

1 **Title:**

2 **Rapid evaluation of COVID-19 vaccine effectiveness against VOC/VOIs by genetic mismatch**

3

4 **Authors:**

5 Lirong Cao^{1,2}, Jingzhi Lou³, Hong Zheng^{1,2}, Shi Zhao^{1,2}, Chris Ka Pun Mok¹, Renee Wan Yi Chan^{4,5}, Marc
6 Ka Chun Chong^{1,2}, Zigui Chen⁶, Eliza Lai Yi Wong¹, Paul Kay Sheung Chan⁶, Benny Chung-Ying Zee^{1,2}, Eng
7 Kiong Yeoh¹, and Maggie Haitian Wang^{1,2*}

8

9 **Affiliations:**

10 ¹ JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong SAR, China

11 ² CUHK Shenzhen Research Institute, Shenzhen, China

12 ³ Beth Bioinformatics Co. Ltd., Hong Kong SAR, China

13 ⁴ Department of Paediatrics, Chinese University of Hong Kong, Hong Kong SAR, China

14 ⁵ CUHK-UMCU Joint Research Laboratory of Respiratory Virus & Immunobiology, Chinese University of
15 Hong Kong, Hong Kong SAR, China

16 ⁶ Department of Microbiology, Stanley Ho Centre for Emerging Infectious Diseases, Li Ka Shing Institute of
17 Health Sciences, Chinese University of Hong Kong, Hong Kong SAR, China

18

19 **Correspondence:** Maggie Haitian Wang, email: maggiew@cuhk.edu.hk, JC School of Public Health and
20 Primary Care, Chinese University of Hong Kong, Hong Kong SAR, China.

21

22 **Abstract**

23 Timely evaluation of the protective effects of COVID-19 vaccines is challenging but urgently needed to
24 inform the pandemic control planning. Based on vaccine efficacy/effectiveness (VE) data of 11 vaccine
25 products and 297,055 SARS-CoV-2 sequences collected in 20 regions, we analyzed the relationship between
26 genetic mismatch of circulating viruses against the vaccine strain and VE. Variations from technology
27 platforms are controlled by a mixed-effects model. We found that the genetic mismatch measured on the RBD
28 is highly predictive for vaccine protection and accounted for 72.0% (p -value < 0.01) of the VE change. The
29 NTD and S protein also demonstrate significant but weaker per amino acid substitution association with VE
30 (p -values < 0.01). The model is applied to predict vaccine protection of existing vaccines against new genetic
31 variants and is validated by independent cohort studies. The estimated VE against the delta variant is 79.3%
32 (95% prediction interval: 67.0 – 92.1) using the mRNA platform, and an independent survey reported a close
33 match of 83.0%; against the beta variant (B.1.351) the predicted VE is 53.8% (95% prediction interval: 39.9 –
34 67.4) using the viral-vector vaccines, and an observational study reported a close match of 48.0%. Genetic
35 mismatch provides an accurate prediction for vaccine protection and offers a rapid evaluation method against
36 novel variants to facilitate vaccine deployment and public health responses.

37 **Keywords:** COVID-19; prediction; vaccine effectiveness; vaccine efficacy; genetic mismatch, sequence
38 analysis

39

40 **Main**

41 Vaccination is a crucial measure to control the transmission scale and mitigate the impact of COVID-19
42 infections. To date, 19 vaccines against SARS-CoV-2 are in early use or have been fully approved for
43 application in mass population¹. However, protective effect of the various vaccine products is under the
44 challenge of new genetic variants. Vaccine efficacy or effectiveness (VE) against COVID-19, which measures
45 the relative reduction of risk a disease outcome in clinical trials or mass population, exhibited a wide range of
46 variation from 10.4% to 97.2%²⁻⁵.

47 A number of reasons contribute to the variation in VE that makes it difficult to directly interpret and
48 inform the protective effect of vaccines. The notable factors include the technology platforms, the target
49 population, differences in study protocols, background risk of COVID-19 and time of study. The various
50 vaccine technology strategies generated non-identical immune correlates of protection for SARS-CoV-2
51 infection⁶. For instance, the LNP-mRNA vaccine (Moderna) induces S-specific IgG, high T_H1 cell responses
52 and low T_H2 cell responses^{7,8}, while the inactivated virus strategy (Sinovac) generates S, RBD and N-specific
53 IgG, without obvious T cell responses^{9,10}. Among all the influencing factors, emerging genetic variants
54 relative to the vaccine strain play a critical role in affecting vaccine effectiveness. Many investigations
55 showed that neutralizing activity in plasma or sera of vaccinated individuals against variants containing
56 E484K and N501Y mutations decreased significantly¹¹⁻¹³. Viral structure studies demonstrated that these
57 amino acid substitutions on the S protein may alter virus-host cell interactions and reshape antigenic surfaces
58 of the major neutralizing sites, thus leading to immune evasion^{14,15}. While the mechanisms of immune escape
59 caused by the new mutations are continuously being elucidated in experimental studies, an integrative
60 framework to quantify the effect of genetic mismatch on VE would be instrumental for efficient evaluation of
61 vaccine protection for any country in real-time. The genetic mismatch from vaccine strains due to evolution of
62 the circulating strains occurred in different time periods and locations could provide a relatively compact
63 approach to account for the spatial-temporal confounding factors for VE and facilitate the interpretation of
64 vaccine protective effect.

