Exploring the coronavirus pandemic with the WashU Virus Genome Browser

The WashU Virus Genome Browser is a web-based portal for efficient visualization of viral 'omics' data in the context of a variety of annotation tracks and host infection responses. The browser features both a phylogenetic-tree-based view and a genomic-coordinate, track-based view in which users can analyze the sequence features of viral genomes, sequence diversity among viral strains, genomic sites of diagnostic tests, predicted immunogenic epitopes and a continuously updated repository of publicly available genomic datasets.

Jennifer A. Flynn, Deepak Purushotham, Mayank N. K. Choudhary, Xiaoyu Zhuo, Changxu Fan, Gavriel Matt, Daofeng Li and Ting Wang

Coronavirus disease 2019 (COVID-19) is a rapidly spreading viral disease that has become a global health crisis. The first case of COVID-19 was reported on 12 December 2019, in Wuhan, China; by 2 August 2020, the disease had spread to more than 215 countries, territories and areas, resulting in at least 17,660,523 cases and 680,894 deaths (<https://covid19.who.int/>). COVID-19 is caused by a virus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is a zoonotic, enveloped virus containing a positive single-stranded RNA genome 29,903 base pairs in size. The virus is one of seven coronaviruses known to infect humans, along with severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV)^{1,2}. To better understand the evolutionary history and pathogenesis of this new virus, large amounts of genomic data have been generated for both SARS-CoV-2 and human host cells. The genomes of thousands of SARS-CoV-2 strains have now been fully sequenced, the SARS-CoV-2 genome and transcriptome have been functionally annotated^{3,4}, and the genomic activity of host cells in response to viral infection is beginning to be elucidated^{5,6}. This explosion of omics data has created a need for platforms to store, process, analyze and visualize these data, to gain insights into the genomic basis of viral infection. Sequencing databases, such as National Center for Biotechnology Information (NCBI) GenBank⁷ and Global Initiative on Sharing All Influenza Data (GISAID)⁸, currently store most of the known genome sequences of individual SARS-CoV-2 strains. The pathogen genomics platform Nextstrain currently analyzes the genomic diversity and epidemiology of a subset of

Fig. 1 | Evaluating PCR primers for accumulating mutations with the WashU Virus Genome Browser.

a, Genome-browser view of the entire SARS-CoV-2 genome with the following tracks loaded (from top to bottom: gene annotations, ruler, sequence diversity, mutation alert and China CDC primer-binding sites). The sequence-diversity and mutation-alert tracks are two different representations of sequence diversity among SARS-CoV-2 strains: the former shows the Shannon entropy at each nucleotide position, and the latter shows the number of strains with a mutation at each nucleotide position. Primer-binding sites (red bars) within the *N* gene appear to overlap a region with a moderate level of sequence diversity among SARS-CoV-2 strains. The browser's magnification tool (circled in red) can be used to quickly zoom in to the region of interest. **b**, A zoomed-in view of the region described in **a**, showing that the binding site of the forward primer ('F-N') is indeed highly variable among SARS-CoV-2 strains. SNV tracks of SARS-CoV-2 strains show three strains with mutations and three strains lacking mutations in the primer-binding site. Clicking on mutations in the mutation-alert track shows the total number of strains mutated at the selected position and the accession numbers of these strains (pop-up display window).

these strains and provides a useful overview of the phylogeography of SARS-CoV-2 transmission⁹. Although these databases

have amassed a rich source of genomic and phylogenetic information for SARS-CoV-2, comprehensive analysis of the SARS-CoV-2

genome will require the use of a high-performance genome browser designed for storing and visualizing various viral and host omic datasets.

To address this need, we created the WashU Virus Genome Browser (<https://virusgateway.wustl.edu>), a web-based portal adapted from the WashU Epigenome Browser^{10–13}, which is specifically designed for efficient visualization of viral genomic sequencing data (Supplementary Note). The browser contains the genomes of four different pathogenic virus species: SARS-CoV-2, SARS-CoV, MERS-CoV and Ebola. A reference-genome sequence is provided for each species, as well as a comprehensive collection of genome sequences of individual strains that have been isolated from patients from different geographical regions and time periods. In total, we have collected genomes of >80,000 SARS-CoV-2 strains, 332 SARS-CoV strains, 551 MERS-CoV strains and 1,574 Ebola strains, and we are continuously updating our database (Supplementary Table 1). The genome of each strain is automatically aligned in pairwise fashion to the reference genome, and sequence variants relative to the reference are visualized through single-nucleotide variant (SNV) tracks, which provide a simple and effective way to visualize sequence variation across multiple viral strains (Figs. 1b, 2 and 3b

