



OPEN

# Rapid evaluation of COVID-19 vaccine effectiveness against symptomatic infection with SARS-CoV-2 variants by analysis of genetic distance

Lirong Cao<sup>1,2</sup>, Jingzhi Lou<sup>3</sup>, See Yeung Chan<sup>1,3</sup>, Hong Zheng<sup>1,2</sup>, Caiqi Liu<sup>1</sup>, Shi Zhao<sup>1,2</sup>, Qi Li<sup>1,2</sup>, Chris Ka Pun Mok<sup>1,4</sup>, Renee Wan Yi Chan<sup>5,6</sup>, Marc Ka Chun Chong<sup>1,2</sup>, William Ka Kei Wu<sup>4,7,8</sup>, Zigui Chen<sup>9</sup>, Eliza Lai Yi Wong<sup>1,10</sup>, Paul Kay Sheung Chan<sup>9,11</sup>, Benny Chung Ying Zee<sup>1,2</sup>, Eng Kiong Yeoh<sup>1,10</sup> and Maggie Haitian Wang<sup>1,2</sup> ✉

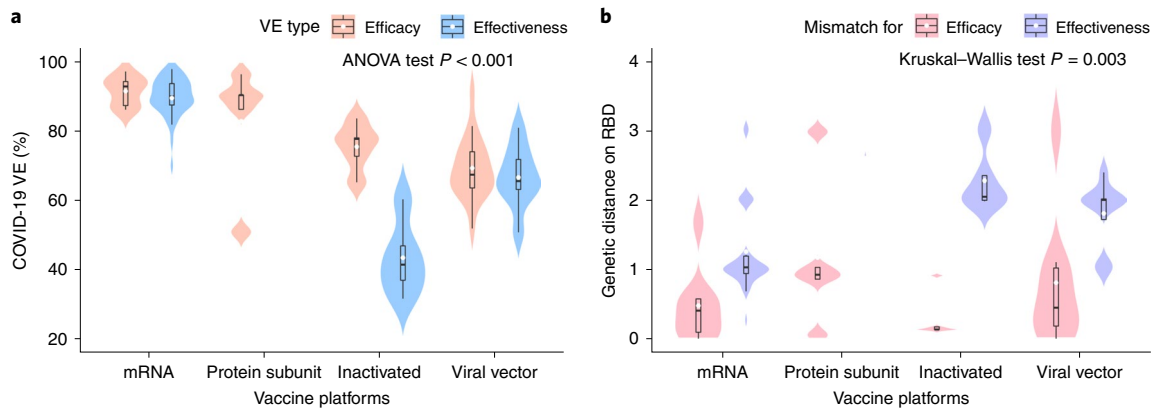
**Timely evaluation of the protective effects of Coronavirus Disease 2019 (COVID-19) vaccines against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants of concern is urgently needed to inform pandemic control planning. Based on 78 vaccine efficacy or effectiveness (VE) data from 49 studies and 1,984,241 SARS-CoV-2 sequences collected from 31 regions, we analyzed the relationship between genetic distance (GD) of circulating viruses against the vaccine strain and VE against symptomatic infection. We found that the GD of the receptor-binding domain of the SARS-CoV-2 spike protein is highly predictive of vaccine protection and accounted for 86.3% ( $P = 0.038$ ) of the VE change in a vaccine platform-based mixed-effects model and 87.9% ( $P = 0.006$ ) in a manufacturer-based model. We applied the VE-GD model to predict protection mediated by existing vaccines against new genetic variants and validated the results by published real-world and clinical trial data, finding high concordance of predicted VE with observed VE. We estimated the VE against the Delta variant to be 82.8% (95% prediction interval: 68.7–96.0) using the mRNA vaccine platform, closely matching the reported VE of 83.0% from an observational study. Among the four sublineages of Omicron, the predicted VE varied between 11.9% and 33.3%, with the highest VE predicted against BA.1 and the lowest against BA.2, using the mRNA vaccine platform. The VE-GD framework enables predictions of vaccine protection in real time and offers a rapid evaluation method against novel variants that may inform vaccine deployment and public health responses.**

Vaccination is a crucial measure to control the scale of SARS-CoV-2 transmission and mitigate the severity of COVID-19. To date, 38 vaccines against SARS-CoV-2 are in early use or have been approved for application in the general population<sup>1</sup>. However, the protective effect of the various vaccine products is challenged by new genetic variants. VE against COVID-19, which measures the relative reduction of risk for a disease outcome in clinical trials or in the general population, exhibited a wide range of variation, from  $-2.7\%$  to  $97.2\%$ <sup>2,3</sup>.

Several factors may contribute to the variations in VE that make it difficult to directly interpret the protective effect of vaccines. The notable contributors include the technology platforms, calendar period of studies, the target population, dosing interval, differences in study protocols and background risk of COVID-19, among others. The various vaccine technology strategies generated non-identical immune responses to provide protection against SARS-CoV-2 infection<sup>4</sup>. For instance, the LNP-mRNA vaccine, mRNA-1273, induces spike (S)-specific IgG, high  $T_H1$  cell

responses, low  $T_H2$  cell responses and  $CD8^+$  T cell responses<sup>5,6</sup>, whereas the inactivated virus vaccine, CoronaVac, elicits robust  $CD4^+$  and  $CD8^+$  T cell responses to the structural proteins, including S, nucleocapsid (N), envelope (E) and matrix (M), in addition to humoral responses<sup>7,8</sup>. Among all the influencing factors, emerging genetic variants relative to the vaccine strain play a critical role in determining vaccine effectiveness. Serology studies showed that neutralizing activity against the Omicron variant decreased substantially in recipients of two COVID-19 vaccine doses<sup>9,10</sup>. Viral structure studies demonstrated that the amino acid substitutions in the receptor-binding domain (RBD) and N-terminal domain (NTD) alter virus–host cell interactions and reshape antigenic surfaces of the major neutralizing sites, leading to immune evasion<sup>9,11–14</sup>. Although the mechanisms of immune escape caused by the new mutations are being elucidated in experimental studies, an integrative framework to quantify the effect of genetic mismatch on VE would be instrumental for efficient evaluation of vaccine protection for any country in real time.

<sup>1</sup>JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>2</sup>CUHK Shenzhen Research Institute, Shenzhen, China. <sup>3</sup>Beth Bioinformatics Co. Ltd., Hong Kong SAR, China. <sup>4</sup>Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>5</sup>Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>6</sup>Hong Kong Hub of Paediatric Excellence, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>7</sup>Department of Anaesthesia and Intensive Care and Peter Hung Pain Research Institute, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>8</sup>State Key Laboratory of Digestive Disease, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>9</sup>Department of Microbiology, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>10</sup>Centre for Health Systems and Policy Research, The Chinese University of Hong Kong, Hong Kong SAR, China. <sup>11</sup>Stanley Ho Centre for Emerging Infectious Diseases, The Chinese University of Hong Kong, Hong Kong SAR, China. ✉e-mail: [maggiew@cuhk.edu.hk](mailto:maggiew@cuhk.edu.hk)



**Fig. 1 | Comparison of COVID-19 VE and genetic mismatch across vaccine platforms. a**, Distribution of the VE estimates for different platforms. The VE of mRNA and protein subunit vaccines are higher than other vaccines (two-sided ANOVA test  $P = 2.2 \times 10^{-14}$ ,  $n = 78$ ). **b**, Distribution of genetic mismatch on RBD for different vaccine technologies. Genetic mismatch is the lowest for mRNA vaccines (two-sided Kruskal-Wallis test  $P = 0.003$ ,  $n = 78$ ). In the box plots, the middle bar indicates the median; the white dot indicates the mean; and the boundaries are Q1 and Q3. Whiskers of the box plot are extended to  $Q3 + 1.5 \times$  interquartile range (IQR) and  $Q1 - 1.5 \times$  IQR.

In this study, we evaluated the link between genetic mismatch of circulating SARS-CoV-2 viruses and reported COVID-19 VE from population studies. Based on our bioinformatics approach previously established for influenza viruses<sup>15,16</sup>, we tailored the VE estimation framework for COVID-19 by controlling the clustered random variation of technology platforms or manufacturers using a mixed-effects model. Through extensive analysis of publicly reported VE studies and genetic sequences, we showed that a substantial proportion of the change in VE could be explained by GD, and we proposed an efficient approach to evaluate vaccine protection against symptomatic COVID-19.