65 In this study, we drew the connection between genetic mismatch of circulating SARS-CoV-2 viruses
66 and reported COVID-19 VE from population studies. Based on previous bioinformatics approach established
67 for the influenza viruses^{16,17}, we further advanced the VE estimation framework for COVID-19 by controlling
68 the clustered random variation of technology platforms using a mixed-effects model. Through extensive
69 analysis of worldwide VE studies and genetic sequences, we showed that a significant proportion of the
70 change in VE could be explained by the genetic factor and provided an efficient framework to evaluate
71 vaccine protection.

72 Results

73 **VEs and genetic mismatch distributions by vaccine platforms** VE and genetic mismatch of the four
74 vaccine platforms are compared in **Figure 1**. Within each vaccine platform, the vaccine effectiveness is
75 generally lower compared to the efficacy outcome (**Figure 1a**); while in terms of genetic mismatch (**Figure**
76 **1b, Supplementary Figure S3.1**), the vaccine effectiveness cohort encompasses larger genetic mismatch
77 relative to the vaccine efficacy cohorts. This result indicates that genetic mismatch had increased during the
78 mass vaccination phase compared to the earlier clinical trial periods. Across the technology platforms, vaccine
79 protection (efficacy/effectiveness) shows significant difference (ANOVA p -value < 0.001, **Figure 1a**). The
80 mRNA vaccines reported the highest mean VE of 89.2% (95% CI: 86.2 – 92.2, N=18), followed by the
81 protein subunit vaccine 77.9% (range: 49.4 – 96.4, N=3), inactivated vaccine 72.3% (95% CI: 64.3 – 80.3,
82 N=8), and viral-vectored vaccines 66.7% (95% CI: 57.5 – 75.6, N=15). Interestingly, the genetic mismatch of
83 these platforms shows a perfect reverse trend, of which the mRNA vaccines cohorts correspond to the
84 smallest mismatch, and the viral-vector the highest. The genetic mismatch summarizes the deviation of
85 genetic variants with respect to the vaccine strains, accounting for time, location and multiple strain co-
86 circulation, for vaccine evaluation at population level using sequencing data.

87 **Relationship between vaccine protection and genetic mismatch** Next, we explored the statistical
88 relationship between vaccine protection and genetic mismatch. Using a mixed-effects model, at most 72.0%
89 of the variations in VE can be explained by the genetic mismatch measure, controlling for the random effect
90 of technology platforms (**Figure 2, Supplementary Table S3.2**). Among the candidate genomic regions,
91 genetic mismatch on the RBD region demonstrated the strongest influence on vaccine protection. For every
92 residue substitution on the RBD, the VE would reduce by an average of 7.2% (95% CI: 3.8 – 10.7, p -value <
93 0.001); the reduction of VE due to one mutation on the NTD and S-protein are 5.4% (95% CI: 2.8 – 7.9) and
94 1.6% (95% CI: 0.6 – 2.6), respectively (**Supplementary Table S3.2**), while mismatch on ORF1ab, ORF3a,
95 ORF8 and N proteins show no association to VE (**Supplementary Figures S3.3-3.4**). When no genetic
96 mismatch is present, VE is the highest for the mRNA vaccine of an expected level of 94.4% (95% CI: 91.2 –
97 97.7), estimated by the RBD region; and the inactivated and viral-vector vaccines show a systematically lower
98 VE by 16% and 18.6% relative to the mRNA vaccines.

99 **Independent validation and estimating VE against specific genetic variants** In **Figure 3a**, the predicted
100 and observed VEs for all independent datasets are summarized. Calibration plot (**Supplementary Figure S3.5**)
101 shows a close matching, and the concordance correlation coefficient reaches a high level of 0.96 (95% CI:
102 0.88 – 0.99). These validation results demonstrated feasibility of using genetic mismatch to estimate vaccine
103 performance. In **Figure 3b**, we further predicted VEs of the mRNA, inactivated and viral-vector vaccines for
104 15 different variants, including VOC and VOI based on the RBD mismatch (**Supplementary Table S1.3**).
105 Among these variants, four of them have observed VE reported while most of the rest variants have not been
106 surveyed for VE. Against the delta variant (B.1.617.2), the estimated VE is 79.3% (95% prediction interval:

107 67.0 – 92.1), 63.2% (95% prediction interval: 50.5 – 76.1) and 61.5% (95% prediction interval: 48.3 – 73.4)
108 for the mRNA, inactivated, and viral-vector vaccines, respectively (**Figure 3a**). These estimates are supported
109 by two independent epidemiological studies against the delta variant: the mRNA vaccine BNT162b2 and the
110 viral-vector vaccine ChAdOx1 provided 83% and 61% protection, respectively¹⁸; and the inactivated vaccine
111 BBV152 conferred 65.2% protection according to¹⁹. Furthermore, against the beta (B.1.351) and gamma (P.1)
112 variant, the estimated VE for viral-vector vaccines is 53.8% (95% prediction interval: 39.9 – 67.4) and 54.1%
113 (95% prediction interval: 40.0 – 67.7), respectively. An independent study of the viral-vector ChAdOx1-S
114 vaccine reported a VE of 48.0% against these variants²⁰.

115 **Depicting trend of VE from the genetic mismatch** VEs are predicted for the major vaccine platforms in
116 California at weekly intervals (**Figure 3c**). In general, an accelerating decreasing trend of VE in California is
117 depicted from the genetic mismatch. We showed that the model can be conveniently applied to track the
118 continuous change of VE. The observed VEs from clinical trials conducted during the period are overlaid on
119 the prediction outcomes for reference^{3,21-26}. During February and March 2021, the predicted VE is 86.3% (95%
120 prediction interval: 73.8 – 98.2) for the mRNA vaccines, and an independent survey in the US reported 91%
121 protection for the same vaccine platform²⁷.