Fig. 2 | Using the WashU Virus Genome Browser to discover conserved immune-epitopes.

a, Browser view of the SARS-CoV-2 gene encoding the nucleocapsid (N) protein with SNV tracks of two SARS-CoV strains (DQ071615.1 and AY278488.2) and five SARS-CoV-2 strains (MN938384.1, MN975262.1, MN985325.1, MN988668.1 and MN988669.1) loaded in the view. The track of putative SARS-CoV immune epitopes is also displayed in density mode. **b**, A zoomed-in view of the orange box in **a**, displaying the first 9 amino acids of the SARS-CoV-2 N protein. SNVs at positions 28296 (T>C) and 28299 (G>A) are silent mutations; however, the 'TCA' insertion at position 28294 in SARS-CoV accession AY278488.2 (BJ01) results in the insertion of a serine residue in the SARS-CoV N protein relative to the SARS-CoV-2 N protein. Of note, this insertion is not present in the other SARS-CoV accession (DQ071615.1). Owing to the variability in the amino acid sequence within this region across SARS-CoV and SARS-CoV-2 strains, this region is unlikely to be a good candidate for epitope design. **c**, A zoomed-in view of the purple box in **a**, displaying a region that is likely to be a good candidate for epitope design because it is fully conserved across SARS-CoV and SARS-CoV-2 strains, and it also encodes several putative antigenic epitopes.

a

Fig. 3 | Using the WashU Virus Genome Browser to study the prevalence of the S-protein p.Asp614Gly alteration. **a**, Default browser view of the SARS-CoV-2 genome. The SARS-CoV-2 genome browser includes various genomic datasets in track format (from top to bottom: gene annotations, sequence diversity, mutation alerts, gene expression, predicted immune epitopes, recombination events and RNA modifications). The p.Asp614Gly alteration within the S protein is circled in red within the sequence-diversity track. **b**, A zoomed-in view of the p.Asp614Gly alteration circled in **a**. The ruler track displays a color-coded nucleotide sequence. The sequence-diversity and mutation-alert tracks reveal a high degree of variation across SARS-CoV-2 strains at this location. Sequence variation within individual strains is displayed below in the form of SNV tracks, which report variations relative to the reference genome in a color-coded format. Notably, mutations at this position appear to be enriched in non-Asian accessions. Beneath these tracks is a dynamic track showing the weekly percentage of strains with the p.Asp614Gly alteration (beginning on 24 December 2019). Finally, the bottom track displays predicted antigenic linear epitopes that were experimentally identified in SARS-CoV by using T-cell, B-cell and major-histocompatibility-complex ligand assays¹⁶. **c**, Plot showing the percentage of SARS-CoV-2 isolates in each country that contain the p.Asp614Gly alteration (data retrieved from GISAID, accessed on 19 August 2020)⁸. Countries are ordered by continent and then by date of virus introduction. Countries with fewer than ten cases are excluded. Sample sizes are displayed in parentheses beneath each country. **d**, Plot showing the percentage of SARS-CoV-2 isolates per week (beginning on 24 December 2019) that contain the p.Asp614Gly alteration. Sample sizes for each week are displayed in parentheses.

and Supplementary Note). Users can search for viral strains in the browser by using the data table (Supplementary Fig. 1 and Supplementary Note). Here, strains can be filtered by several metadata features, including country, continent, data source, collection date, tree-view availability and clade. Additionally, the data table's search feature allows users to locate strains of interest by querying individual accession IDs or mutations. Strains of interest can be added to the user's cart and displayed in the form of SNV tracks in the track-based browser view or highlighted in the tree-based view (Supplementary Note). To accommodate users who wish to visualize sequence variation within their own viral strains, the browser supports user upload of viral sequences in SNV format. Users are provided with easy-to-use scripts to convert their alignment results into SNV format for visualization (Supplementary Note). Additionally, the browser offers various precomputed genome-annotation tracks, such as gene annotations, protein annotations, recombination sites, sequence diversity, mutation frequencies, comparative tracks between species and GC density. Along with the traditional track-based view, our genome browser also features a phylogenetic-tree-based view, which allows users to analyze the evolutionary relationships and history of viral strains (Supplementary Fig. 2 and Supplementary Note). In this view, users can annotate strains on the phylogenetic tree with different metadata categories (for example, country of origin, collection date or Nextstrain clade designations) and can also highlight strains of interest within the tree to determine their relative relationships to other strains or clades (Supplementary Note).