## Results

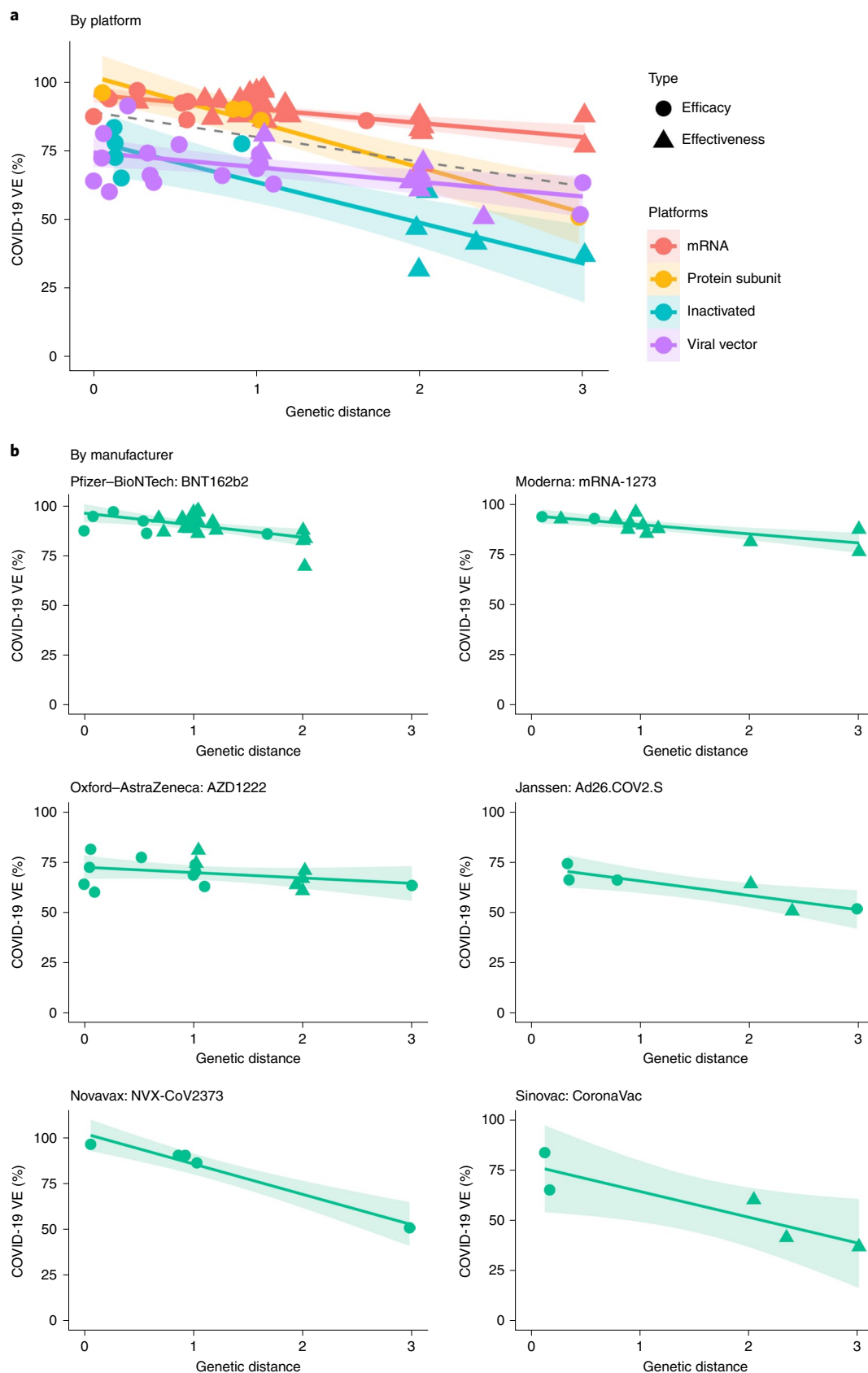
GD, or genetic mismatch, is calculated by the average Hamming distance on the RBD of the genome of the circulating viruses to the vaccine strain during the timeframe of VE studies. VE data used are detailed in Supplementary Table 1. The prediction method for VE was constructed through a mixed-effects model using GD as the main predictor, controlling for the confounding variables, including the midpoint (days) since the second dose and age group of the study. Particularly, variations in VE caused by technology platform or manufacturer were controlled by random effect in the mixed model (see Methods for details). In the following, we will first describe the variations in VE and GD by vaccine platform and then investigate their relationship.

**VE and GD distributions by vaccine platform.** VE and GD of the four vaccine platforms with authorized use are compared in Fig. 1. Within each vaccine platform, the vaccine effectiveness is generally lower compared to the efficacy outcome (Fig. 1a), whereas, in terms of genetic mismatch (Fig. 1b and Extended Data Fig. 1), the vaccine effectiveness cohort encompasses larger genetic mismatch relative to the vaccine efficacy cohorts. The result indicates that genetic mismatch had increased during the mass vaccination phase compared to the earlier clinical trial periods. This could be due to the accumulation of virus mutations through time, as well as the generally longer evaluation period of the effectiveness studies compared to the efficacy trials. Across the technology platforms, vaccine protection (efficacy/effectiveness) shows considerable difference (ANOVA test  $P < 0.001$ ; Fig. 1a). The mRNA vaccines reported the highest mean VE of 90.0% (95% confidence interval (CI): 88.2–91.8,  $n = 39$ ), followed by the protein subunit vaccine (82.9%) (95% CI: 67.0–98.8,  $n = 5$ ), viral vector vaccines (68.5%) (95% CI: 64.8–72.1,

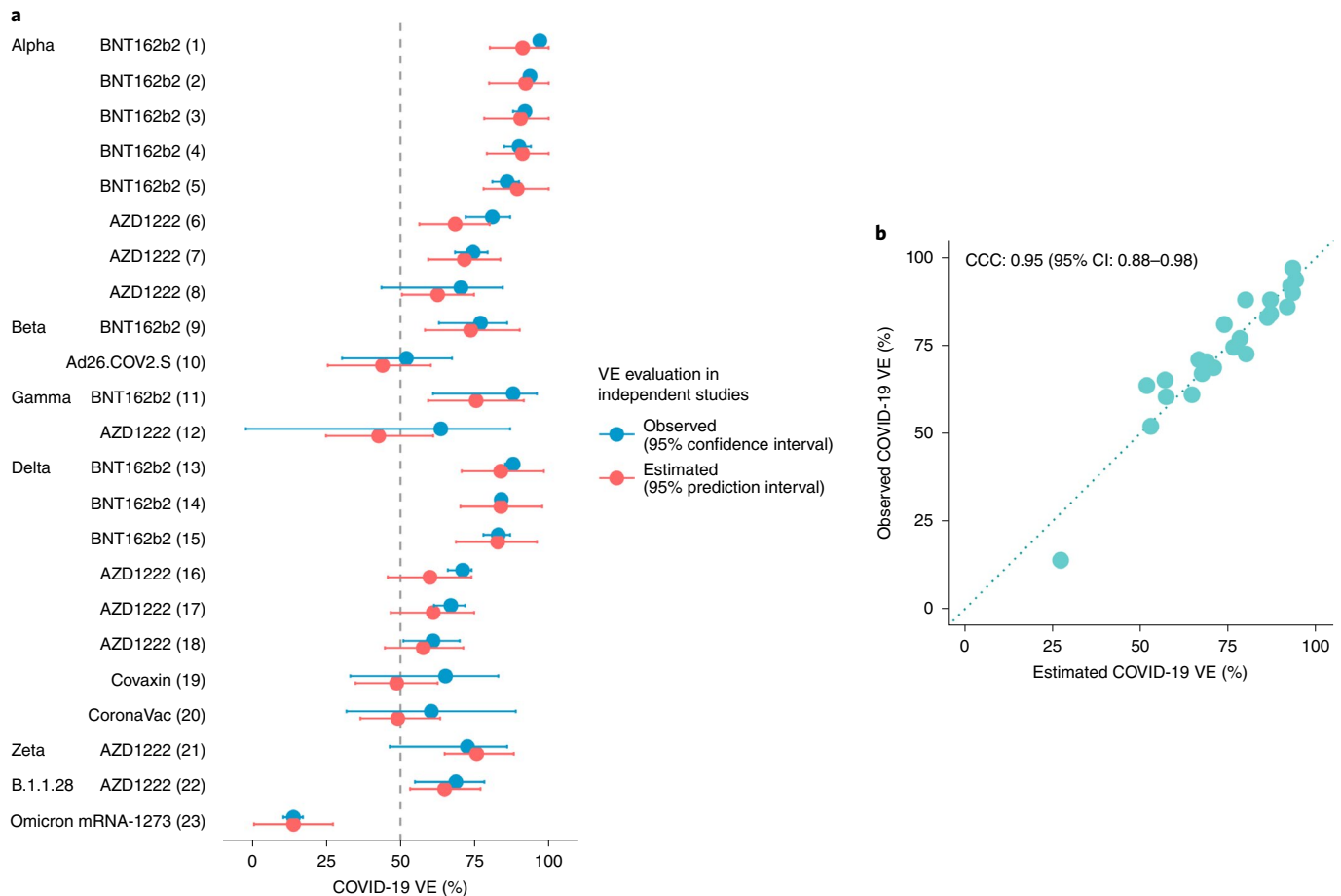
$n = 24$ ) and inactivated vaccine (59.6%) (95% CI: 47.8–71.3,  $n = 10$ ). Interestingly, the genetic mismatch of these platforms shows a perfect reverse trend, of which the mRNA vaccines cohorts correspond to the smallest mismatch, and the other platforms exhibit larger mismatches. This might also be contributed by the timeframe of the vaccine evaluations for these platforms, in which the mRNA trials were the earliest to complete and corresponded to a more homogeneous viral population. The genetic mismatch summarizes the deviation of genetic variants with respect to the vaccine strains, accounting for time, locations and multiple strain co-circulation, for vaccine evaluation at population level using sequencing data.

**Relationship between vaccine protection and GD.** Next, we explored the effect of GD on vaccine protection. At most, 86.3% of the variations in VE can be explained by the GD measure, controlling for the random effects of vaccine technology platforms (Fig. 2a and Supplementary Table 2), and 87.9% of the variation can be explained (Fig. 2b and Supplementary Table 3) when the random effects of six major vaccine products (BNT162b2, mRNA-1273, AZD1222, Ad26.COVS.2, NVX-CoV2373 and CoronaVac) were controlled. Among the candidate genomic regions, genetic mismatch on the RBD demonstrates the strongest influence on vaccine protection, whereas the GD of the non-S proteins shows no association with VE (Extended Data Figs. 2 and 3). For every residue substitution on the RBD, the VE would reduce by an average of 5.2% (95% CI: 2.4–8.0) for mRNA vaccines, 6.8% (95% CI: 4.2–9.4) for viral vector vaccines, 14.3% (95% CI: 9.4–19.2) for protein subunit vaccines and 15.8% (95% CI: 12.4–19.3) for inactivated vaccines ( $P = 0.038$ ) (Supplementary Table 2). The NTD and S protein demonstrate weaker per-amino-acid substitution association with VE ( $P = 0.086$  and  $P = 0.082$ , respectively) (Extended Data Fig. 4 and Supplementary Table 4). When no genetic mismatch is present, VE for the mRNA vaccines is expected to be 95.8% (95% CI: 92.0–99.5), estimated by the RBD region; the protein subunit vaccine's expected VE is similar; and the inactivated and viral vector vaccines are expected to exhibit a systematically lower VE by 17.3% and 20.6% compared to the mRNA vaccines. The estimates using the manufacturer-based model can be found in Supplementary Table 3.