122 **Discussion**

123 As novel variants of SARS-CoV-2 keep emerging in the ongoing pandemic, rapid assessment of vaccine
124 performance in populations is crucial to inform public health and clinical responses. This study established an
125 efficient computational framework to estimate COVID-19 VE using virus sequencing data. The predicted VEs
126 against the VOCs are close to outcomes of independent cohort studies. The framework has several advantages.
127 First, it enables prediction of VE against novel variants using existing virus surveillance network to derive a
128 rapid estimate, thus it could inform timely hospital resource allocation and preparedness. Second, it provides
129 an integrated measure to facilitate the interpretation of vaccine effects, which takes account of the
130 confounding effect of time and location related to genetic evolution. Third, through mixed-effects modelling,
131 the framework controls for the random effects in technology platforms, providing a consistent and adaptable
132 prediction framework for inclusion of multiple vaccine platforms.

133 Among the candidate genomic regions, the RBD and NTD regions exhibit the strongest statistical
134 association with VE. These findings are also supported by biological evidence. The RBD is the major target
135 for neutralizing antibodies that interfere with viral receptor binding^{28,29}, and the NTD is reported to be the
136 target of 5-20% of S-specific monoclonal antibodies from memory B cells against SARS-CoV-2^{30,31}.

137 Recent studies have investigated the use of the neutralization titer as a predictor of vaccine efficacy³²,
138 however the neutralizing results against COVID-19 genetic variants showed varying outcomes. The vaccine
139 protection against the B.1.351 variant reduced from 95.0%³ to 75.0%³³ by BNT162b2. Due to lack of

140 standardized neutralization assays and different protocols, one neutralization study showed that the titer
141 against the B.1.351 variant is 7.6- and 9-fold lower compared to the early Wuhan-related Victoria variant in
142 the BNT162b2 vaccine serum and ChAdOx1 vaccine serum, respectively¹²; while another experiment
143 reported a 2.7-fold decrease in neutralization titers against the B.1.351 strain in the BNT162b2-elicited
144 Serum³⁴. The varying neutralization results increase the challenge of inferring vaccine performance solely by
145 neutralization levels. In addition, the association of neutralization with protection across studies showed that
146 neutralizing antibodies might not be deterministic in mediating protection, and the effect of other vaccine-
147 induced immune responses also need to be quantified. This work uses an alternative angle to bridge the link
148 between molecular activities and population level vaccine responses. Further investigations are needed to
149 integrate potential correlates of vaccine protection and improve the existing framework.

150 The global pandemic of COVID-19 and virus evolution have caused regions in the world to
151 encompass diversified virus populations. We explored the possibility of developing region-specific vaccines
152 and how well they would match the circulating virus profiles. We investigated optimal candidate vaccine
153 strains for 13 regions, including the United Kingdom (UK), Germany, South Africa, Russia, India, Hong
154 Kong (HK), Malaysia, Japan, California, New York, Mexico, Peru and Brazil. Based on the genetic mismatch
155 between vaccine strains and observed viruses circulating in the region and period, hierarchical clustering of
156 the regions was performed to show the similarity of vaccine mismatches (**Figure 4**). We found that no single
157 strain can match to the epidemic viruses in all regions during March-April or May-Jun 2021. Particularly, for
158 the new Moderna vaccine mRNA-1273.351 adopting the B.1.351 variant³⁵, the mean genetic discrepancy to
159 local circulating strains is wider compared to either the Wuhan strain or the dominating region strains. This
160 result suggests that updating the vaccine compositions with a single genetic variant might not be sufficient. As
161 manufacturing of region-specific vaccines may not economically feasible, a reconciling strategy might be to
162 provide optimal vaccine candidates for country-clusters that share similar compositions of circulating viruses,
163 or to provide multivalent vaccines.

164 This study has several limitations. Although the current model reached good statistical significance,
165 the complexity of the model is restricted by the sample size of the available VE studies. Thus, population
166 characteristics and study design factors that may influence VE cannot be included. Secondly, the waned
167 immunity in host was not accounted for in the current model. Thus, the current estimate only suggests the
168 mean protection level within weeks since vaccination based on the data used for model training, and should be
169 interpreted with caution of potentially optimistic estimates. Further study will be sought to consider
170 penalization of the VE according to the time elapsed since last vaccination, as more longitudinal data of
171 immune correlates are available. Thirdly, bias might occur if sequences in databases disproportionately
172 represented regions with known circulation of a given variant. Enhanced efforts are needed to ensure better
173 geographical representativeness of available SARS-CoV-2 sequences. Despite these limitations, the
174 relationship of genetic mismatch and VE observed in multiple countries showed robust outcomes and were

175 validated by independent data. The framework could further pool VE outcomes by various manufacturers
176 using one additional layer of structured modelling, when enough data is available in the future.