The WashU Virus Genome Browser also hosts a large set of publicly available genomic datasets for both SARS-CoV-2 and human host cells. These datasets are organized as data hubs consisting of one or more tracks of data related to a specific

aspect of viral biology, host biology, disease diagnosis or disease therapeutics. The browser currently hosts data hubs for viral transcription, viral recombination sites, viral RNA modifications, host transcriptional responses to infection, predicted antigenic epitopes, binding sites of diagnostic primers and genomic targets of CRISPR-based diagnostic tests (Supplementary Note and Supplementary Table 1). These data hubs can be easily loaded into the genome browser, providing users with an efficient way to analyze multiple data tracks of interest in tandem. As new omics data become publicly available, these datasets are promptly uploaded to the browser in data hubs. Emerging evidence now links host genomic variation with disease severity¹⁴, and further research efforts are focused on understanding the host response to viral infection, thus necessitating a platform for visualizing cross-species genomic data. The compatibility between the WashU Virus Genome Browser and the WashU Epigenome Browser provides a platform for visualizing both viral and host genomic data. Our browser provides the unique capability of seamlessly visualizing viral genomics data and the corresponding host genomic response data, also in a modular data-hub format (Supplementary Fig. 3).

As the SARS-CoV-2 virus continues to evolve, one major task is studying how mutations accumulate within diagnostic PCR-primer-binding sites, which could potentially decrease test efficacy. To efficiently track mutation hotspots within a viral species, we developed two annotation tracks: a mutation-alert track, which displays the number of strains with a mutation at each genomic position, and a sequence-diversity (Shannon entropy) track, which displays the Shannon entropy (variation) at each genomic position across strains (Fig. 1a). We also provide a data hub of the genomic binding sites of diagnostic PCR primers from the Centers for Disease Control and Prevention (CDC) and World

Health Organization, which users can load into the genome-browser view to determine whether primers overlap with mutation hotspots or mutations in certain strains (Supplementary Note). As shown in Fig. 1a, after loading the track for China-CDC detection primers, users can readily see a mutation hotspot overlapping primer locations within the *N* gene. Zooming into the region of interest and adding SNV tracks for individual strains offers a color-coded display of individual mutations and accession IDs of strains with a mutation at the given position (Fig. 1b). In addition to the preloaded primer-location tracks, the browser supports user upload of novel primer locations in standard bed format¹⁵, thereby allowing users to determine whether their primers overlap with mutations.

The WashU Virus Genome Browser can also be used to identify antigenic epitopes that are conserved across viral species. We demonstrate this utility by performing a similar analysis to that in Extended Data Fig. 5 in ref. ³, showing a genomic alignment of two SARS-CoV strains and five SARS-CoV-2 strains to the reference SARS-CoV-2 *N* gene, along with a track of putative immune epitopes identified in SARS-CoV (Fig. 2). Peptides in SARS-CoV-2 that are homologous to putative antigenic epitopes in SARS-CoV serve as useful targets for SARS-CoV-2 vaccine development. However, the presence of sequence variants within these peptides may compromise vaccine effectiveness (Fig. 2b,c). Motivated by the overall high sequence similarity between the SARS-CoV and SARS-CoV-2 genomes, we analyzed the full list of experimentally identified linear epitopes from the Immune Epitope Database and Analysis Resource (IEDB)¹⁶, and identified a list of 320 high-confidence linear epitopes whose amino acid sequences are identical to those of the predicted translated products from the SARS-CoV-2 reference strain (Supplementary Table 2). This list provides a catalog of epitopes for

Fig. 4 | Using the WashU Virus Genome Browser to study sequence conservation across viral species. Genome-browser view of the 5' end of the SARS-CoV-2 S gene. Tracks loaded in the view are as follows from top to bottom: gene annotations, ruler, transcription regulatory sequence (TRS) sites, pangolin coronavirus (EPI_ISL_410721) comparison tracks (SNV track, genome-alignment track and read-alignment track) and bat coronavirus (EPI_ISL_402131) comparison tracks (SNV track, genome-alignment track and read-alignment track). In the SNV track (displayed in density mode), the y axis represents the density of sequence variation in bat or pangolin coronaviruses compared with the SARS-CoV-2 reference. In the genome-alignment track, the SARS-CoV-2 reference is represented by the solid blue band, and the query sequence (in this case, the bat or pangolin coronavirus genome) is represented by the purple band. Black bars between the reference and the query represent matches, whereas the absence of a bar represents a variant (mismatch or gap). A gap in the reference or query is represented by a gap in the blue or purple band, respectively. In the read-alignment track, data are loaded in bam format. Each read is displayed as a colored bar (blue, plus strand; red, minus strand), and gaps within a bar represent mismatches to the reference. The alignment of a read to the reference can be displayed by selecting a particular read (pop-up display window). This genome-browser view shows that the 5' end of the S gene is highly variable between SARS-CoV-2 and the pangolin coronavirus, but not between SARS-CoV-2 and the bat coronavirus.

researchers testing immunological targets that can potentially elicit T-cell and B-cell responses to SARS-CoV-2.

