Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource

The emergence of the COVID-19 pandemic has spurred a global rush to uncover basic biological mechanisms to inform effective vaccine and drug development. Despite the novelty of the virus, global sequencing efforts have already identified genomic variation across isolates. To enable easy exploration and spatial visualization of the potential implications of SARS-CoV-2 mutations in infection, host immunity and drug development, we have developed COVID-3D (http://biosig.unimelb.edu.au/covid3d/).

Stephanie Portelli, Moshe Olshansky, Carlos H. M. Rodrigues, Elston N. D'Souza, Yoochan Myung, Michael Silk, Azadeh Alavi, Douglas E. V. Pires and David B. Ascher

Declared a global pandemic on 11 March 2020, COVID-19 has become the most recent modern-day global health challenge, infecting 10 million people and claiming more than 500,000 lives within 6 months of being reported to the World Health Organization. Consequently, the scale of its humanitarian and economic impact has driven academic and pharmaceutical efforts to develop vaccines and antiviral treatments. Current efforts include more than 118 active vaccine candidates and numerous additional endeavors to identify biologics and small-molecule treatments.

One further challenge in controlling COVID-19 is the accumulation of variation across genes. Sources indicate that SARS-CoV-2 is mutating at approximately two variants per month, but the potential effects of the accumulation of these variants (Supplementary Fig. 1) on molecular diagnostics and the development of candidate vaccines and treatments remain poorly explored. Fortunately, the continual rapid increase in the amount of SARS-CoV-2 genome sequence data and structural information available provides an opportunity to analyze both data sources concomitantly, thus presenting a unique opportunity to not only understand how variants might affect patient outcomes, but also anticipate and minimize their potential roles in viral escape through early incorporation of this information within the development pipeline.

To facilitate such an understanding, we have developed a comprehensive online resource, COVID-3D, to enable analysis and interpretation of more than 11,000 variants detected in circulating SARS-CoV-2 genomic sequences (Supplementary Fig. 2). We have mapped these circulating variants and their frequencies to the corresponding protein sequences (Supplementary Table 1) and structures of the SARS-CoV-2 proteins derived from available experimental information (Supplementary Table 2), thus permitting direct comparison of variant clustering between the sequence and structural representations, along with the identification of coevolutionary relationships and potential compensatory mutations. Beyond these circulating variants, we have identified mutations from the longer-circulating related viruses BAT RaTG13 and SARS-CoV, to enable further investigation of the mutations that drove the species jump from RaTG13 and that increased the infectivity and mortality beyond those of SARS-CoV. Our interactive three-dimensional viewer enables fast and intuitive spatial visualization of SARS-CoV-2 variants, highlighting their potential effects on protein structure and interactions[1–7] (Supplementary Figs. 3–6). This viewer is particularly useful for analyzing sites that are currently being targeted by potential therapeutics. A built-in mutation-analysis tool allows users to contrast properties and identify patterns in the data, plotting correlations and distributions (Supplementary Fig. 7).

To further enhance therapeutic discovery efforts, we have included maps of the fragment-binding hotspots to capture likely drug-binding sites[8,9], as well as predicted antigenicity maps[10,11] on the structures, which permit rational selection of target sites and compound design, specifically avoiding already circulating variants (Supplementary Fig. 4). Finally, combining this structural information with evolutionary and population variation analysis can further aid in identifying sites that are relatively less likely to accommodate mutations in the future. To facilitate this analysis, COVID-3D also allows users to go from analyzing a protein pocket to virtual screening in several clicks[12]. In an illustrative example, we have used COVID-3D to provide insights into the two main therapeutic targets: the spike protein and main proteinase.

The SARS-CoV-2 spike protein binds human angiotensin-converting enzyme 2 (ACE2), which mediates cell entry. Subsequently, the spike protein's ACE2-receptor-binding domain has been the main target of most vaccine programs. Measures of selective pressure suggest that the spike protein is one of the viral proteins most tolerant to the introduction of mutations[13,14] (Supplementary Table 1). Closer inspection (http://biosig.unimelb.edu.au/covid3d/protein/QHD43416/CLOSED) indicates that although SARS-CoV-2 was discovered only 6 months before the time of analysis, substantial variation can already be seen across the protein surface, including in predicted epitope regions in the receptor-binding domain (Fig. 1). Of these variants, QHD43416 p.Asp614Gly is present in two-thirds of the sequenced strains, although its actual importance remains unclear, despite initial suggestions that it may increase transmissibility[15]. The residue is located far from the ACE2 interface (73 Å) and has been predicted to have a mildly stabilizing effect on protein stability (0.5 kcal mol^{-1} according to DUET[3] and 2.3 kcal mol^{-1} according to SDM[2] analyses) and hence a minimal fitness cost[16]. However, it has been predicted to alter protein dynamics and the interactions between the subunits (4.4 Å from the interface; –0.5 kcal mol^{-1} for

