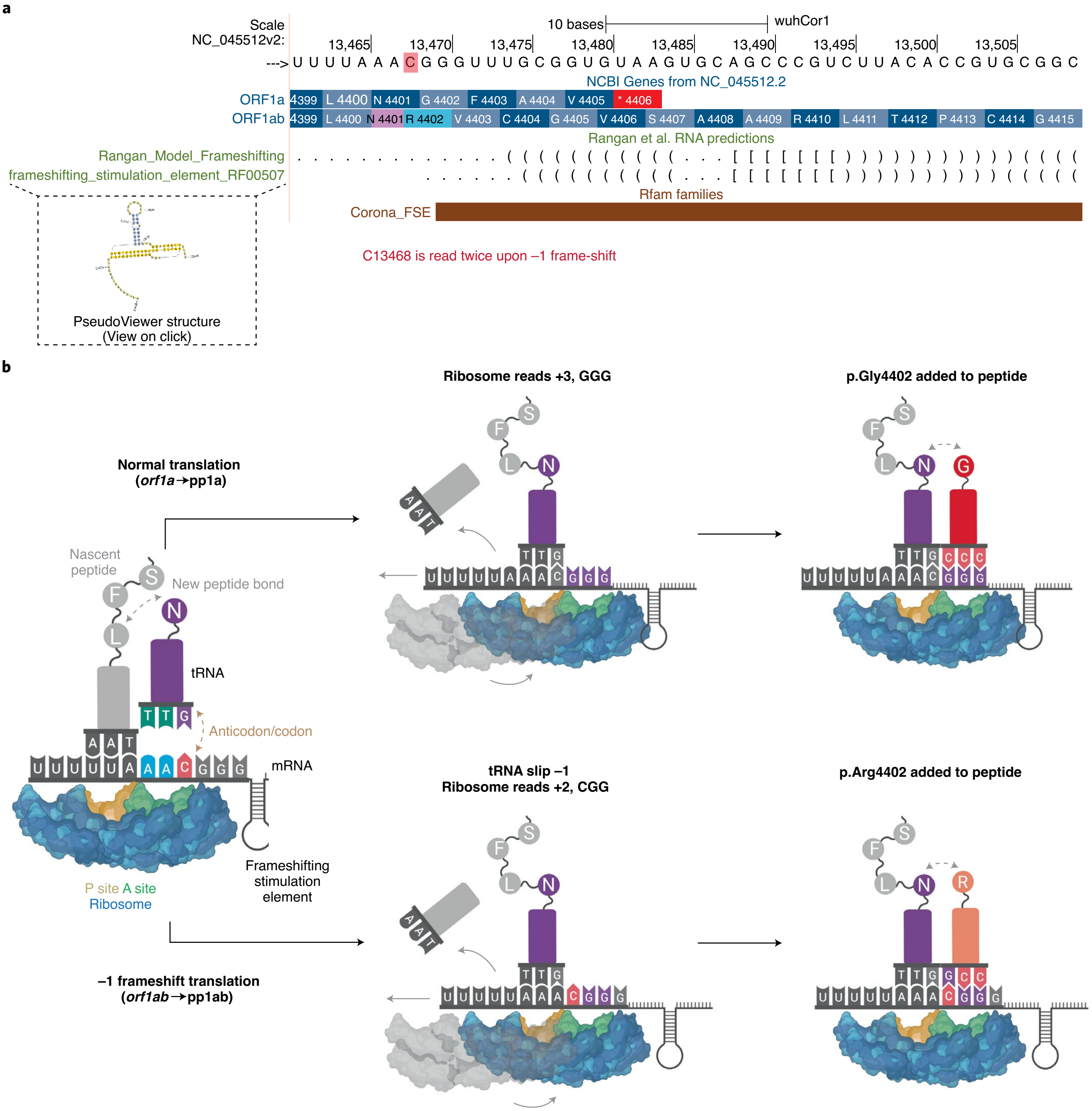

Fig. 3 | *orf1a/orf1ab* ribosomal frameshifting. a, Genome Browser annotations detailing translation of *orf1a* and *orf1ab* via ribosomal frameshifting. The red highlighted C is read twice by the ribosome, owing to the upstream poly(U) tract and the downstream frameshifting RNA structure annotated in the RFAM and RNA predictions track. Predicted base-pairing reported by Rangan et al.[30] and putative secondary structures are visible after clicking on annotations. Tertiary interactions are not shown (live interactive session: http://genome.ucsc.edu/s/SARS_CoV2/Figure3a). **b**, Schematic representation of ribosomal frameshifting to generate distinct protein products from *orf1a* and *orf1ab*. After the AAC in the aminoacyl (A) site of the ribosome recognizes its cognate transfer RNA, p.Asn4401 is added to the nascent peptide, and the ribosome prepares to move p.Asn4401 to the peptidyl (P) site of the ribosome and read the next codon at the A site. Normally, this process occurs canonically (above), with the ribosome moving +3 nucleotides on the mRNA (GGG), thus leading to addition of p.Gly4402 and normal translation of *orf1*. However, ~10% of the time in SARS-CoV reporter constructs[21], the ribosome slips, owing to the highly structured frameshifting element (depicted as a simple cartoon stem-loop) with the bound tRNAs slipping –1 nucleotide and causing C13468 to remain in the A site of the ribosome. This results in +2 movement along the mRNA and an overall –1 frameshift. Therefore, the next codon read is CGG, and p.Arg4402 is added. Because the *orf1a* stop codon at position 4406 is no longer in frame, ~2,700 additional amino acids encoding nsp12–16 are added to the polyprotein.