**VE-GD model assessment by validation data.** The VE-GD relationship can be used to make predictions on VE by vaccine type. A total of 57 VE data were used for model training and 23 variant-specific VE data for validation (Supplementary Table 5).



**Fig. 2 | The relationship between VE and GD of the circulating SARS-CoV-2 strains to the vaccine strain on RBD. a**, Negative linear relationships between VE and GD for different vaccine platforms ( $P=0.038$ ,  $R^2=86.3\%$ ). The dashed line was fitted by all data points. **b**, Negative linear relationship between VE and GD for each vaccine product ( $P=0.006$ ,  $R^2=87.9\%$ ). The two-sided  $P$  value was obtained from the mixed-effects model. The colored lines were fitted by data points of each platform. The shaded area indicates 95% CI.



**Fig. 3 | Prediction of VE based on GD.** The VE-GD model was trained using all non-variant-specific VEs, and estimations were made on all ( $n=23$ ) variant-specific VEs. **a**, Estimated and observed VE for variants by vaccine product. Source of the VE data is indexed behind vaccine product names, available in Supplementary Table 5. **b**, Calibration plot for the prediction outcome in validation data. The predicted VEs are close to the observed VEs with a CCC of 0.95 (95% CI: 0.88–0.98).

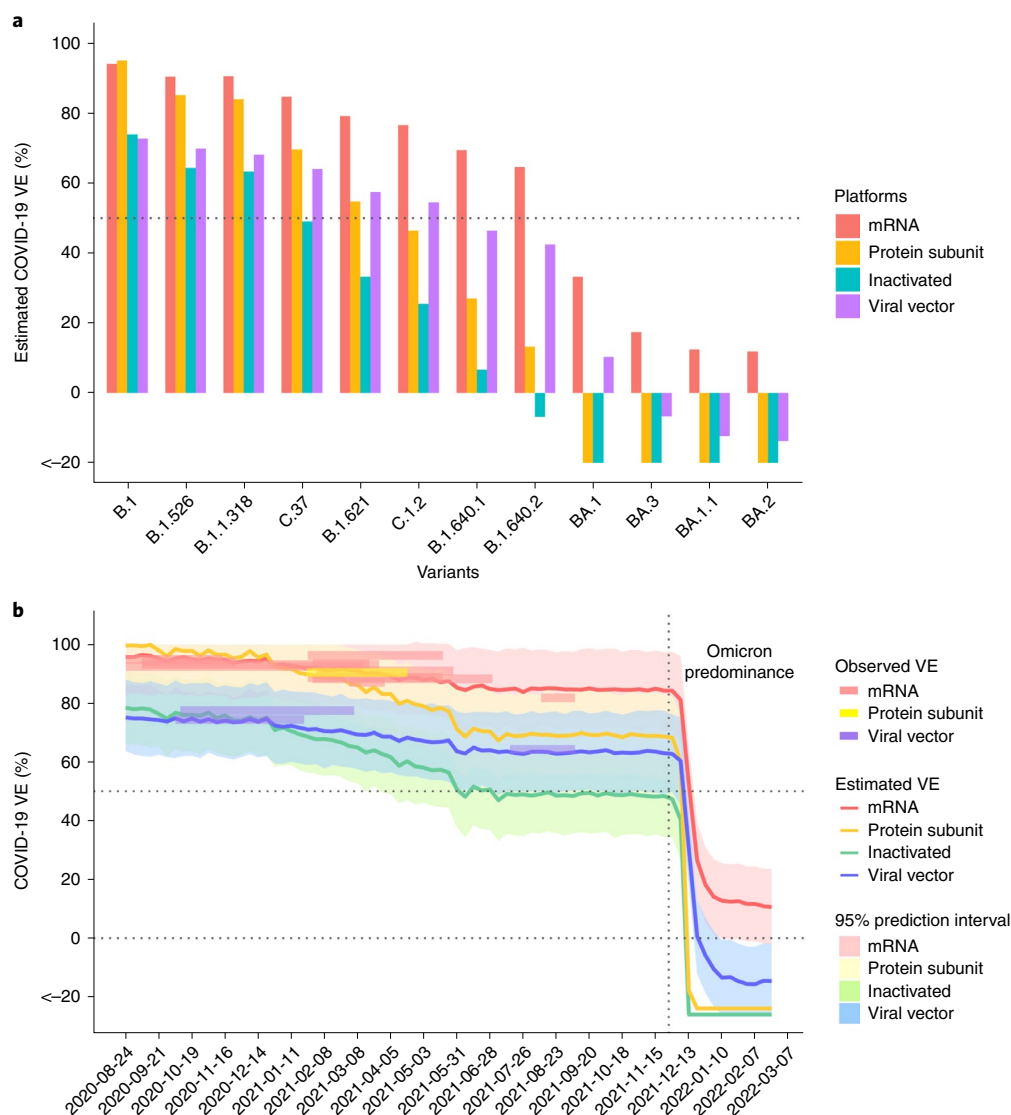
In Fig. 3a, the predicted and observed VEs for the genetic variants are overlaid. The calibration plot (Fig. 3b) shows a close matching, and the concordance correlation coefficient (CCC) reaches a high level of 0.95 (95% CI: 0.88–0.98). Against the Delta variant (B.1.617.2), the estimated VE is 82.8% (95% prediction interval: 68.7–96.0) and 61.1% (95% prediction interval: 46.7–74.9) by the mRNA and viral vector vaccines, respectively (Fig. 3a). These estimates are supported by the observed VE against the Delta variant: the mRNA vaccine BNT162b2 and the viral vector vaccine AZD1222 provided 83% (95% CI: 78–87) and 67.0% (95% CI: 61.3–71.8) protection, respectively<sup>17,18</sup>. The predicted VE is 89.4% (95% prediction interval: 78.1–100.0) for the Alpha variant and 73.7% (95% prediction interval: 58.3–90.2) for the Beta and Gamma variants by the BNT162b2 and mRNA-1273 vaccines, close to the observed VE of 86% (95% CI: 81–90) and 77% (95% CI: 63–86), respectively<sup>19</sup>. Against the Omicron variant, the model predicted an expected VE of 14.0% (95% prediction interval: 0.7–27.3) in California in late 2021, and the observed value was 13.9% (95% CI: 10.5–17.1) for the mRNA-1273 vaccine<sup>20</sup>. These validation results demonstrate high predictive feasibility of using genetic mismatch to estimate vaccine performance.

**Prediction for variants and Omicron sublineages without known VEs.** Next, we fitted the model with all available data and predicted VE against circulating variants as well as the Omicron sublineages for which there are no observed VE data at the time of writing (Fig. 4a).

Interestingly, among the four sublineages of Omicron (BA.1, BA.1.1, BA.2 and BA.3), the expected VEs vary between 11.9% for BA.1 and 33.3% for BA.2, using the mRNA vaccines. This might contribute to the considerable variations in VEs for Omicron reported from observational studies, whose cohorts might have been infected by divergent Omicron sublineages, in addition to differences in immune history. The model predicts that VEs against variants of concern (VOCs) or variants of interest (VOIs) other than the Omicron, such as the Lambda and Mu variants, are expected to be above 50% within 3 months after the second dose of an mRNA vaccine; however, the VEs of inactivated vaccines against symptomatic infection are predicted to wane most under the challenge of new genetic variants.

#### Depicting trend of VE in serial cross-sectional sequencing data.

We demonstrated the application of predicting VE in real time against the circulating virus in a given geographical region, using California as an example. Sequencing data of virus isolates from California were downloaded from public databases. VEs were estimated for the major vaccine platforms at weekly intervals by GD in the serial cross-sectional sequencing data (Fig. 4b). In general, a decreasing trend of VE is depicted, with a sharp drop after the Omicron predominance since December 2021. The observed VEs from clinical trials and observational studies conducted during the period in the United States are overlaid on the prediction outcomes for reference<sup>2,21–32</sup>.

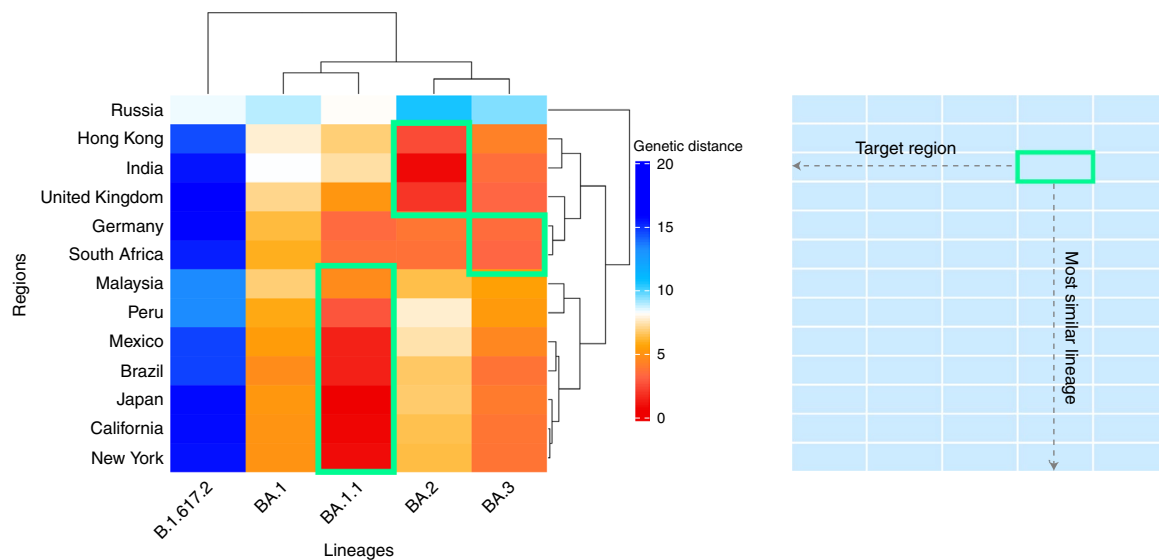


**Fig. 4 | Prediction for SARS-CoV-2 genetic variants with unknown VEs including the Omicron sublineages and in serial cross-sectional sequencing data.** **a**, Predicted VEs for specific variants/sublineages without observed VEs. Omicron sublineages: BA.1, BA.1.1, BA.2 and BA.3. **b**, VEs in California were estimated at weekly intervals for different vaccine platforms. The surveyed VEs from clinical trials or observational studies during the same period are overlaid on the trend curve as colored rectangles for reference. The declining trend of estimated VE captures the influence of virus evolution on population-level immune protection. During the Omicron predominance period, a cliff-like drop of VE is depicted. The shaded areas are 95% prediction interval. The vertical dashed line marks the date of 26 November 2021, which is the earliest time of the Omicron appearance in these data. The top horizontal dashed line marks the 50% efficacy threshold.