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

181 **Methods**

182 **Vaccine Efficacy and Effectiveness Data** Vaccine efficacy is the relative proportion of vaccine protection
183 measured in clinical trials, and vaccine effectiveness is the quantity obtained from observational studies. Both
184 quantities are calculated by $(1-RR) \times 100$, where RR is the relative risk of a COVID-19 outcome in the
185 vaccine group compared to the placebo group. VE reports before May 17, 2021 were collected from journal
186 articles, documents of World Health Organization and Food and Drug Administration, and other government
187 reports. Studies that related to human subjects and contained clear investigation time were included. A total
188 number of 44 VEs were obtained for model building, among which 24 reported vaccine efficacy and 20
189 surveyed vaccine effectiveness. The vaccine efficacy studies contain thirteen Phase III trials, two Phase II
190 trials, and two Phase II/III trials. Inclusion criteria for the vaccine effectiveness studies are: target population
191 is all age group without special conditions; and the primary outcome is symptomatic COVID-19 infections or
192 confirmed infections requiring medical care. We also extracted 14 VEs from subsequent independent research
193 for validation study. The detailed information of VE for model building and validation is available in
194 **Supplementary Table S1.1-1.2.**

195 **Genetic Sequences** In the first part of analysis, relationship between genetic mismatch and VE was
196 modelled. Human SARS-CoV-2 strains with collection dates ranging from April 23, 2020 to May 16, 2021
197 were retrieved from the global initiative on sharing all influenza data (GISAID) EpiCoV database³⁶. All
198 available sequences that matched to the period and location of the clinical trials or observational studies were
199 downloaded. A total number of 297,055 full-length genome sequences were sampled from 20 geographical
200 regions for model development. For model validation, a total of 331,116 complete SARS-CoV-2 genome
201 sequences were retrieved from the GISAID.

202 The source of all SARS-CoV-2 sequences involved in this study was acknowledged in the **Supplementary**
203 **Acknowledgement Table**. Strains with duplicated names were removed. Multiple sequence alignment was
204 performed using MAFFT (version 7). The ‘Wuhan-Hu-1’ genome (GenBank ‘NC_045512.2’, or GISAID
205 ‘EPI_ISL_402125’) was set as the reference sequence. The variants involved in this study were summarized
206 in **Supplementary Tables S1.3-1.4**. Lineage classification for sequences was referenced from the GISAID.

207 **Statistical Methods** Following the previous framework developed for influenza virus¹⁶, let $X = \{x_{ij}\}$
208 denote the i -th sample from the GISAID database collected for a target population, where $i=1, \dots, n, j=1, \dots, J$;
209 and $V = \{v_j\}$ denote the vaccine strain applied in the target population, where index j indicates the j -th
210 codon position in the sequence. Denote the amino acids in a given genome region as $W = \{w_k\}$, where k is
211 the index for codon positions contained in the segment, $k = 1, \dots, K, 0 \leq K \leq J$. Suppose the Hamming
212 distance is used as a basic measure of dissimilarity between two sequences, the vaccine genetic mismatch
213 statistic (d) calculated for the target population is,

$$214 \quad d = \sum_{i=1}^n d_i / n = \sum_{i=1}^n \sum_{k=1}^K I(v_{w_k} \neq x_{w_k}) / n \quad . \quad \text{Eq. 1}$$

215 Thus, the d summarized the average amino acids mismatch of circulating strains versus the vaccine strain
216 based on a given genome segment in a target population. In this study, we considered a range of candidate W ,
217 including the receptor-binding domain (RBD), N-terminal domain (NTD), spike (S), nucleocapsid (N),
218 ORF1ab, ORF3 and ORF8 proteins. A schematic representation of SARS-CoV-2 genome and the structure of
219 the S protein are available in **Supplementary Figures S2.1-2.2**. All vaccine strains are based on the Wuhan
220 strain isolated in January 2020. When the target population is composed of subjects infected with multiple co-
221 circulating variants, the d captures the viral diversity in the cohort; while when the target population is
222 composed of subjects infects by a specific genetic variant, the mismatch measures variant-specific distance.

223 In view of the differences in vaccine platforms, a two-level mixed-effects model was adopted to account for
224 the random effect associated with technology platform. We specified the following random-intercept model
225 for a VE outcome (Y_{ij}) of technology j and study/trial i ,

$$226 \quad Y_{ij} = \beta_0 + \beta_1 \times d_{ij} + u_j + \varepsilon_{ij}, \quad \text{Eq. 2}$$

227 The fixed intercept parameter β_0 represents the expected value of VE when genetic mismatch is zero, that is,
228 the maximum protection of a vaccine. β_1 represents the fixed effect of genetic mismatch; and u_j denotes the
229 random effect associated with the intercept for platform j , which is assumed to follow a normal distribution
230 with zero mean and constant variance σ_p^2 . The ε_{ij} denotes the residual of observation from experiment i of
231 platform j , which follows a normal distribution of zero mean and constant variant σ^2 . The model was fitted
232 using R package lmerTest³⁷. The protein subunit vaccines were excluded in the mixed-effects model as the
233 sample size for this platform is only three. One VE (10.4%) of viral vector vaccine was considered as an
234 outlier and excluded, which is reported from a secondary analysis against the B.1.351 variant in a small-scale
235 South African trial². All analyses were performed using **R** statistical software (version 4.0.3). Statistical
236 significance was declared if p -value < 0.05 .

237 Validation study compared the estimated VE of a given platform by using specific lineage sequences or
238 sequences of circulating viruses in the respective regions and periods with fourteen VE outcomes pulling out

239 from independent observational studies. As an application example, we predicted VE against major variants of
240 concern (VOC) and variants of interest (VOI). We also estimated VE of all existing vaccines at weekly
241 intervals from July 20 2020 to July 19 2021 in California to depict the trend of VE through time. The
242 prediction interval of mixed-effects model was calculated using R package merTools³⁸.

243 **Data availability**

244 All data used in this study is publicly available. The detailed information of VE outcomes is available in the
245 supplementary materials. Viral sequence data were downloaded from the global initiative on sharing all
246 influenza data (GISAID) at <http://platform.gisaid.org/> and the accession numbers were provided in the
247 supplementary acknowledgment table.