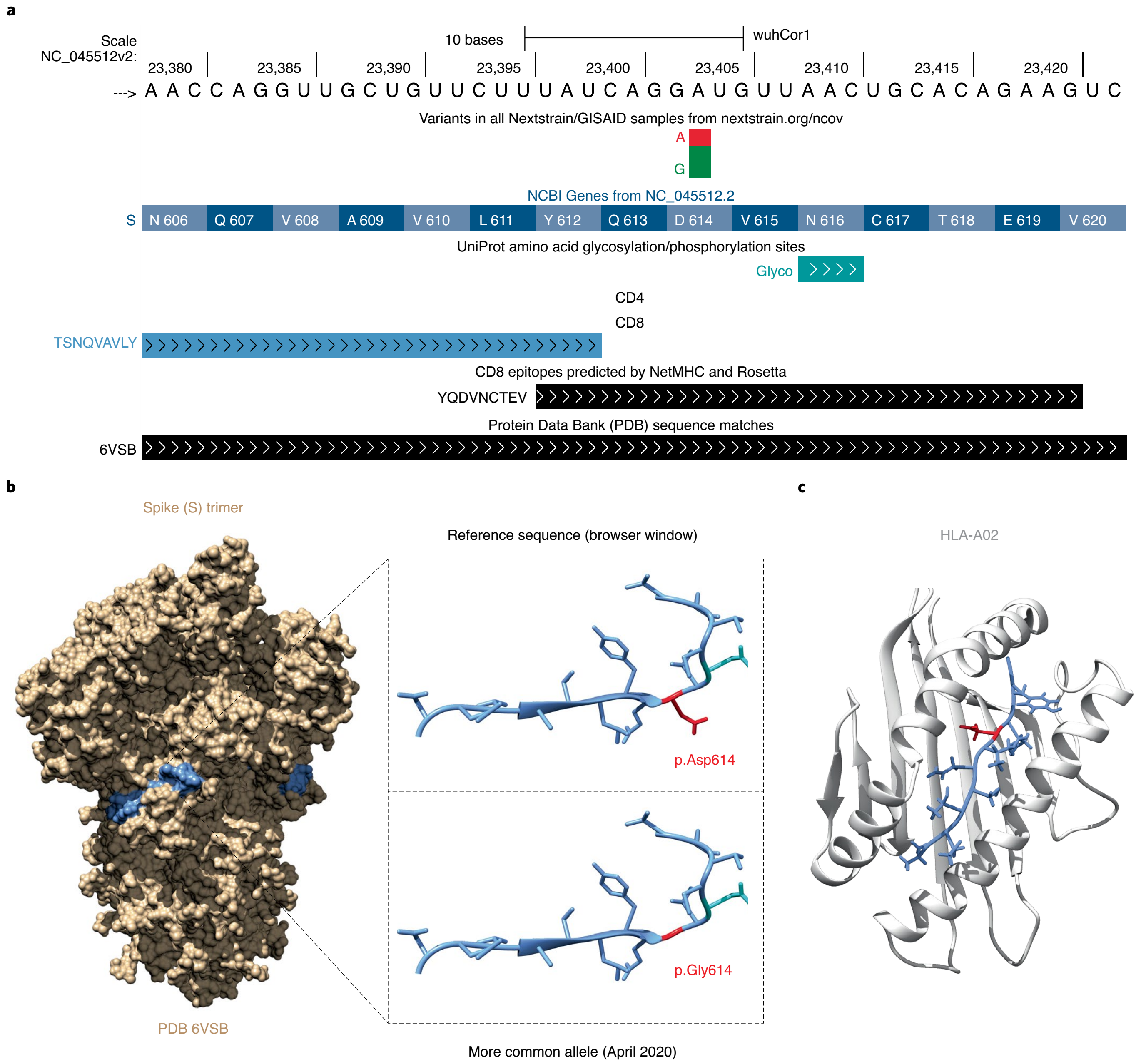
a
Scale
NC_045512v2:
10 bases
wuhCor1
23,380
23,385
23,390
23,395
23,400
23,405
23,410
23,415
23,420
---> A A C C A G G U U G C U G U U C U U U A U C A G G A U G U U A A C U G C A C A G A A G U C
Variants in all Nextstrain/GISAID samples from nextstrain.org/ncov
A
G
NCBI Genes from NC_045512.2
S
N 606
Q 607
V 608
A 609
V 610
L 611
Y 612
Q 613
D 614
V 615
N 616
C 617
T 618
E 619
V 620
UniProt amino acid glycosylation/phosphorylation sites
Glyco
CD4
CD8
TSNQVAVLY
CD8 epitopes predicted by NetMHC and Rosetta
YQDVNCTEV
Protein Data Bank (PDB) sequence matches
6VSB
b
Spike (S) trimer
Reference sequence (browser window)
p.Asp614
p.Gly614
PDB 6VSB
More common allele (April 2020)
c
HLA-A02

Fig. 4 | Combining data tracks to generate hypotheses. **a**, Browser view of a region of the viral genome encoding part of the S protein. The variants track shows an A>G mutation that causes the amino acid change p.Asp614Gly, which is now found more commonly than the reference nucleotide from the original Wuhan outbreak[41,42]. Additional tracks display peptides within the virus that are predicted to be immunogenic. The p.Asp614Gly mutation is contained within a predicted immunogenic peptide. In addition, an annotated glycosylation site at amino acid 616 (highlighted in aqua), which can affect epitope recognition, is shown (live interactive session: http://genome.ucsc.edu/s/SARS_CoV2/Figure4a). **b**, Structure of the S protein (PDB 6VSB) trimer. The amino acid sequence viewed in **a** is highlighted in blue. Inset, close-up view of the blue region with amino acid side chains. Highlighted in red are p.Asp614 (top; the product of the allele present in the original reference genome) and p.Gly614 (bottom) substituted in the structure with UCSF Chimera[43]. **c**, Structure of the immunogenic peptide YQDVNCTEV in complex with HLA-A*02:01. Residue p.Asp614 (red) is nestled within the binding groove, thus leading to a hypothesis that the p.Gly614 alteration may alter binding. Although browser-based comparisons of these data provide insight into possible models for the increased frequency of p.Gly614, further evolutionary and experimental analyses are required to make definitive statements about the functional consequences of this mutation. Recent work has suggested that p.Asp614Gly may increase the transmissibility of SARS-CoV-2 but has little effect on neutralization by antibodies generated from natural infection[44].