**Exploration of candidate vaccine strains.** We further explored the possibility of developing region-specific vaccines and how well they would match the circulating virus profiles. We investigated the optimal candidate vaccine strains for 13 regions, including the United Kingdom, Germany, South Africa, Russia, India, Hong Kong, Malaysia, Japan, California, New York, Mexico, Peru and Brazil. Based on the GD between the vaccine strain and observed viruses circulating in a given region and period, hierarchical clustering of regions was performed to show the similarity of vaccine mismatches (Fig. 5 and Extended Data Fig. 5). We found that, although the Omicron sublineages can match to epidemic viruses in all investigated regions except for Russia during January and February 2022, the dominant sublineages were not the same in these regions. This suggests that updating vaccine compositions with a single genetic variant might not be sufficient for matching the distribution of global viral population.

## Discussion

As novel variants of SARS-CoV-2 keep emerging in the ongoing pandemic, rapid assessment of vaccine performance in populations is crucial to inform public health and clinical responses. This study established an efficient computational framework to estimate COVID-19 VE against symptomatic infection using viral sequence data. We show that the predicted VEs against genetic variants are close to the observed outcomes. The framework has several advantages. First, it enables prediction of VE against novel variants using existing virus surveillance networks to derive a rapid estimate; thus, it could inform timely public health preparedness. Second, it provides an integrated measure to facilitate the interpretation of vaccine effects, which accounts for the potential confounding effects of time and location related to genetic evolution. Third, through mixed-effects modeling, the framework controls for variations by vaccine type, providing a consistent and



**Fig. 5 | Clustering of regions by GD between circulating strains and candidate vaccine strains during January and February 2022.** The candidate strain that gives a lowest genetic mismatch for geographical regions is highlighted in the green box. Rows: target geographical regions; columns: candidate vaccine strain (B.1.617.2: Delta; BA.1–BA.3: Omicron sublineages). The figure shows that the Omicron sublineages can match to the epidemic viruses in most regions, but the dominant sublineages are not the same.

adaptable prediction framework for inclusion of multiple vaccine platforms and manufacturers.

Among candidate genomic regions, the RBD region exhibits the strongest statistical association with VE. Weaker associations between VE and GD were detected for NTD and the entire S protein. These findings are also supported by biological evidence. The RBD is the major target for neutralizing antibodies that interfere with viral receptor binding<sup>33,34</sup>. The NTD is reported to be the target of 5–20% of S-specific monoclonal antibodies from memory B cells against SARS-CoV-2 (refs. 35,36).

Recent studies have investigated the use of neutralization titer as a predictor of vaccine efficacy<sup>37–39</sup>; however, the neutralizing results against SARS-CoV-2 genetic variants showed varying outcomes. The vaccine protection against the B.1.351 variant reduced from 95.0%<sup>2</sup> to 75.0%<sup>40</sup> for BNT162b2 in early 2021. Due to differences in standardization and cohorts, one neutralization study showed that the titer against B.1.351 is 7.6-fold and nine-fold lower compared to the early Wuhan-related Victoria variant in the BNT162b2 vaccine serum and AZD1222 vaccine serum, respectively<sup>41</sup>, whereas another experiment reported a 2.7-fold decrease in neutralization titers against the B.1.351 lineage in the BNT162b2-elicited serum<sup>42</sup>. Similar results have also been observed for the Omicron variant<sup>43–45</sup>. The varying neutralization results increase the challenge of inferring vaccine performance solely by neutralization levels. The association of neutralization with protection across studies showed that neutralizing antibodies might not be deterministic in mediating protection, and the effect of other vaccine-induced immune responses also need to be quantified. This work uses an alternative angle to bridge the link between genetic variations and population-level vaccine responses. Further investigations are needed to integrate potential correlates of vaccine protection and improve the existing framework.

Although 42% of the world population has not completed the full vaccine primary series up to this date<sup>46</sup>, additional booster doses of vaccine are being rolled out in many places. Neutralization activity after the booster can be restored to a higher level for a short period of time. BNT162b2 immune sera of individuals who received only two doses had a low ability to neutralize the Omicron variant, whereas a third dose of the BNT162b2 increased the Omicron-neutralizing titer 23-fold relative to their level at 21 days after the second dose<sup>47</sup>. Similar results have been reported for the

mRNA-1273 vaccine<sup>48</sup>. The booster-enhanced neutralizing level against Omicron was lower than that against the Beta, Delta and Wuhan strains and declined faster than those against the D614G variant<sup>47,48</sup>. Recent studies showed that the VE against symptomatic infection of Omicron is restored up to near 50% after the booster. In Qatar, the VE against symptomatic Omicron infection was 56.6% and 53.1% for the BNT162b2 and mRNA1732 vaccines, respectively, 1 month after the third dose<sup>49</sup>; and, in Israel, the VE against symptomatic Omicron infection was 43% and 31% for BNT162b2 and mRNA1273, respectively, 1 month after the fourth dose among healthcare workers<sup>50</sup>. The flexible VE–GD framework proposed here could be further extended to account for the booster's protection as more effectiveness data of homologous and heterologous booster studies are available.

VE against infection is generally lower compared to the VE against symptomatic infection. For instance, in the Coronavirus Efficacy (COVE) phase 3 trial of the mRNA-1732 vaccine, the VEs for infection and symptomatic infection are 82.0% and 93.2%, respectively<sup>51</sup>. In view of waning immunity, a systematic review including 78 VE studies up to 2 December 2021 showed that the VE dropped by 21.0% (95% CI: 13.9–29.8) and 24.9% (95% CI: 13.4–41.6) against infection and symptomatic infection, respectively, 6 months after the second dose, aggregating the data from several vaccine platforms<sup>52</sup>. VE against severe disease or hospitalization showed longer preservation compared to the protection against symptomatic infection. In Qatar and Canada, the VE against hospitalization due to infection with the Alpha, Beta and Delta variants among all age groups was above 90% after the second dose of the mRNA-1273, BNT162b2 and AZD1222 vaccines<sup>53–56</sup>. VE against hospitalization with Delta infection remained at above 80% in the United Kingdom 20 weeks after vaccination with the BNT162b2 and AZD1222 vaccines<sup>57</sup>. In Qatar and South Africa, VE against hospitalization was in the range of 70–80% during the Omicron predominance within 6 months after the second dose for mRNA vaccines<sup>49,58</sup>.

Previously, the effect of genetic diversity on vaccine efficacy was investigated by sieve analysis, originating in the study of the human immunodeficiency virus 1 (HIV-1) vaccines<sup>59–61</sup>. Sieve analysis compares the infection strains between vaccinated and unvaccinated individuals and estimates the odds ratio of a viral strain type to penetrate the vaccine protection barrier. The sieve method

requires individual-level data of virus isolate sequences and infection outcome of trial participants, whereas the model proposed in this study uses viral sequences in the general population and integrates multiple VE studies. Other studies have considered the proportion of genetic mismatch in the dominant epitope region to account for variations in the VE against influenza viruses<sup>62,63</sup>, whereas the VE-GD model in this report provides a unified framework to account for multiple genes and vaccine types.

This study has several limitations. The scope of inference is subject to the range of VE studies included in model fitting; thus, the VE estimated is presumably for a time close to the second vaccine dose. In model estimation, the effect of waning immunity on VE was controlled by a proxy time variable at population level, and the VE decline corresponding to time was estimated to be 2.4% (95% CI: 1.0–3.8) per 30 days for mRNA vaccines. This estimation is in line with the phase 2/3 efficacy trial of the BNT162b2 vaccine through 6 months of follow-up<sup>32</sup>, which showed an average decline of 2.5% per month by comparing the VE after 4–6 months to VE within 2 months since the second dose. The exact relationship between time and waning of host immunity will be calibrated in individual-level data, in which the main variable of interest is time-to-infection. For these analyses, including the genetic mismatch information would be helpful to control for the genetic variant's effect on vaccine breakthrough alongside waning of host immunity. Second, VE prediction in this study only considered the GD of vaccine strain to circulation strains, and the effect of prior infection on vaccine protection was not captured. Studies showed that natural infection, either before or after vaccination, substantially increased vaccine protection for symptomatic infection and hospitalization during the Beta-predominant and Delta-predominant periods<sup>64</sup> and against the Omicron variant by the mRNA vaccine<sup>65,66</sup>. As more hybrid immunity data become available, the mixed-effects prediction model could be extended to account for this additional level of variation. Moreover, bias might occur if sequences in databases disproportionately represented regions with known circulation of a given variant. Enhanced efforts are needed to ensure better geographical representativeness of available SARS-CoV-2 sequences. Despite these limitations, we demonstrated a robust relationship between genetic mismatch and VE, which we validated using independent data.