248 **Code availability**

249 The code is available upon request from the corresponding author.

250 **Acknowledgements**

251 A complete GISAID acknowledgement table could be found in online supplementary materials. We thank the
252 contributions of all the health care workers and scientists, the GISAID team, and the submitting and the
253 originating laboratories. This work was partially supported by the Health and Medical Research Fund, the
254 Food and Health Bureau, the Government of the Hong Kong Special Administrative Region [COVID190103,
255 INF-CUHK-1], the National Natural Science Foundation of China [31871340, 71974165], and the Chinese
256 University of Hong Kong Grant [PIEF/Ph2/COVID/06, 4054600]. We appreciate the constructive comments
257 from reviewers that improved this work.

258 **Author Contributions**

259 M.H.W conceived the study, L.C and M.H.W wrote the manuscript. L.C and H.Z. collected data. L.C
260 processed data, carried out the analysis and wrote the first draft. J.L., S.Z., C.K.P.M, R.W.Y.C., M.K.C.C.,
261 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
262 approval for publication.

263 **Competing interests**

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

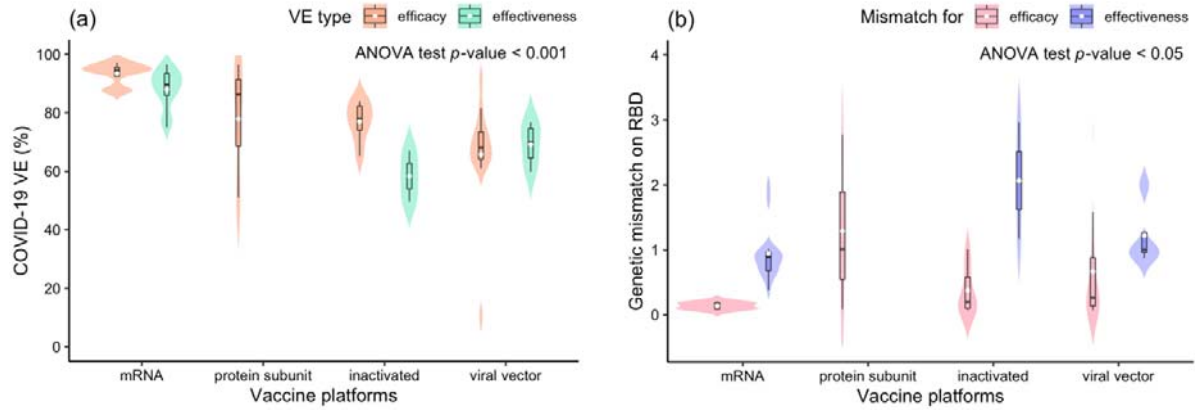
266

267 References

- 268 1. COVID-19 VACCINE TRACKER. Vol. 2021.
- 269 2. Madhi, S.A., *et al.* Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant.
270 *N Engl J Med* (2021).
- 271 3. Polack, F.P., *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med*
272 **383**, 2603-2615 (2020).
- 273 4. Hitchings, M.D.T., *et al.* Effectiveness of CoronaVac in the setting of high SARS-CoV-2 P.1 variant
274 transmission in Brazil: A test-negative case-control study. *medRxiv*, 2021.04.07.21255081 (2021).
- 275 5. Haas, E.J., *et al.* Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2
276 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination
277 campaign in Israel: an observational study using national surveillance data. *Lancet* **397**, 1819-1829
278 (2021).
- 279 6. Dai, L. & Gao, G.F. Viral targets for vaccines against COVID-19. *Nat Rev Immunol* **21**, 73-82 (2021).
- 280 7. Jackson, L.A., *et al.* An mRNA Vaccine against SARS-CoV-2 - Preliminary Report. *N Engl J Med*
281 **383**, 1920-1931 (2020).
- 282 8. Corbett, K.S., *et al.* Evaluation of the mRNA-1273 Vaccine against SARS-CoV-2 in Nonhuman
283 Primates. *N Engl J Med* **383**, 1544-1555 (2020).
- 284 9. Zhang, Y., *et al.* Safety, tolerability, and immunogenicity of an inactivated SARS-CoV-2 vaccine in
285 healthy adults aged 18-59 years: a randomised, double-blind, placebo-controlled, phase 1/2 clinical
286 trial. *Lancet Infect Dis* **21**, 181-192 (2021).
- 287 10. Wang, H., *et al.* Development of an Inactivated Vaccine Candidate, BBIBP-CorV, with Potent
288 Protection against SARS-CoV-2. *Cell* **182**, 713-721.e9 (2020).
- 289 11. Wang, Z., *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature*
290 (2021).
- 291 12. Zhou, D., *et al.* Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-
292 induced sera. *Cell* **184**, 2348-2361.e6 (2021).
- 293 13. Supasa, P., *et al.* Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine
294 sera. *Cell* **184**, 2201-2211.e7 (2021).
- 295 14. Cai, Y., *et al.* Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants.
296 *Science* **373**, 642-648 (2021).
- 297 15. Gobeil, S.M., *et al.* Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and
298 antigenicity. *Science* **373**(2021).
- 299 16. Cao, L., *et al.* In silico prediction of influenza vaccine effectiveness by sequence analysis. *Vaccine*
300 (2021).
- 301 17. Cao, L.R., *et al.* Differential Influence of Age on the Relationship between Genetic Mismatch and
302 A(H1N1)pdm09 Vaccine Effectiveness. *Viruses-Basel* **13**(2021).
- 303 18. Sheikh, A., McMenamin, J., Taylor, B. & Robertson, C. SARS-CoV-2 Delta VOC in Scotland:
304 demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **397**, 2461-2462 (2021).
- 305 19. Ella, R., *et al.* Efficacy, safety, and lot to lot immunogenicity of an inactivated SARS-CoV-2 vaccine
306 (BBV152): a, double-blind, randomised, controlled phase 3 trial. *medRxiv*, 2021.06.30.21259439
307 (2021).
- 308 20. Nasreen, S., *et al.* Effectiveness of COVID-19 vaccines against variants of concern, Canada. *medRxiv*,
309 2021.06.28.21259420 (2021).