(https://sites.google.com/view/opencovid19), nanopore sequencing primers from the ARTIC Network (https://artic.network) and high-scoring and validated CRISPR–Cas13 guides[25,26]. The crowd-sourced annotations track (described below) contains CRISPR guides used in SARS-CoV-2 detection via Cas12 (ref. [27]) and Cas13 (refs. [26,28]), as well as loop-mediated isothermal amplification (LAMP) primers[29]. These tracks can be used in combination with the variants tracks to determine whether specific primers or detection methods might be less effective in detecting certain viral clades (Supplementary Fig. 2).

Genes and gene predictions. These tracks contain information centered around the genes in the viral genome, as illustrated in Fig. 1. For instance, the NCBI Genes track contains annotations of viral gene models from NCBI. Because individual viral genes often have many names (for example, *nsp12*, *RdRp* or *Pol*), many of these tracks list synonyms or notes in additional fields (viewable by clicking an annotation) so that researchers can compare the annotation nomenclature. Additional tracks contain information such as interactions between viral proteins and human proteins from affinity-purification and mass-spectrometry experiments[22] (protein interact), PDB structures and Rfam and other predicted RNA structure annotations[30] (Rangan RNA).

UniProt protein annotations. Protein annotations from SwissProt/UniProt[31] are an essential complement to the NCBI RefSeq gene annotations[32]. These tracks display a variety of protein annotation data including highlights, special regions highlighted by SwissProt curators (for example, the region of the S protein that binds the human receptor protein ACE2); mature products, the mature proteins that result from polypeptide cleavage (Fig. 2c); protein domains; and glycosyl/phosphoryl sites of post-translational modification.

Immunology. These tracks contain SARS-CoV-2 protein epitopes reported in the literature. Included are epitopes that have been predicted and/or validated to be immunogenic. These data can be overlaid with structural and variation information to track mutations that overlap with potential therapeutic targets (Fig. 4). The data feature both linear epitopes recognized by B-cell receptors or antibodies as well as information on conformational epitopes[33,34] (recorded in the crowd-sourced-annotations track; data not shown). These tracks (IEDB predictions, Poran HLA I and Poran HLA II) also display epitopes recognized by CD8$^+$ or CD4$^+$ T cells when presented by human leukocyte antigen (HLA) molecules on host cells[35,36]. When possible, the latter are organized according to the HLA allele of the host used in their presentation. For the track CD8 RosettaMHC, interactive 3D models from Rosetta are available through clicking on the annotated epitope[37]. We will update this track group as validation and identification of epitopes continues.

Comparative genomics. This group contains three tracks that show multiple alignments built from sequences provided by NCBI/INSDC: (1) 7 human CoVs, an alignment of the seven coronavirus sequences that infect humans, (2) 119 vertebrate CoVs, an alignment of 119 sequences, most of which are human coronaviruses (though not necessarily SARS-CoV-2), with various animal viruses and (3) 44 Bat CoVs, an alignment of 44 various bat coronaviruses most closely related to human SARS CoV-2. The mutations in the alignments are colored according to their effects on the protein (white, no difference; red, nonsynonymous; green, synonymous; blue, noncoding; yellow, missing data due to unalignable, absent or unknown sequences). Analysis of evolutionary rates derived from comparative genomics, including insertions and deletions, can pinpoint functionally interesting sections of the viral genome. For instance, comparison of coronaviruses across species and across host genomes can clearly illustrate evolutionary features such as accelerated evolution at receptor spike-binding regions[38,39] (Supplementary Fig. 3).