To conclude, this work developed a modeling framework integrating data from genetics and epidemiological studies for estimating COVID-19 vaccine effectiveness against a specific variant or for a particular cohort in a given period and region. Rapid assessment of VE against an evolving pathogen can be a useful instrument to inform vaccine development, distribution and public health responses.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-022-01877-1>.

Received: 16 September 2021; Accepted: 20 May 2022;

Published online: 16 June 2022

### References

- Basta, N. E. & Moodie, E. M. M., on behalf of the VIPER (Vaccines, Infectious disease Prevention, and Epidemiology Research) Group COVID-19 Vaccine Development and Approvals Tracker Team. COVID-19 Vaccine Tracker. <https://covid19.trackvaccines.org/>
- Polack, F. P. et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N. Engl. J. Med.* **383**, 2603–2615 (2020).
- Andrews, N. et al. Covid-19 vaccine effectiveness against the Omicron (B.1.1.529) variant. *N. Engl. J. Med.* **386**, 1532–1546 (2022).
- Dai, L. & Gao, G. F. Viral targets for vaccines against COVID-19. *Nat. Rev. Immunol.* **21**, 73–82 (2021).
- Jackson, L. A. et al. An mRNA vaccine against SARS-CoV-2—preliminary report. *N. Engl. J. Med.* **383**, 1920–1931 (2020).
- Anderson, E. J. et al. Safety and immunogenicity of SARS-CoV-2 mRNA-1273 vaccine in older adults. *N. Engl. J. Med.* **383**, 2427–2438 (2020).
- Mok, C. K. P. et al. Comparison of the immunogenicity of BNT162b2 and CoronaVac COVID-19 vaccines in Hong Kong. *Respirology* **27**, 301–310 (2022).
- Melo-Gonzalez, F. et al. Recognition of variants of concern by antibodies and T cells induced by a SARS-CoV-2 inactivated vaccine. *Front. Immunol.* **12**, 747830 (2021).
- Dejnirattisai, W. et al. SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses. *Cell* **185**, 467–484 (2022).
- Cheng, S. M. S. et al. Neutralizing antibodies against the SARS-CoV-2 Omicron variant BA.1 following homologous and heterologous CoronaVac or BNT162b2 vaccination. *Nat. Med.* **28**, 486–489 (2022).
- McCallum, M. et al. Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. *Science* **375**, 864–868 (2022).
- Cui, Z. et al. Structural and functional characterizations of infectivity and immune evasion of SARS-CoV-2 Omicron. *Cell* **185**, 860–871 (2022).
- Cai, Y. et al. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *Science* **373**, 642–648 (2021).
- Gobeil, S. M.-C. et al. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science* **373**, eabi6226 (2021).
- Cao, L. et al. In silico prediction of influenza vaccine effectiveness by sequence analysis. *Vaccine* **39**, 1030–1034 (2021).
- Cao, L. et al. Differential influence of age on the relationship between genetic mismatch and A(H1N1)pdm09 vaccine effectiveness. *Viruses* **13**, 619 (2021).
- Sheikh, A., McMenamin, J., Taylor, B. & Robertson, C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **397**, 2461–2462 (2021).
- Lopez Bernal, J. et al. Effectiveness of Covid-19 vaccines against the B.1.617.2 (Delta) variant. *N. Engl. J. Med.* **385**, 585–594 (2021).
- Charmet, T. et al. Impact of original, B.1.1.7, and B.1.351/P.1 SARS-CoV-2 lineages on vaccine effectiveness of two doses of COVID-19 mRNA vaccines: results from a nationwide case-control study in France. *Lancet Reg. Health Eur.* **8**, 100171 (2021).
- Tseng, H.F. et al. Effectiveness of mRNA-1273 against SARS-CoV-2 Omicron and Delta variants. *Nat. Med.* **28**, 1063–1071 (2022).
- Pilishvili, T. et al. Effectiveness of mRNA Covid-19 vaccine among U.S. health care personnel. *N. Engl. J. Med.* **385**, e90 (2021).
- Sadoff, J. et al. Safety and efficacy of single-dose Ad26.COV2.S vaccine against Covid-19. *N. Engl. J. Med.* **384**, 2187–2201 (2021).
- Dunkle, L. M. et al. Efficacy and safety of NVX-CoV2373 in adults in the United States and Mexico. *N. Engl. J. Med.* **386**, 531–543 (2022).
- Falsey, A. R. et al. Phase 3 safety and efficacy of AZD1222 (ChAdOx1 nCoV-19) Covid-19 vaccine. *N. Engl. J. Med.* **385**, 2348–2360 (2021).
- Baden, L. R. et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.* **384**, 403–416 (2021).
- El Sahly, H. M. et al. Efficacy of the mRNA-1273 SARS-CoV-2 vaccine at completion of blinded phase. *N. Engl. J. Med.* **385**, 1774–1785 (2021).
- Pilishvili, T. et al. Interim estimates of vaccine effectiveness of Pfizer-BioNTech and Moderna COVID-19 vaccines among health care personnel—33 U.S. sites, January–March 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 753–758 (2021).
- Cavanaugh, A. M. et al. COVID-19 outbreak associated with a SARS-CoV-2 R.1 lineage variant in a skilled nursing facility after vaccination program—Kentucky, March 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 639–643 (2021).
- Kim, S. S. et al. mRNA vaccine effectiveness against COVID-19 among symptomatic outpatients aged ≥16 years in the United States, February–May 2021. *J. Infect. Dis.* **jiab451** (2021).
- Lin, D. Y. et al. Effectiveness of Covid-19 vaccines over a 9-month period in North Carolina. *N. Engl. J. Med.* **386**, 933–941 (2022).
- Bruxvoort, K. J. et al. Real-world effectiveness of the mRNA-1273 vaccine against COVID-19: interim results from a prospective observational cohort study. *Lancet Reg. Health Am.* **6**, 100134 (2022).
- Thomas, S. J. et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine through 6 months. *N. Engl. J. Med.* **385**, 1761–1773 (2021).
- Ju, B. et al. Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115–119 (2020).
- Piccoli, L. et al. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* **183**, 1024–1042 (2020).
- McCallum, M. et al. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347 (2021).
- McCallum, M. et al. SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern. *Science* **373**, 648–654 (2021).

37. Khoury, D. S. et al. Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. *Nat. Med.* **27**, 1205–1211 (2021).
38. Bergwerk, M. et al. Covid-19 breakthrough infections in vaccinated health care workers. *N. Engl. J. Med.* **385**, 1474–1484 (2021).
39. Feng, S. et al. Correlates of protection against symptomatic and asymptomatic SARS-CoV-2 infection. *Nat. Med.* **27**, 2032–2040 (2021).
40. Abu-Raddad, L. J., Chemaitelly, H. & Butt, A. A. Effectiveness of the BNT162b2 Covid-19 vaccine against the B.1.1.7 and B.1.351 variants. *N. Engl. J. Med.* **385**, 187–189 (2021).
41. Zhou, D. et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* **184**, 2348–2361 (2021).
42. Liu, Y. et al. Neutralizing activity of BNT162b2-elicited serum. *N. Engl. J. Med.* **384**, 1466–1468 (2021).
43. Cele, S. et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. *Nature* **602**, 654–656 (2022).
44. Muik, A. et al. Neutralization of SARS-CoV-2 Omicron by BNT162b2 mRNA vaccine-elicited human sera. *Science* **375**, 678–680 (2022).
45. Nemet, I. et al. Third BNT162b2 vaccination neutralization of SARS-CoV-2 Omicron infection. *N. Engl. J. Med.* **386**, 492–494 (2022).
46. Ritchie, H. et al. Coronavirus Pandemic (COVID-19). <https://ourworldindata.org/coronavirus>
47. Muik, A. et al. Neutralization of SARS-CoV-2 Omicron by BNT162b2 mRNA vaccine—elicited human sera. *Science* **375**, 678–680 (2022).
48. Pajon, R. et al. SARS-CoV-2 Omicron variant neutralization after mRNA-1273 booster vaccination. *N. Engl. J. Med.* **386**, 1088–1091 (2022).
49. Chemaitelly, H. et al. Duration of protection of BNT162b2 and mRNA-1273 COVID-19 vaccines against symptomatic SARS-CoV-2 Omicron infection in Qatar. Preprint at <https://www.medrxiv.org/content/10.1101/2022.02.07.22270568v1> (2022).
50. Regev-Yochay, G. et al. Efficacy of a fourth dose of Covid-19 mRNA vaccine against Omicron. *N. Engl. J. Med.* **386**, 1377–1380 (2022).
51. El Sahly, H. M. et al. Efficacy of the mRNA-1273 SARS-CoV-2 vaccine at completion of blinded phase. *N. Engl. J. Med.* **385**, 1774–1785 (2021).
52. Feikin, D. R. et al. Duration of effectiveness of vaccines against SARS-CoV-2 infection and COVID-19 disease: results of a systematic review and meta-regression. *Lancet* **399**, 924–944 (2022).
53. Tang, P. et al. BNT162b2 and mRNA-1273 COVID-19 vaccine effectiveness against the SARS-CoV-2 Delta variant in Qatar. *Nat. Med.* **27**, 2136–2143 (2021).
54. Abu-Raddad, L. J., Chemaitelly, H., Butt, A. A. & National Study Group for, C.-V. Effectiveness of the BNT162b2 Covid-19 vaccine against the B.1.1.7 and B.1.351 variants. *N. Engl. J. Med.* **385**, 187–189 (2021).
55. Chung, H. et al. Effectiveness of BNT162b2 and mRNA-1273 covid-19 vaccines against symptomatic SARS-CoV-2 infection and severe covid-19 outcomes in Ontario, Canada: test negative design study. *BMJ* **374**, n1943 (2021).
56. Nasreen, S. et al. Effectiveness of COVID-19 vaccines against symptomatic SARS-CoV-2 infection and severe outcomes with variants of concern in Ontario. *Nat. Microbiol.* **7**, 379–385 (2022).
57. Andrews, N. et al. Duration of protection against mild and severe disease by Covid-19 vaccines. *N. Engl. J. Med.* **386**, 340–350 (2022).
58. Collie, S., Champion, J., Moultrie, H., Bekker, L. G. & Gray, G. Effectiveness of BNT162b2 vaccine against Omicron variant in South Africa. *N. Engl. J. Med.* **386**, 494–496 (2022).
59. Gilbert, P., Self, S., Rao, M., Naficy, A. & Clemens, J. Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *J. Clin. Epidemiol.* **54**, 68–85 (2001).
60. Rolland, M. et al. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat. Med.* **17**, 366–371 (2011).
61. Rolland, M. & Gilbert, P. B. Sieve analysis to understand how SARS-CoV-2 diversity can impact vaccine protection. *PLoS Pathog.* **17**, e1009406 (2021).
62. Gupta, V., Earl, D. J. & Deem, M. W. Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine* **24**, 3881–3888 (2006).
63. Munoz, E. T. & Deem, M. W. Epitope analysis for influenza vaccine design. *Vaccine* **23**, 1144–1148 (2005).
64. Abu-Raddad, L. J. et al. Association of prior SARS-CoV-2 infection with risk of breakthrough infection following mRNA vaccination in Qatar. *JAMA* **326**, 1930–1939 (2021).
65. Andeweg, S. P. et al. Protection of COVID-19 vaccination and previous infection against Omicron BA.1, BA.2 and Delta SARS-CoV-2 infections. Preprint at <https://www.medrxiv.org/content/10.1101/2022.02.06.22270457v1> (2022).
66. Altarawneh, H. N. et al. Effect of prior infection, vaccination, and hybrid immunity against symptomatic BA.1 and BA.2 Omicron infections and severe COVID-19 in Qatar. Preprint at <https://www.medrxiv.org/content/10.1101/2022.03.22.22272745v1> (2022).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022