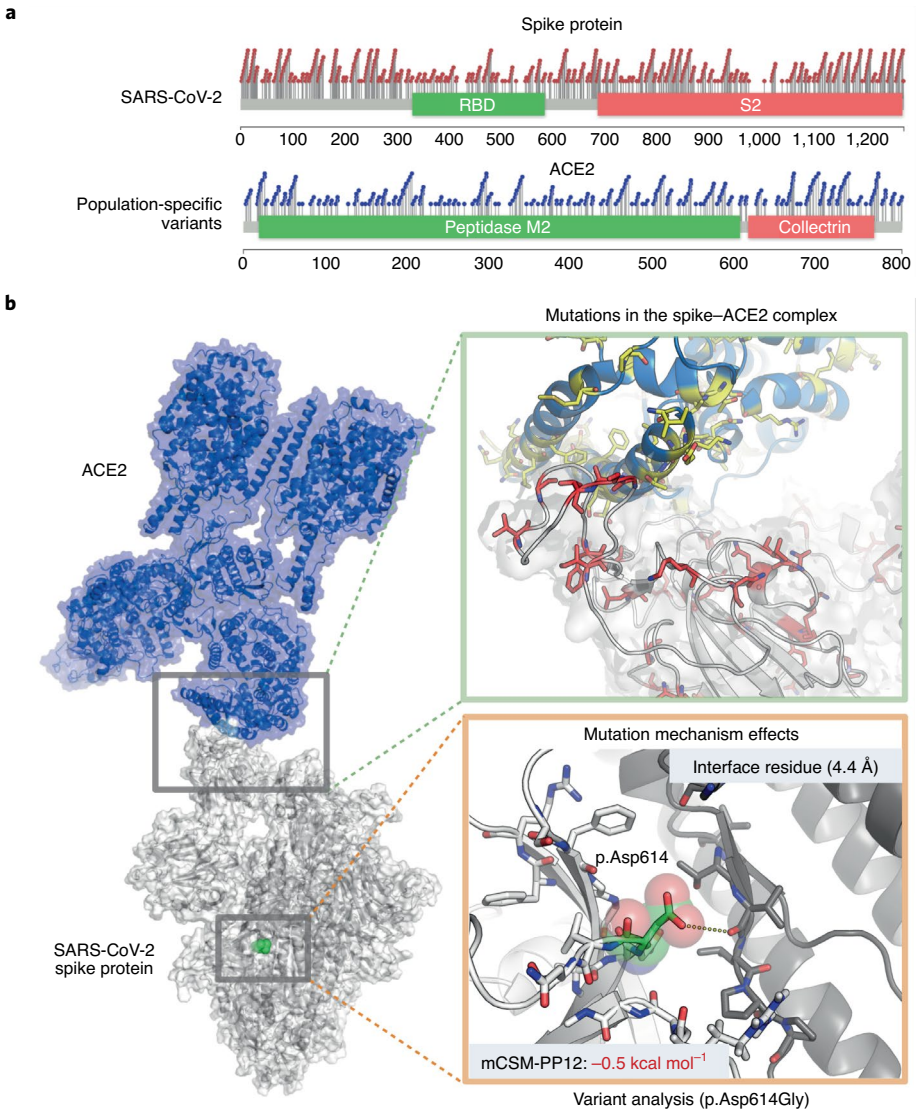

Fig. 1 | Population variation across the spike–ACE2 complex. a, Lollipop plots of circulating missense variants in the SARS-CoV-2 spike protein and population-specific missense variants in human ACE2 illustrate the broad distribution of variants across the proteins. **b**, When visualized spatially, several variants seen at the ACE2–spike interface are predicted to affect the binding affinity. One of the most prevalent circulating SARS-CoV2 spike variants, p.Asp614Gly, is located far from the ACE2 interface but close to the spike-trimer interface and is predicted to lead to structural perturbations.

the closed form versus –0.35 kcal mol^{-1} for the open form, according to mCSM-PPI2 analysis[6]), thus potentially affecting the equilibrium between open and closed states.