Variant and repeats. These tracks contain information on the variation and evolutionary patterns observed in SARS-CoV-2 sequences in samples from around the world. NextstrainVars (Supplementary Fig. 4) contains the time-stamped molecular phylogenetic tree produced by the Nextstrain team[6], on the basis of complete and quality-controlled viral genomes from Global Initiative on Sharing All Influenza Data (GISAID)[40]. The tree is shown in pack and squish views. Dense and full modes display the frequencies at which these variants are found. In the tree, samples are sorted according to their order in the JSON file produced by Nextstrain describing the phylogenetic tree, and they appear in different colors according to clades identified by Nextstrain. Tools are provided to filter these data to show only well-supported mutation calls, set thresholds for minor-allele frequency and display data for specific clades.

Crowd-sourced data. To encourage shared community annotation of the viral genome without requiring submitters to have detailed knowledge of genomic data type formats, we have created a crowd-sourced-annotations track. Its documentation includes a link (http://bit.ly/cov2annots) to a spreadsheet in which anyone can enter the start and end positions of some annotations, descriptive text, and links to websites or articles containing the source information. Every night, all new user annotations from the previous day undergo a brief manual check and are added to the public crowd-sourced-annotations track.

Custom tracks. Users can add annotations for their own use or to share with other groups by clicking the ‘custom tracks’ button. Annotation data can be pasted directly into a text box or uploaded in a standard genomic file format. To share a custom track, users can create a stable link via the menu My Data → My Sessions. Session links can be shared multiple times and will always load the data exactly as originally saved.

Discussion

The rapid pace of SARS-CoV-2 research is generating a wealth of molecular and genomic data across a variety of databases. The UCSC Genome Browser is an established and highly accessed web-based viewer and standardized repository of genomic data with extensive functionality and a 20-year track record of serving the scientific community.

The browser, together with its underlying data organization, is a familiar environment for hundreds of thousands of biomedical researchers who study the human genome, using it to explore and download standardized data formats from a single source for custom analyses. Here, we hope to introduce these tools to virologists, epidemiologists, vaccinologists, antiviral-therapy developers and those seeking to repurpose existing biomedical resources and therapies to combat the virus and its pathological effects. To leverage the expertise of scientists who lack familiarity with genomic analyses but possess expertise in other areas, we have developed a crowd-sourced-data track in which users can simply enter the coordinate and name of a feature. This annotation is then displayed and shared with others on the browser, where it will be viewed and compared with the many existing data tracks. Together, these features make the SARS-CoV-2 browser a simple but powerful tool for researchers to track developments in SARS-CoV-2 science, to detail the changes in the viral genome over the course of the pandemic, and to develop testable hypotheses and novel strategies to combat it.

All SARS-CoV-2 data shown at https://genome.ucsc.edu can be downloaded as customizable spreadsheet tables via our tool UCSC Table Browser at https://genome.ucsc.edu/cgi-bin/hgTables or can be downloaded as raw data from our data downloads server, https://hgdownload.soe.ucsc.edu. In accordance with the GISAID license, we cannot allow downloading of any mutation data derived from these databases, but the raw sequences can be downloaded from https://gisaid.org after registration. Researchers who face problems downloading data from our website are invited to contact the help desk at genome@soe.ucsc.edu.

As scientists continue to generate SARS-CoV-2 data, we will continue to rapidly process, display and share these data in the SARS-CoV-2 browser. We urge authors to contact us at genome- www@soe.ucsc.edu for help in properly citing, annotating and displaying their data in a clear, accurate and intuitive manner in the browser so that it can reach the widest possible audience of researchers. Through this type of open collaboration, we believe the SARS-CoV-2 browser will facilitate the analysis and display of the collective molecular information needed to defeat the virus.