## Methods

**VE data.** VE is calculated by  $(1 - RR) \times 100$ , where RR is the relative risk of a disease outcome in the vaccinated group compared to the unvaccinated group. Vaccine efficacy is measured in randomized controlled trials, whereas vaccine effectiveness is obtained from observational studies. VE reports before 24 December 2021 were collected from published articles and preprint articles. Inclusion criteria for the vaccine effectiveness studies include: target population is a cohort without special conditions; the primary outcome is symptomatic COVID-19 infection after the second vaccine dose; and the study period of VE evaluation is clearly reported. A total of 78 VE data from 49 studies were obtained for estimating the effect size of GD, among which were 33 efficacy data and 45 were effectiveness data. The vaccine efficacy studies include 28 phase 3 trials, one phase 2 trial and four phase 2/3 trials. The vaccine effectiveness studies include 16 cohorts and 29 case-control studies. Detailed information of VE studies is available in Supplementary Table 1.

**Genetic sequences.** Human SARS-CoV-2 strains with collection dates ranging from 4 August 2020 to 6 March 2022 were retrieved from the Global Initiative on Sharing All Influenza Data (GISAID) EpiCoV database<sup>57</sup>. All available sequences that matched to the period and locations of the clinical trials or observational studies totaled 1,984,241 full-length genome sequences from 31 geographical regions. The sources of SARS-CoV-2 sequences involved in this study are reported in the Supplementary Acknowledgement Table. Strains with duplicated names and unclear collection time of samples were removed. Multiple sequence alignment was performed using MAFFT (version 7). The ‘Wuhan-Hu-1’ genome (GenBank NC\_045512.2 or GISAID EPI\_ISL\_402125) was set as the reference sequence. The variants involved in this study are summarized in Supplementary Tables 6 and 7. Lineage classification for sequences was referenced from the GISAID.

**Statistical methods.** GD. Following our previous framework developed for influenza virus<sup>15</sup>, let  $X = \{x_{ij}\}$  denote the  $i$ -th sample from the GISAID database collected for a target population, where  $i = 1, \dots, n$ ,  $j = 1, \dots, J$ ; and let  $V = \{v_j\}$  denote the vaccine strain applied in the target population, where index  $j$  indicates the  $j$ -th codon position in the sequence. Denote the amino acids in a given genomic region as  $W = \{w_k\}$ , where  $k$  is the index for codon positions contained in the segment,  $k = 1, \dots, K$ ,  $0 \leq K \leq J$ . Suppose the Hamming distance is used as a basic measure of dissimilarity between two sequences, the vaccine genetic distance ( $d$ ) calculated for the target population is:

$$d = \sum_{i=1}^n d_i/n = \sum_{i=1}^n \sum_{k=1}^K I(v_{w_k} \neq x_{w_k})/n. \quad (1)$$

Thus, the  $d$  summarized the average amino acids mismatch of circulating strains versus the vaccine strain based on a given genomic segment in a target population. In this study, we considered a wide range of candidate  $W$ , including the RBD, NTD and S, E, M, N, ORF1ab and accessory proteins. A schematic representation of the SARS-CoV-2 genome and the structure of S protein are available in Supplementary Figs. 1 and 2. All vaccine strains are based on the Wuhan strain isolated in January 2020. When the target population is composed of individuals infected with multiple co-circulating variants, the  $d$  captures the average mismatch over all co-circulating variants in the cohort, whereas, when the target population is a single genetic variant,  $d$  captures the variant-specific distance.

**The VE-GD mixed-effects model.** A two-level mixed-effects model was adopted to account for the random effect associated with vaccine type (technology platform or manufacturer). The genetic distance,  $d_{ij}$ , is the main predictor variable for study  $i$  and vaccine type  $j$ ,  $i = 1, \dots, n_j$ , and  $n_j$  is the number of studies for vaccine type  $j$ . Therefore, the following random intercept and random slope model is specified for the VE response  $Y_j$ :

$$Y_j = X_j\beta + Z_ju_j + \varepsilon_j \quad (2)$$

In the equation,  $X_j$  is the covariate matrix of fixed factors, and  $\beta$  is the fixed effect vector.  $Z_j = [1, \mathbf{d}_j]$  is the matrix containing a unit vector and the  $n_j$ -length genetic distance vector  $\mathbf{d}_j$ , and  $u_j = (u_{0j}, u_{1j})^T$  is composed of a random intercept variable  $u_{0j}$  and a random slope variable  $u_{1j}$ .  $u_j \sim N(0, D)$ , where  $D$  is a variance component matrix. The fixed factors include the age category of the study, midpoint (days) after the second dose extracted from each study and the genetic distance  $\mathbf{d}_j$ .  $\varepsilon_j \sim N(0, R_j)$  is the error term of the mixed-effects model,  $R_j = \sigma^2 I_{n_j}$ . The model was fitted using the R package lmerTest<sup>68</sup>. The prediction interval of the mixed-effects model was calculated using the R package merTools<sup>69</sup>. All analyses were performed using R statistical software (version 4.0.3). Statistical significance was declared if  $P < 0.05$ .

Model assessment was performed in a training-validation setting. A total of 23 variant-specific VEs were extracted from the data as the validation set (Supplementary Table 5). The model was fitted using the remaining 57 VEs (non-variant specific), and predictions were made for the genetic variants. The agreement between the predicted and observed VEs is measured by the CCC<sup>70</sup>.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All data used in this study are publicly available. Detailed information of VE outcomes is available in the Supplementary Materials. Viral sequence data were downloaded from the GISAID at <http://platform.gisaid.org/>, and the accession numbers are provided in the online Supplementary Acknowledgement Table (<https://github.com/VaccineEffectivenessPrediction/COVID-19-Vaccine-Effectiveness>).

## Code availability

All code is freely available at <https://github.com/VaccineEffectivenessPrediction/COVID19-Vaccine-Effectiveness>.

## References

- Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data—from vision to reality. *Euro. Surveill.* **22**, 30494 (2017).
- Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
- Knowles, J. E., Frederick, C. & Knowles, M. J. E. merTools: Tools for Analyzing Mixed Effect Regression Models. <https://cran.r-project.org/web/packages/merTools/index.html> (2020).
- Lin, L. I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268 (1989).

## Acknowledgements

A complete GISAID acknowledgement table is in the online Supplementary Materials. We thank the contributions of all the healthcare workers and scientists, the GISAID team and the submitting and originating laboratories. This work was partially supported by the Health and Medical Research Fund, the Food and Health Bureau, the Government of the Hong Kong Special Administrative Region (COVID190103; M.H.W. and INF-CUHK-1; E.K.Y.), the National Natural Science Foundation of China (31871340; M.H.W. and 71974165; M.K.C.C.), Chinese University of Hong Kong grants PIEF/Ph2/COVID/06 and 4054600 (M.H.W.), the Guangdong-Hong Kong-Macau Joint Laboratory of Respiratory Infectious Disease (20191205; C.K.P.M.) and the visiting scientist scheme from the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (C.K.P.M.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We appreciate the constructive comments from the editor and reviewers that improved this work.

## Author contributions

M.H.W. conceived the study. L.C. and M.H.W. wrote the manuscript. L.C., H.Z. and S.Y.C. collected data. L.C. processed data, carried out the analysis and wrote the first draft. J.L., C.L., S.Z., Q.L., C.K.P.M., R.W.Y.C., M.K.C.C., W.K.K.W., Z.C., E.L.Y.W., P.K.S.C., B.C.Y.Z. and E.K.Y. critically read and revised the manuscript and gave final approval for publication.

## Competing interests

M.H.W. and B.C.Y.Z. are shareholders of Beth Bioinformatics Co., Ltd. B.C.Y.Z. is a shareholder of Health View Bioanalytics Ltd. S.Y.C. and J.L. are employees of Beth Bioinformatics Co., Ltd. All other authors declare no competing interests.