Interestingly, when we examined population-specific variants across ACE2, we observed several population-specific variants across the interface recognized by the spike protein (Fig. 1a). Evaluation of the consequences of these variants with mCSM-PPI[6], which has been experimentally validated on this protein system[17], shows potential significant effects on the binding affinity of spike protein, thus paving the way for further work exploring the influence of these variants on COVID-19 severity and progression.

Apart from the spike protein, the main proteinase (http://biosig.unimelb.edu.au/covid3d/protein/QHD43415_5/APO) has also attracted many therapeutic development efforts as a target for the development of small-molecule inhibitors. The main proteinase, however, is not particularly intolerant to missense variants (Supplementary Table 1), thus potentially promoting the emergence of resistant variants. The structures show that several circulating variants already present in the drug-binding site may have effects on efficacy (Fig. 2a). Using COVID-3D, we leveraged the abundance of SARS-CoV-2 genomic sequences to calculate measures of mutational tolerance, and we identified several genes under strong purifying selection (Supplementary Table 1). These include the genes encoding helicase, RNA polymerase, NSP4, NSP9 and ExoN, which may serve as novel, promising drug targets with few circulating variants seen near the druggable pockets (Fig. 2b).

COVID-3D provides an easy-to-use bridge between genomic information and structural insights to better guide biological understanding and treatment efforts. The data and code (http://biosig.unimelb.edu.au/covid3d/code) are freely available via the web interface (http://biosig.unimelb.edu.au/covid3d/). As new structural and sequence data become available, COVID-3D will be periodically updated to enable

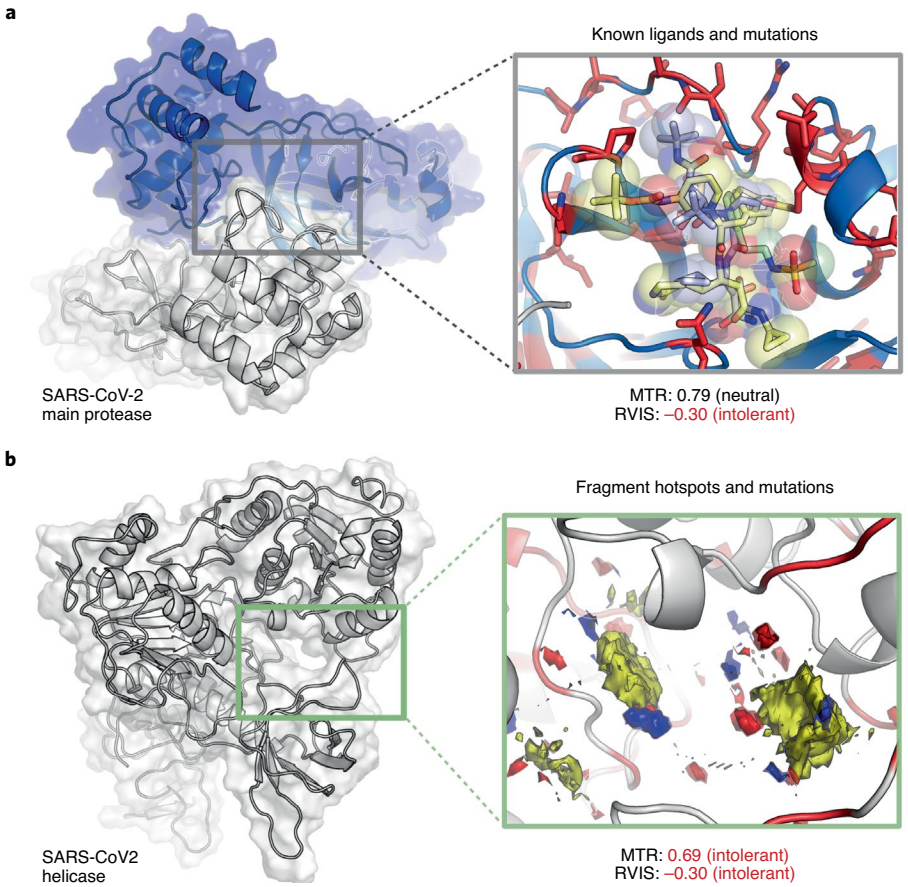

Fig. 2 | Visualization of SARS-CoV-2 circulating variants relative to druggable pockets. a, The gene encoding the main proteinase is neutral to the introduction of missense variants, with an overall missense tolerance score (MTR) and residual variation intolerance score (RVIS) both indicating that the gene is tolerant to genetic variation. Some circulating variants (red sticks) have already been observed to lead to alterations near binding sites of known inhibitors (boceprevir shown in yellow) and are likely to affect drug binding. Therefore, resistance mutations could be selected for with widespread use. **b**, The gene encoding helicase is among the SARS-CoV-2 genes most intolerant to missense variation, with low MTR and RVIS scores. Mapping the fragment-binding hotspots of the protein shows pockets with apolar (yellow), hydrogen-bond-donor (blue) and hydrogen-bond-acceptor (red) potential. Although some variation has been observed near this region, optimization of interactions to avoid these sites could decrease the potential for future resistance.