Jason D. Fernandes[1,2], Angie S. Hinrichs[1], Hiram Clawson[1], Jairo Navarro Gonzalez[1], Brian T. Lee[1], Luis R. Nassar[1], Brian J. Raney[1], Kate R. Rosenbloom[1], Santrupti Nerli[1], Arjun A. Rao[3], Daniel Schmelter[1], Alastair Fyfe[1], Nathan Maulding[1], Ann S. Zweig[1], Todd M. Lowe[1,5], Manuel Ares Jr[4,5], Russ Corbet-Detig[1], W. James Kent[1], David Haussler[1,2,5] ✉ and Maximilian Haeussler[1] ✉

[1]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [2]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [3]ImmunoX Initiative, University of California San Francisco, San Francisco, CA, USA. [4]Molecular, Cell and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA, USA. [5]Center for Molecular Biology of RNA, University of California Santa Cruz, Santa Cruz, CA, USA.

✉*e-mail: maxh@ucsc.edu; haussler@ucsc.edu*

Published online: 9 September 2020
https://doi.org/10.1038/s41588-020-0700-8

References

1. Kent, W. J. et al. *Genome Res.* **12**, 996–1006 (2002).
2. Flynn, J.A. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2020.02.07.939124 (2020).
3. Buels, R. et al. *Genome Biol.* **17**, 66 (2016).
4. Stalker, J. et al. *Genome Res.* **14**, 951–955 (2004).
5. Wu, F. et al. *Nature* **579**, 265–269 (2020).
6. Hadfield, J. et al. *Bioinformatics* **34**, 4121–4123 (2018).
7. Rambaut, A. et al. *Nat. Microbiol.* https://doi.org/10.1038/s41564-020-0770-5 (2020).
8. Zhao, W.-M. et al. *Yi Chuan* **42**, 212–221 (2020).
9. Raney, B. J. et al. *Bioinformatics* **30**, 1003–1005 (2014).
10. Chen, Y., Liu, Q. & Guo, D. *J. Med. Virol.* **92**, 418–423 (2020).
11. Nakagawa, K., Lokugamage, K. G. & Makino, S. *Adv. Virus Res.* **96**, 165–192 (2016).
12. Sola, I., Almazán, F., Zúñiga, S. & Enjuanes, L. *Annu. Rev. Virol.* **2**, 265–288 (2015).
13. Yount, B., Roberts, R. S., Lindesmith, L. & Baric, R. S. *Proc. Natl Acad. Sci. USA* **103**, 12546–12551 (2006).
14. Kim, D. et al. *Cell* **181**, 914–921.e10 (2020).
15. Nomburg, J., Meyerson, M. & DeCaprio, J.A. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.28.066951 (2020).
16. Brian, D. A. & Baric, R. S. *Curr. Top. Microbiol. Immunol.* **287**, 1–30 (2005).
17. Fehr, A. R. & Perlman, S. *Methods Mol. Biol.* **1282**, 1–23 (2015).
18. Barretto, N. et al. *J. Virol.* **79**, 15189–15198 (2005).
19. Zhang, L. et al. *Science* **368**, 409–412 (2020).
20. Bekaert, M. & Rousset, J.-P. *Mol. Cell* **17**, 61–68 (2005).
21. Plant, E. P. & Dinman, J. D. *RNA* **12**, 666–673 (2006).
22. Gordon, D. E. et al. *Nature* **583**, 459–468 (2020).
23. Schaecher, S.R. & Pekosz, A. in *Molecular Biology of the SARS-Coronavirus* (ed. Lal, S. K.) 153–166 (Springer, 2010).
24. Davidson, A. D. et al. *Genome Med.* **12**, 68 (2020).
25. Abbott, T. R. et al. *Cell* **181**, 865–876.e12 (2020).
26. Wessels, H.-H. et al. *Nat. Biotechnol.* **38**, 722–727 (2020).
27. Broughton, J. P. et al. *Nat. Biotechnol.* **38**, 870–874 (2020).
28. Metsky, H.C., Freije, C.A., Kosoko-Thoroddsen, T.-S.F., Sabeti, P.C. & Myhrvold, C. Preprint at *bioRxiv* https://doi.org/10.1101/2020.02.26.967026 (2020).
29. Park, G.-S. et al. *J. Mol. Diagn.* **22**, 729–735 (2020).
30. Rangan, R., Watkins, A. M., Kladwang, W. & Das, R. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.14.041962 (2020).
31. UniProt Consortium. *Nucleic Acids Res.* **47**, D506–D515 (2019).
32. Pruitt, K. D., Tatusova, T. & Maglott, D. R. *Nucleic Acids Res.* **35**, D61–D65 (2007).
33. Pinto, D. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.07.023903 (2020).
34. Yuan, M. et al. *Science* **368**, 630–633 (2020).
35. Grifoni, A. et al. *Cell Host Microbe* **27**, 671–680.e2 (2020).
36. Poran, A. et al. *Genome Med.* **12**, 70 (2020).
37. Nerli, S. & Sgourakis, N.G. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.23.004176 (2020).
38. Demogines, A., Farzan, M. & Sawyer, S. L. *J. Virol.* **86**, 6350–6353 (2012).
39. Damas, J. et al. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.2010146117 (2020).
40. Shu, Y. & McCauley, J. *Euro Surveill.* **22**, 30494 (2017).
41. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. *Nat. Med.* **26**, 450–452 (2020).
42. Korber, B. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.29.069054 (2020).
43. Pettersen, E. F. et al. *J. Comput. Chem.* **25**, 1605–1612 (2004).
44. Grubaugh, N. D., Hanage, W. P. & Rasmussen, A. L. *Cell* **182**, 794–795 (2020).
45. Kent, W. J. *Genome Res.* **12**, 656–664 (2002).