The browser supports multiple file formats, thus allowing users to visualize sequencing data in a variety of ways to better understand viral sequence evolution. Recent work suggests that the amino acid change from aspartate to glycine at position 614 within the spike (S) protein of SARS-CoV-2 has become dominant in non-Asian countries¹⁷. Although this mutation is not located within the core receptor-binding domain of the protein, it is nonetheless thought to contribute to the transmissibility of the virus¹⁷. In the browser track view, the prevalence of the p.Asp614Gly alteration among SARS-CoV-2 strains is immediately evident when viewing the preloaded sequence-diversity (Shannon entropy) track and the mutation-alert track, which highlight the entropy across strains and the number of strains with mutations,

respectively (Fig. 3a). A finer-scale visualization of the p.Asp614 codon within the browser shows that some accessions of non-Asian origin contain the p.Asp614Gly alteration, and a dynamic track of the mutation's weekly prevalence shows that more than 90% of strains isolated during the week of 27 April 2020 contained the p.Asp614Gly alteration (Fig. 3b). Further characterization of the prevalence confirmed the higher frequency of the p.Asp614Gly alteration in non-Asian accessions (Fig. 3c) and the increase in prevalence of the p.Asp614Gly alteration over time (Fig. 3d).

Using a similar approach, we were also able to confirm the recent observation of a heightened mutation rate at the amino terminus of the S protein among betacoronaviruses³. The genomes of pangolin and bat coronaviruses were aligned to the SARS-CoV-2 reference genome (Supplementary Note), and

sequence variation was displayed in the browser in the form of SNV tracks and genome-comparison tracks (Fig. 4 and Supplementary Note). This view showed that the 5' end of the S gene is highly divergent between the pangolin coronavirus and SARS-CoV-2 (Fig. 4). This finding was further supported by displaying raw next-generation sequence reads in bam format in the browser. Coronavirus reads from a pangolin viral metagenomic dataset¹⁸ were extracted and aligned to the SARS-CoV-2 reference genome (Supplementary Note). Most pangolin reads (1,468 of 2,288) aligned to the SARS-CoV-2 genome, and the average mismatch rate between pangolin-CoV and the SARS-CoV-2 sequences was only ~7% (Supplementary Note), thus indicating that the pangolin-CoV dataset is closely related to the SARS-CoV-2 dataset. Nevertheless, we found that no reads from the pangolin library aligned to the 5' end of the S gene of

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 betacoronaviruses³.

All analyzed viral sequences are available from the NCBI GenBank⁷ (<https://www.ncbi.nlm.nih.gov/nuccore/>), GISAID⁸ (<https://www.gisaid.org/>) and Nextstrain⁹ (<https://nextstrain.org/sars-cov-2>) public repositories, with the exception of the pangolin viral metagenomic dataset¹⁸, 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 here¹⁹, 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} ✉

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

- de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
- Cui, J., Li, F. & Shi, Z. L. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
- Zhou, P. et al. *Nature* **579**, 270–273 (2020).
- Kim, D. et al. *Cell* **181**, 914–921.e10 (2020).
- Blanco-Melo, D. et al. *Cell* **181**, 1036–1045.e9 (2020).
- Bojkova, D. et al. *Nature* **583**, 469–472 (2020).
- NCBI Resource Coordinators. *Nucleic Acids Res.* **46**, D8–D13 (2018).
- Shu, Y. & McCauley, J. *Eur. Surveill.* **22**, 30494 (2017).
- Hadfield, J. et al. *Bioinformatics* **34**, 4121–4123 (2018).

- Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. *Nucleic Acids Res.* **47**, W158–W165 (2019).
- Zhou, X. et al. *Nat. Biotechnol.* **33**, 345–346 (2015).
- Zhou, X. et al. *Nat. Methods* **10**, 375–376 (2013).
- Zhou, X. et al. *Nat. Methods* **8**, 989–990 (2011).
- van der Made, C. I. et al. *JAMA* **324**, 1–11 (2020).
- Kent, W. J. et al. *Genome Res.* **12**, 996–1006 (2002).
- Vita, R. et al. *Nucleic Acids Res.* **47**, D339–D343 (2019).
- Korber, B. et al. *Cell* **182**, 812–827.e819 (2020).
- Liu, P., Chen, W. & Chen, J. P. *Viruses* **11**, 979 (2019).
- 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>.

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 Browser¹, 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