their integration into ongoing efforts to understand and combat SARS-CoV-2. ❐

Stephanie Portelli[1,2,5], Moshe Olshansky[1,2,5], Carlos H. M. Rodrigues[1,2,5], Elston N. D'Souza[1,2,5], Yoochan Myung[1,2], Michael Silk[1,2], Azadeh Alavi[1,2], Douglas E.V. Pires[1,2,3,5] and David B. Ascher[1,2,4]✉

[1]*Structural Biology and Bioinformatics, Department of Biochemistry, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia.* [2]*Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia.* [3]*School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, Australia.* [4]*Department of Biochemistry, University of Cambridge, Cambridge, UK.* [5]*These authors contributed equally: Stephanie Portelli, Moshe Olshansky, Carlos H. M. Rodrigues, Elston N. D'Souza, Douglas E. V. Pires.*
✉*e-mail: david.ascher@unimelb.edu.au*

Published online: 9 September 2020
https://doi.org/10.1038/s41588-020-0693-3

References

1. Jubb, H. C. et al. *J. Mol. Biol.* **429**, 365–371 (2017).
2. Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B. & Blundell, T. L. *Nucleic Acids Res.* **45**, W229–W235 (2017).
3. Pires, D. E., Ascher, D. B. & Blundell, T. L. *Nucleic Acids Res.* **42**, W31–W319 (2014).
4. Pires, D. E., Ascher, D. B. & Blundell, T. L. *Bioinformatics* **30**, 335–342 (2014).
5. Rodrigues, C. H., Pires, D. E. & Ascher, D. B. *Nucleic Acids Res.* **46**, W350–W355 (2018).
6. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. *Nucleic Acids Res.* **47**, W338–W344 (2019).
7. Pires, D. E., Blundell, T. L. & Ascher, D. B. *Sci. Rep.* **6**, 29575 (2016).
8. Radoux, C. J., Olsson, T. S. G., Pitt, W. R., Groom, C. R. & Blundell, T. L. *J. Med. Chem.* **59**, 4314–4325 (2016).
9. Kawabata, T. *Proteins* **78**, 1195–1211 (2010).
10. Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. *Nucleic Acids Res.* **45**, W24–W29 (2017).
11. Ponomarenko, J. et al. *BMC Bioinforma.* **9**, 514 (2008).
12. Pires, D. E. V. et al. *Bioinformatics* **36**, 4200–4202 (2020).
13. Silk, M., Petrovski, S. & Ascher, D. B. *Nucleic Acids Res.* **47**, W121–W126 (2019).
14. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. *PLoS Genet.* **9**, e1003709 (2013).
15. Korber, B. et al. *Cell* **182**, 812–827.e19 (2020).
16. Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G. & Furnham, N. *Sci. Rep.* **8**, 15356 (2018).
17. MacGowan, S. A. & Barton, G. J. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.03.074781 (2020).

Acknowledgements

We thank N. Thanh Binh (Bioinformatics Institute, A*STAR, Singapore) for help with molecular dynamics simulations. S.P., C.H.M.R. and Y.M. were supported by a Melbourne Research Scholarship. D.B.A. and D.E.V.P. were funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG; MR/M026302/1) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). D.B.A. was funded by the Jack Brockhoff Foundation (JBF 4186, 2016); the Wellcome Trust (200814/Z/16/Z) and an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405). This work was supported in part by the Victorian Government's OIS Program. This research has been conducted using the UK Biobank Resource under Application Number 50000.

Author contributions

S.P. was responsible for structure curation, homology modeling and structural characterization. M.O. was responsible for curating SARS-CoV-2 variants. C.H.M.R. was responsible for developing the website. Y.M. performed the molecular dynamics analysis and assisted with the website. E.N.D. was responsible for curating the human population variants. E.N.D. and M.S. were responsible for calculating intolerance scores. A.A. assisted with SARS-CoV-2 genomic curation. D.E.V.P. was responsible for figures, normal mode analysis and implementation of virtual screening, and assisted with fragment hotspot calculations, website development and supervision. D.B.A. conceived, designed and supervised all aspects of the project and website, and wrote the manuscript. All authors assisted with manuscript writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0693-3.