## Additional information

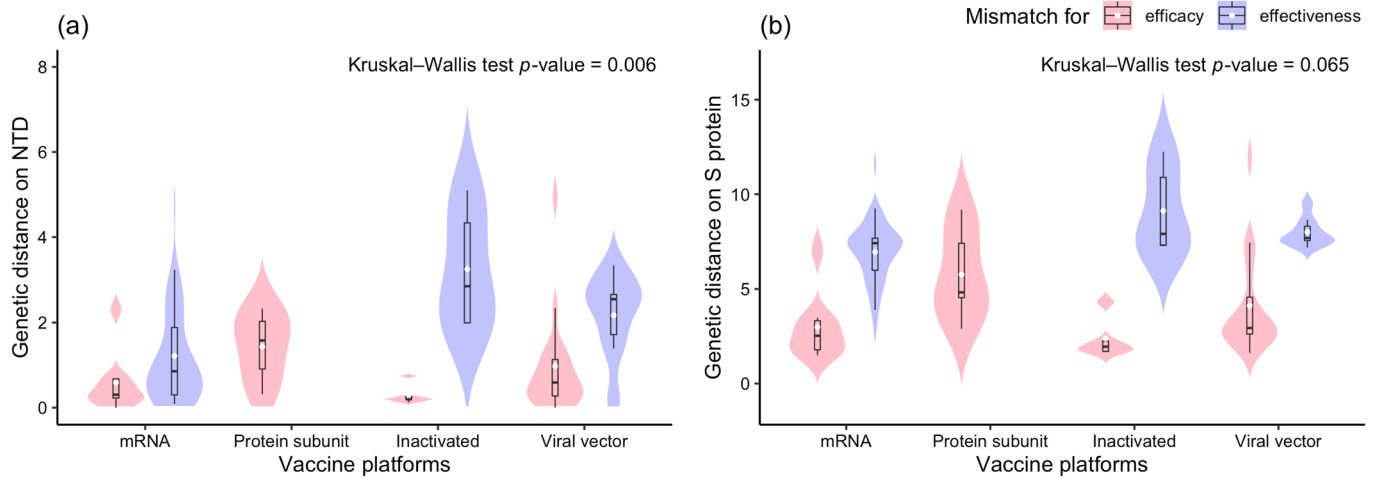
**Extended data** is available for this paper at <https://doi.org/10.1038/s41591-022-01877-1>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41591-022-01877-1>.

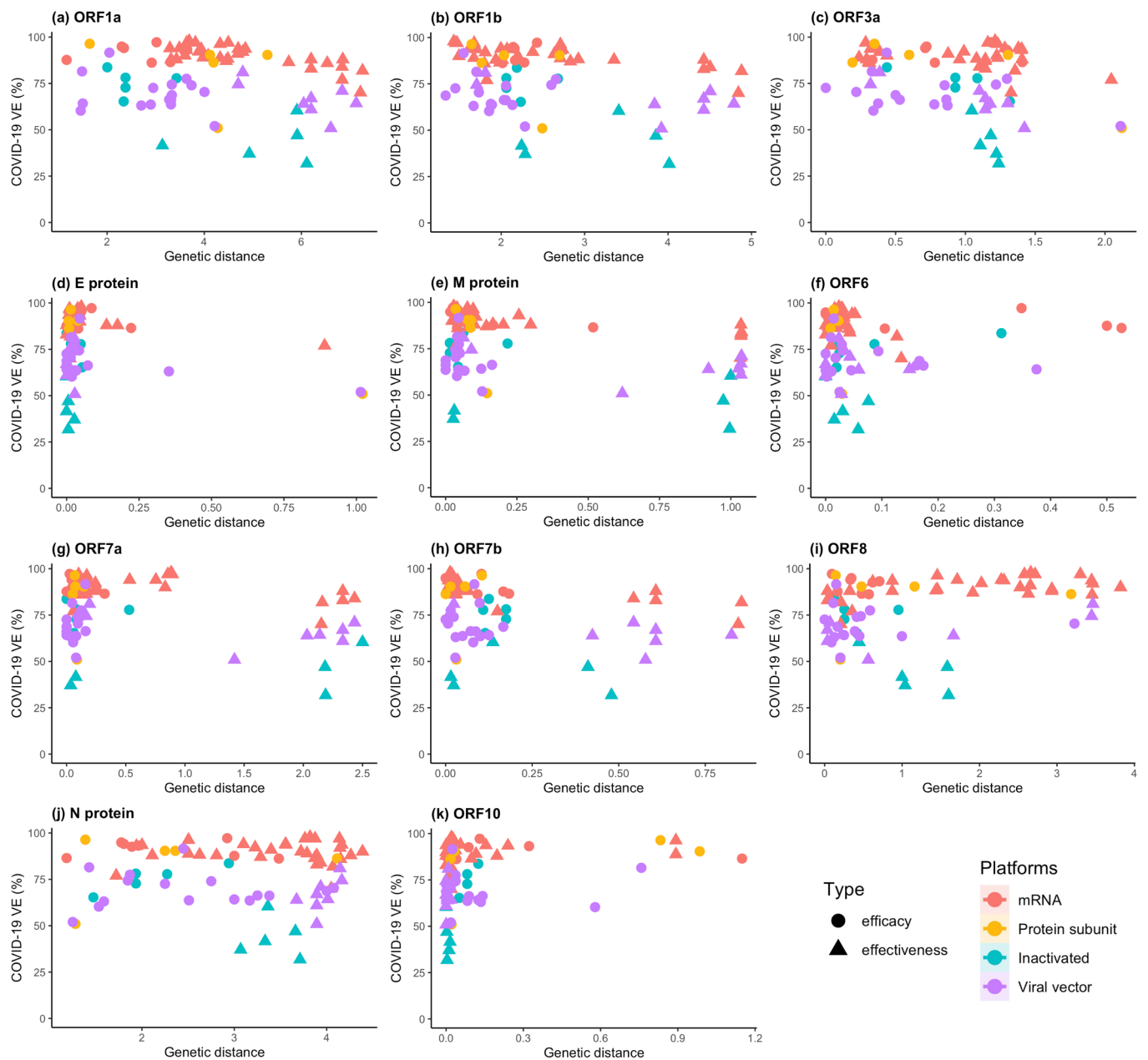
**Correspondence and requests for materials** should be addressed to Maggie Haitian Wang.

**Peer review information** *Nature Medicine* thanks Dean Follmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary handling editor: Alison Farrell, in collaboration with the *Nature Medicine* team.

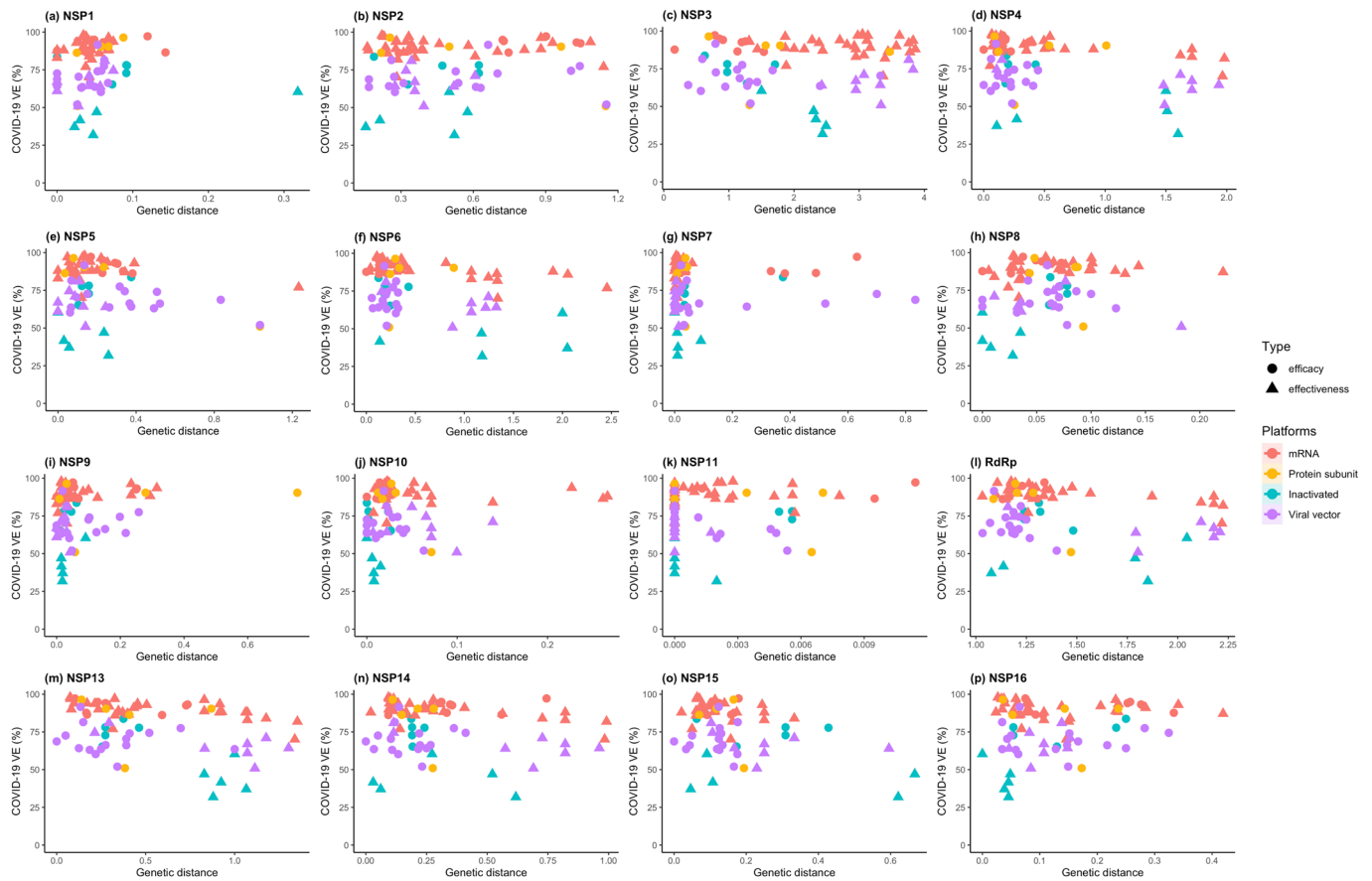
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