- 310 21. Corchado-Garcia, J., *et al.* Real-world effectiveness of Ad26.COV2.S adenoviral vector vaccine for
311 COVID-19. *medRxiv*, 2021.04.27.21256193 (2021).
- 312 22. Andrejko, K., *et al.* Early evidence of COVID-19 vaccine effectiveness within the general population
313 of California. *medRxiv*, 2021.04.08.21255135 (2021).
- 314 23. Thompson, M.G., *et al.* Interim Estimates of Vaccine Effectiveness of BNT162b2 and mRNA-1273
315 COVID-19 Vaccines in Preventing SARS-CoV-2 Infection Among Health Care Personnel, First
316 Responders, and Other Essential and Frontline Workers - Eight U.S. Locations, December 2020-
317 March 2021. *MMWR Morb Mortal Wkly Rep* **70**, 495-500 (2021).
- 318 24. CDC. Largest CDC COVID-19 Vaccine Effectiveness Study in Health Workers Shows mRNA
319 Vaccines 94% Effective. (2021).
- 320 25. Sadoff, J., *et al.* Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. *N Engl*
321 *J Med* **384**, 2187-2201 (2021).
- 322 26. Baden, L.R., *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med* **384**,
323 403-416 (2021).
- 324 27. Kim, S.S., *et al.* mRNA Vaccine Effectiveness against COVID-19 among Symptomatic Outpatients
325 Aged ≥ 16 Years in the United States, February – May 2021. *medRxiv*, 2021.07.20.21260647 (2021).
- 326 28. Ju, B., *et al.* Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115-119
327 (2020).
- 328 29. Piccoli, L., *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike
329 Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183**, 1024-1042.e21
330 (2020).
- 331 30. McCallum, M., *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-
332 CoV-2. *Cell* **184**, 2332-2347.e16 (2021).
- 333 31. McCallum, M., *et al.* SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern.
334 *Science* **373**, 648-654 (2021).
- 335 32. Khoury, D.S., *et al.* Neutralizing antibody levels are highly predictive of immune protection from
336 symptomatic SARS-CoV-2 infection. *Nat Med* **27**, 1205-1211 (2021).
- 337 33. Abu-Raddad, L.J., Chemaitelly, H. & Butt, A.A. Effectiveness of the BNT162b2 Covid-19 Vaccine
338 against the B.1.1.7 and B.1.351 Variants. *N Engl J Med* **385**, 187-189 (2021).
- 339 34. Liu, Y., *et al.* Neutralizing Activity of BNT162b2-Elicited Serum. *N Engl J Med* **384**, 1466-1468
340 (2021).
- 341 35. Wu, K., *et al.* Variant SARS-CoV-2 mRNA vaccines confer broad neutralization as primary or
342 booster series in mice. *bioRxiv* (2021).
- 343 36. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to
344 reality. *Euro Surveill* **22**(2017).
- 345 37. Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H. lmerTest package: tests in linear mixed effects
346 models. *Journal of statistical software* **82**, 1-26 (2017).
- 347 38. Knowles, J.E., Frederick, C. & Knowles, M.J.E. Package ‘merTools’. (2016).

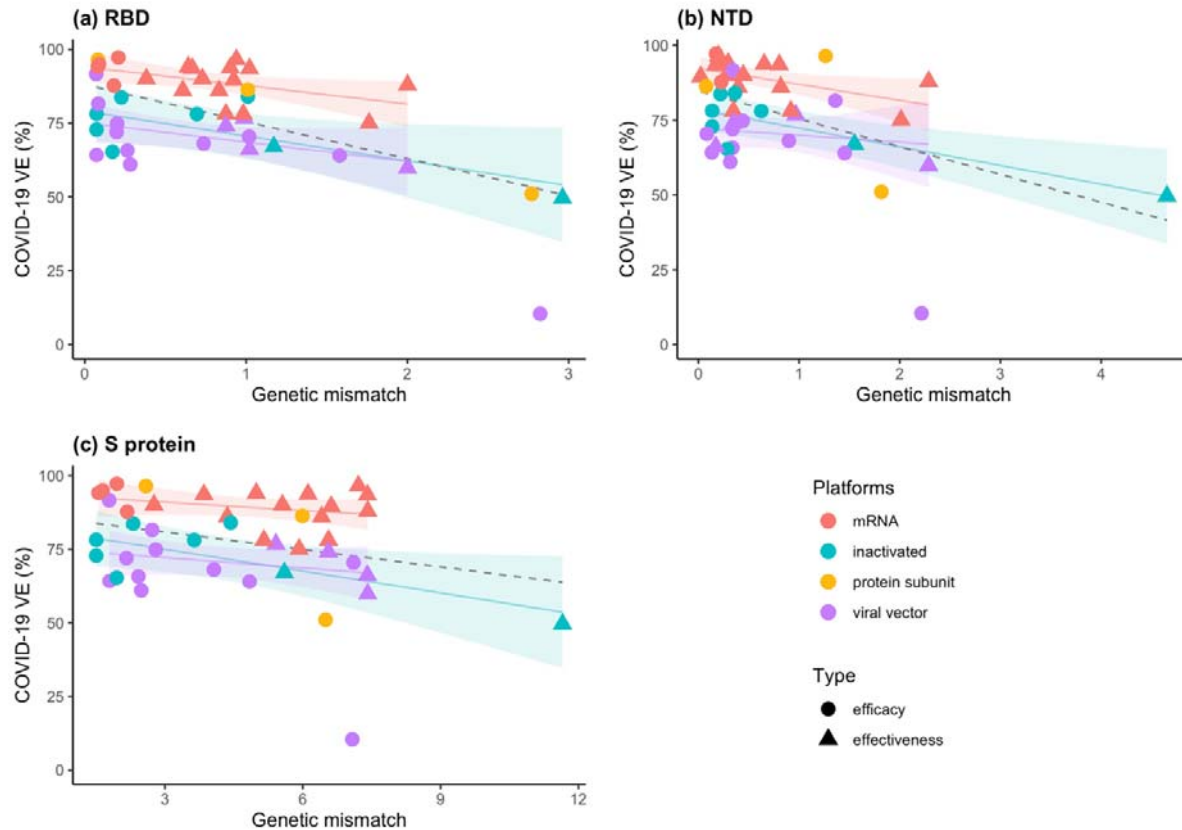
348



349

350 **Figure 1.** Comparison of COVID-19 vaccine efficacy and effectiveness (VE) and genetic mismatch across
351 vaccine platforms. Panel (a): distribution of the VE estimates for different platforms. The VE of mRNA
352 vaccines is higher than other vaccines (ANOVA p -value < 0.001). Panels (b): distribution of genetic mismatch
353 on RBD for different vaccine technologies. Genetic mismatch is the lowest for mRNA vaccines (ANOVA p -
354 value < 0.05).

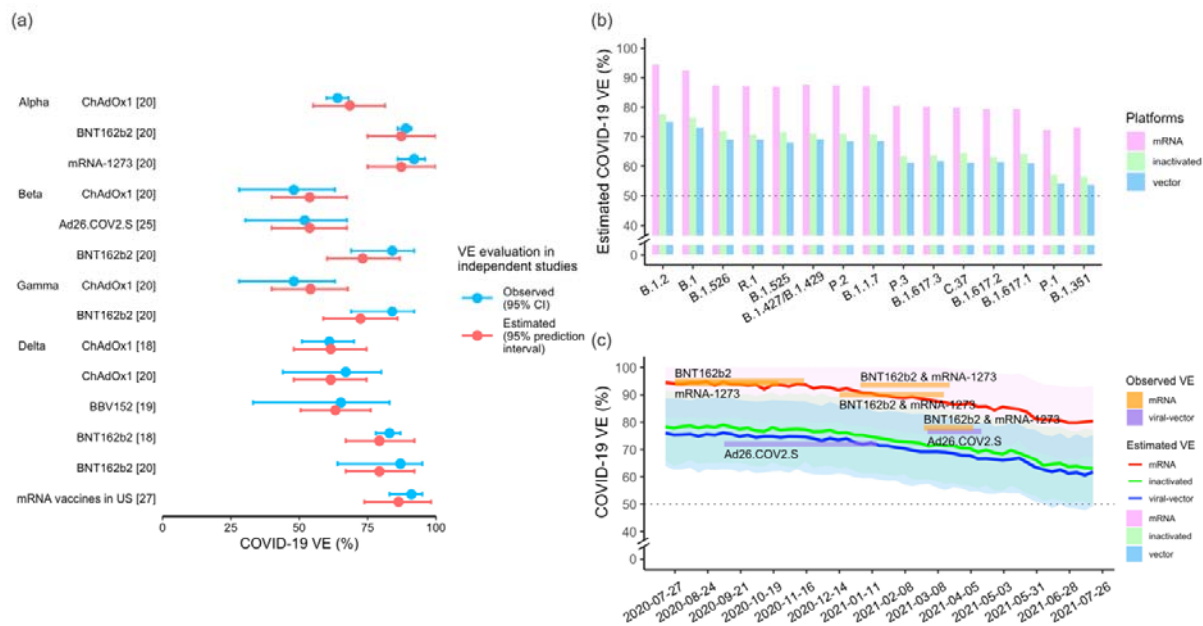
355



356

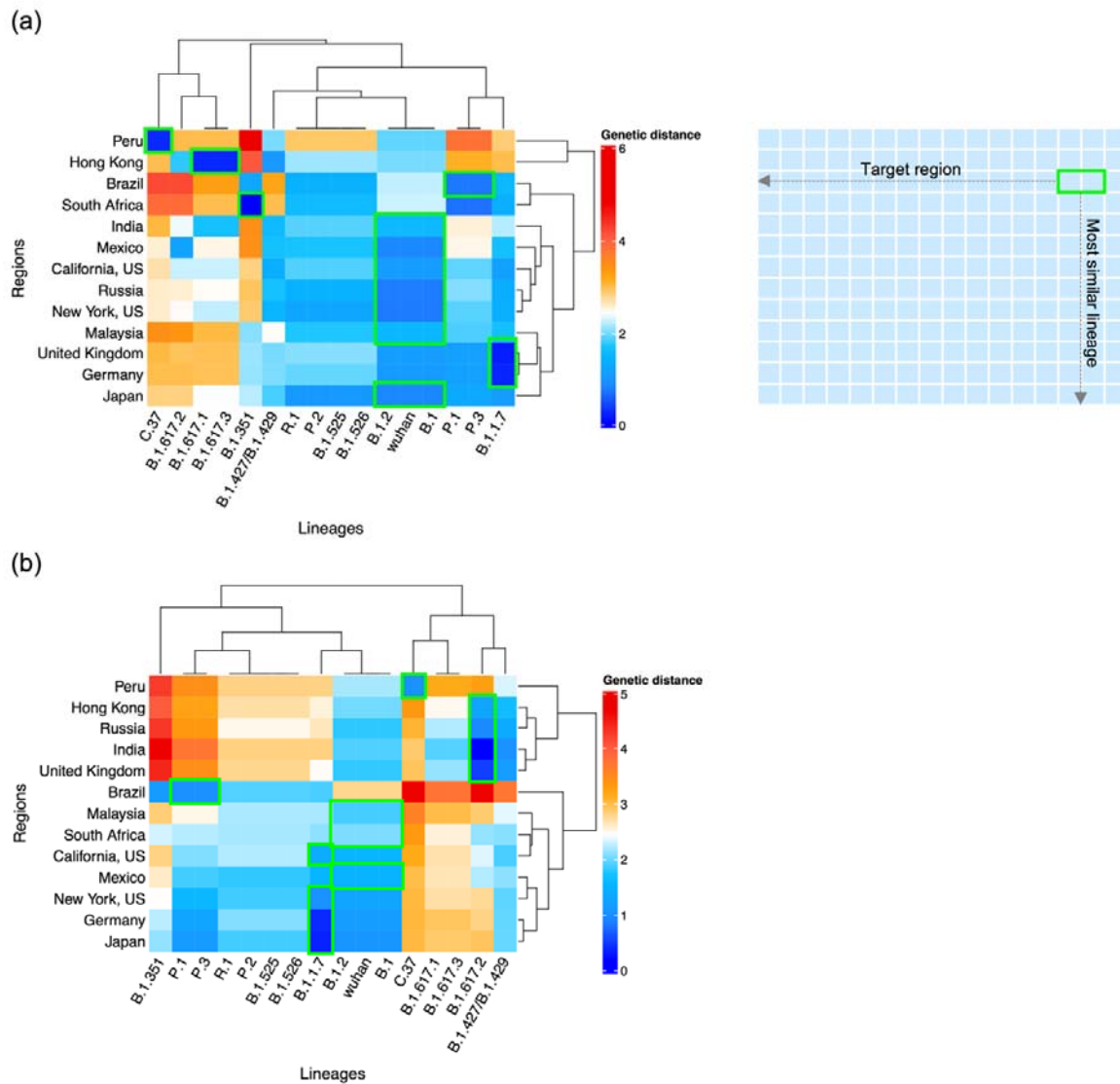
357 **Figure 2.** The relationship between VE and genetic mismatch of the circulating SARS-CoV-2 strains to the
358 vaccine strain on S protein. Panels (a-c): negative linear relationships between VE and genetic mismatch for
359 RBD (p -value < 0.01, R-sq = 72.0%), NTD (p -value < 0.001, R-sq = 68.8%), and full-length sequence (p -value
360 < 0.01, R-sq = 69.0%), respectively. The dashed line was fitted by all data points. The colored lines were fitted
361 by data points of each platform. The shaded area indicates 95% confidence interval.

362



363

364 **Figure 3.** Prediction of the VE based on the genetic distance. Panel (a): Validation outcome of estimated VE
 365 and observed VE in independent datasets. The predicted VEs against VOC are close to outcomes of cohort
 366 studies observations with concordance correlation coefficient 0.96 (95% CI: 0.88 – 0.99) (Supplementary
 367 Figure S3.5). Panel (b): Estimation of the variant-specific VE for mRNA vaccines (pink bar), inactivated
 368 vaccines (green bar) and viral-vector vaccines (blue bar). Panel (c): VEs in California were predicted at
 369 weekly intervals for different vaccine platforms. The plot indicates that the VE is declining at an accelerating
 370 speed. The surveyed VE from clinical trials or observational studies during the same period are overlaid on
 371 the trend curve as colored rectangles for reference, and only the mRNA and viral-vector platform vaccines are
 372 available. The shaded areas are 95% prediction interval. The dashed line marks the 50% efficacy threshold.



373

374 **Figure 4.** Clustering of regions by circulating strains similarities to VOC/VOIs. Panel (a): Genetic mismatch
 375 of genetic variants to the local circulating virus during March and April 2021. The best candidate vaccine
 376 antigen for a region measured by genetic distance is highlighted by a green box. Rows: target regions;
 377 Columns: candidate vaccine antigens (VOC/VOIs). Panel (b): Genetic mismatch during May and June 2021.
 378 For example, in Panel (b), the dark blue of B.1.1.7 in Japan means that the average genetic mismatch between
 379 the circulating viruses to the B.1.1.7 is lowest compared to using other variants as vaccine strains, suggesting
 380 that the B.1.1.7 is the most optimal vaccine antigen in Japan during May-Jun 2021. The figure shows that no
 381 single strain can match to the epidemic viruses in all regions, and the solution might be to provide optimal
 382 vaccine candidates for country-clusters that share similar compositions of circulating viruses, or to develop
 383 multivalent vaccines.