Acknowledgements

We thank K. Kober and K. Charbonneau from the Kober laboratory at UCSF for their contributions to the crowd-sourced data track, as well as H. Beale, J. Sim, A. Gonzalez Armenta, P. Berman, N. Sgourakis, E. Jung, P. Angulo and the rest of the scientists at UCSC and around the world for making these tracks of molecular information possible.

The UCSC Human Genome Browser software, quality control and training are funded by the NIH National Human Genome Research Institute, currently with grant 5U41HG002371. The SARS-CoV-2 genome browser and data annotation tracks are funded by generous individual donors including Pat & Rowland Rebele, Eric and Wendy Schmidt by recommendation of the Schmidt Futures program, the Center for Information Technology Research in the Interest of Society (CITRIS, 2020-0000000020) and University of California Office of the President (UCOP) Emergency COVID-19 Research Seed Funding Grant R00RG2456. Funding for open-access charges was provided by the National Human Genome Research Institute (5U41HG002371). R.B.C. was supported by NIGMS/R35GM128932.

Author contributions

A.S.H., H.C., J.N.G., B.T.L., L.R.N., B.J.R., K.R.R., D.S., A.S.Z., W.J.K. and M.H. built the SARS-CoV-2 Browser. J.D.F., A.S.H., S.N., A.A.R., B.J.R., M.H., T.M.L. N.M., A.F. and M.A. developed tracks for the browser. J.D.F., T.M.L., M.A., R.C.-D., W.J.K., D.H. and M.H. provided general guidance on aspects of virology, RNA biology and genomics. J.D.F., D.H. and M.H. wrote the manuscript.

Competing interests

A.S.H., H.C., J.N.G., B.T.L., L.R.N., B.J.R., K.R.R., D.S., A.S.Z., W.J.K., D.H. and M.H. receive royalties from the sale of UCSC Genome Browser source code, LiftOver, GBiB and GBiC licenses to commercial entities. W.J.K. owns Kent Informatics.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0700-8.