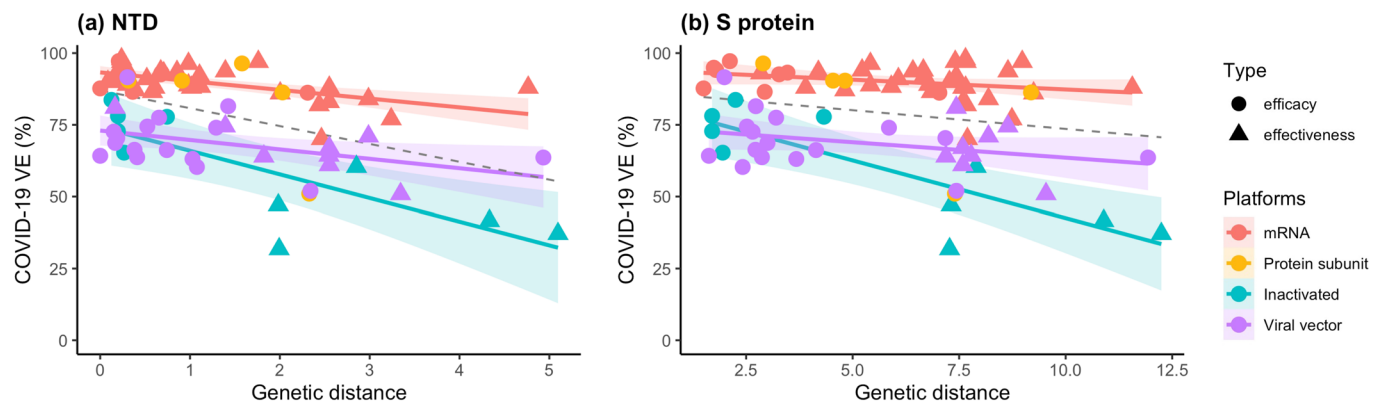
**Extended Data Fig. 1 | Distribution of genetic mismatch on the NTD and complete S protein.** The genetic mismatch on the NTD and S protein was measured. The results show that the genetic mismatch is lowest for mRNA vaccines (Kruskal-Wallis test: two-sided  $P$  of NTD = 0.006, two-sided  $P$  of S protein = 0.065;  $n = 78$ ). In the box plots, the middle bar indicates the median, the white dot indicates the mean, and the boundaries are Q1 and Q3. Whiskers of the box plot are extended to  $Q3 + 1.5 \times IQR$  and  $Q1 - 1.5 \times IQR$ .



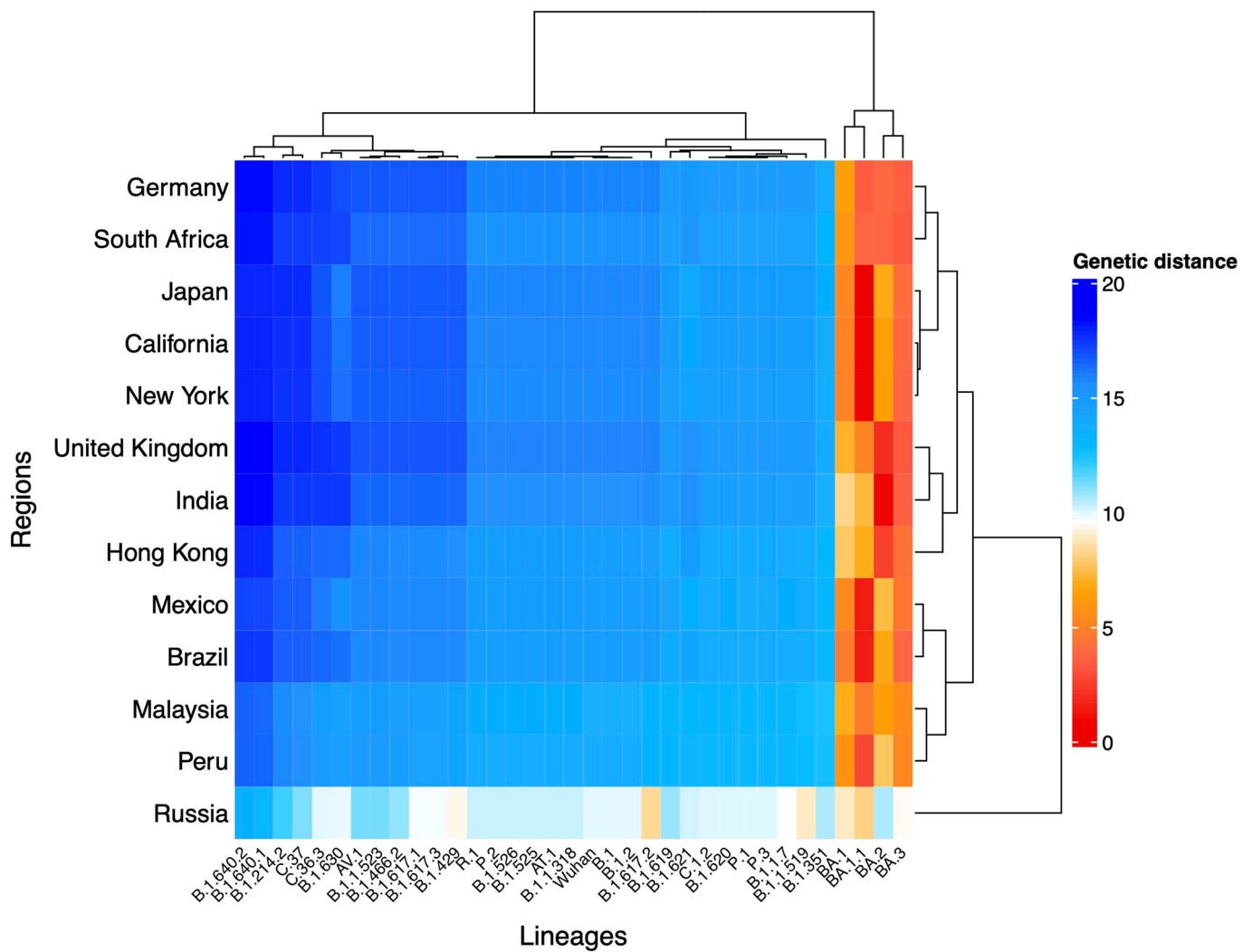
**Extended Data Fig. 2 | Scatterplot of the observed VE and genetic distance on the non-S proteins of SARS-CoV-2.** The same analysis under the null hypothesis to explore the association of VE with genetic distance was performed on the structural proteins containing envelope (E), membrane (M) and nucleocapsid (N); ORF1ab; accessory proteins containing ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10 proteins. No significant relationship with VE was observed.



**Extended Data Fig. 3 | Scatterplot of the observed VE and genetic distance on the non-structural proteins (NSPs).** The ORF1ab polyprotein is composed of 16 non-structural proteins (NSPs). The genetic distance of each NSP was also calculated and no relationship with VE was observed.



**Extended Data Fig. 4 | The relationship between VE and genetic distance on the NTD and S protein.** Panels (a-b): negative linear relationships between VE and genetic mismatch for NTD ( $P = 0.086$ ,  $R^2 = 75.8\%$ ), and full-length sequence ( $P = 0.082$ ,  $R^2 = 78.4\%$ ), respectively. The two-sided  $P$  was obtained from the mixed-effects model. The colored lines were fitted by data points of each platform. The shaded area indicates 95% CI.



**Extended Data Fig. 5 | Clustering of regions by circulating strains similarities to SARS-CoV-2 genetic variants.** Genetic mismatch of genetic variants to the local circulating virus during January and February 2022. The best candidate vaccine antigen for a geographical region measured by genetic distance is shown by dark red. Rows: target geographical regions; columns: candidate vaccine strains. The figure shows that the Omicron sublineages can match to the epidemic viruses in most regions, but the dominant sublineages are not the same.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- |                                     |                                     |  |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A description of all covariates tested   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data used in this study is publicly available. The detailed information of VE outcomes is available in the Supplementary Information. Viral sequence data were downloaded from the global initiative on sharing all influenza data (GISAID) at <http://platform.gisaid.org/> and the accession numbers are provided in the online Supplementary Acknowledgment Table (<https://github.com/VaccineEffectivenessPrediction/COVID19-Vaccine-Effectiveness>).

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The study extracted vaccine efficacy or vaccine effectiveness (VE) data before 24 Dec, 2021 from published articles and preprint articles. A total of 78 VE data were obtained for model building. All available sequences that matched to the period and locations of the clinical trials or observational studies totaled 1,984,241 full-length genome sequences from 31 geographical regions.
Data exclusions	For VE data, exclusion criteria include: target population has special conditions; the primary outcome is not symptomatic COVID-19 infection after the second vaccine dose; and the study period of VE evaluation is not reported. For sequence data, strains with duplicated names and unclear collection time of samples were removed.
Replication	This study demonstrated a clear relationship between COVID-19 VE and genetic distance on RBD, NTD and entire S protein. Our findings can be supported by biological experiments. We first collected data before June 2021 and determined genetic distance is associated with VE against symptomatic infection. After adding subsequent data before March 2022, the results are consistent with previous results. Such relationships exist in different vaccine platforms and vaccine products. The prediction results were validated by independent data. All attempts at replication were successful. Additionally, this bioinformatics framework has been applied to influenza A/H1N1pdm09, H3N2 and influenza B viruses and such a relationship was also detected.
Randomization	Randomization is not applicable in our study design. The vaccine efficacy outcomes included in this study were based upon clinical trials. The vaccine effectiveness outcomes were obtained from observational studies. All available sequences that matched to the period and locations of the clinical trials or observational studies were collected.
Blinding	Blinding is not relevant to the study. This study used population-level data and did not involve individual participants.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging