Evaluating the Effect of SARS-CoV-2 Spike Mutations by Causal Inference

Xin Wang  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Mingda Hu  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Bo Liu  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Huifang Xu  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Yuan Jin  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Boqian Wang  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Yunxiang Zhao  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Jun Wu  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Junjie Yue  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Hongguang Ren (✉ bioren@163.com)  
Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences (AMMS)

Article
Abstract

Driven by various mutations on the viral Spike protein, diverse variants of SARS-CoV-2 have emerged and prevailed repeatedly, which necessitates the identification of key Spike mutations for fitness enhancement. To address the need, this manuscript formulates a principled framework of causal inference for evaluating Spike mutations. In the context of large-scale genomes of SARS-CoV-2, it estimates the contribution of mutations to viral fitness across lineages and validates mutational effects on the Spike stability, receptor-binding affinity, and potential for immune escape. Key fitness-enhancing mutations and protein regions are recognized and studied. The transmission capacity of any new variant possessing these mutations can be predicted based on our model, solely based on the viral sequence. This research produces an innovative and systematic insight into SARS-CoV-2 and promotes functional studies of its key mutations.

Main Text

As of Oct 2022, the coronavirus disease 2019 (COVID-19) pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1), has been ongoing for nearly three years, resulting in over 624 million infections with 6.6 million deaths worldwide (https://covid19.who.int/). As a paramount characteristic of SARS-CoV-2, diverse variants have emerged and prevailed repeatedly, driven by numerous mutations, especially on the viral Spike protein (2, 3). Emerging variants of SARS-CoV-2 have substantially prolonged the pandemic by frequently repeated epidemics, which continually threatens public health across the world (4).

During the pandemic, the Spike protein of SARS-CoV-2 has attracted particular attention because it functionally mediates viral entry into host cells (5), and is the target of antibody-mediated immunity (6–8). Meanwhile, various mutations have accumulated in the Spike protein, including the receptor-binding domain (RBD, amino acid position 319–541), and those mutations may enhance viral fitness and give rise to new variants (3). For instance, the D614G mutation can increase viral infectivity (9–11) and has been found in almost all the following VoCs (Variant of Concern). Therefore, identifying key Spike mutations that likely elevate viral fitness is of vital importance for the research on SARS-CoV-2.

Up until now, millions of genome sequences of SARS-CoV-2 have been submitted and shared globally (12), making it feasible for computing analysis on viral mutations. As a novel computing method, causal inference enjoys broad prospects for applications (13, 14). It produces an unbiased estimation of the effect of a given intervention with confounders (13–15), and is considerably applicable to mutational analysis on SARS-CoV-2 in which mutations act as confounding factors for each other. Benetted from causal inference, Spike mutations can be evaluated according to the statistical contributions to viral fitness, in the context of large-scale genomes of SARS-CoV-2. Subsequently, key fitness-enhancing mutations can be distinguished from numerous mutations, and further applied to downstream analysis.
This manuscript formulates a principled framework, which utilizes causal inference to estimate the statistical contribution of Spike mutations to viral fitness across lineages. To the best of our knowledge, this research is the first to apply causal inference to mutational analysis on large-scale genomes of SARS-CoV-2. This work, schematically depicted in Fig. 1, is described in detail in Methodology, including the Data Preprocessing, the Effect Estimation, the Validation and Application, etc. In the Data Preprocessing, 7.7 million high-quality SARS-CoV-2 complete genome sequences as of May 11, 2022 are retrieved from GISAID website (12), aligned for Spike amino acid mutations, and mapped into mutation combinations with the corresponding basic reproduction number (R0), as row vectors in the feature matrix. In the Effect Estimation, causal inference is utilized for an unbiased estimation of the average treatment effect (ATE) of each mutation on the outcome R0, and the estimated ATE serves as the effect score of mutations. Mutations are computationally validated for functional influences, including the Spike protein stability, the host cell-surface receptor (the human angiotensin-converting enzyme 2, ACE2) binding affinity, and the potential for immune escape. Based on effect scores as the quantitative assessment of mutations, important mutations and protein regions can be recognized and studied. Besides, the fitness of SARS-CoV-2 variants is estimated, and the transmission capacity of any new variant can be predicted solely based on the viral sequence. Moreover, secondary results of causal inference models can likewise assist further analysis, which may reveal potential interactions between mutations.

Results

Effect Score And Validation Of Mutations

Based on 7.7 million genomes, 107 mostly frequent amino acid mutations on the Spike protein are identified. Meanwhile, the basic reproduction number (R0) of each sequence is quantified according to the corresponding variant type (16, 17). By causal inference methods, mutations are evaluated on their contribution to R0 across lineages, namely the effect score. As validations, mutations are studied by computing methods for detailed influences, including the Spike stability, ACE2 binding affinity, and immune escape. Based on effect scores as the quantitative assessment of mutations, important mutations and protein regions can be recognized and studied. Besides, the fitness of SARS-CoV-2 variants is estimated, and the transmission capacity of any new variant can be predicted solely based on the viral sequence. Moreover, secondary results of causal inference models can likewise assist further analysis, which may reveal potential interactions between mutations.

Effect score of the top twenty mutations, with computing validations and references for mutational effects, is presented in Table 1. All mutations in Table 1, except P26- mutation, possess one or more validated positive functional influences, supported by either computing results or literature references, or both. For instance, the T478K mutation may stabilize the Spike protein and can significantly enhance the binding affinity between Spike and ACE2 (18, 19). The D614G mutation, which has been found in VoCs since early 2020, may be involved in the Spike stability, viral replication, and Spike conformation shifting, thus improving viral infectivity and transmissibility (9–11). As a key mutation in BA.2.12.1 strains (20), S704L is another high-scoring mutation. Computing results have demonstrated its contributive effect in all three perspectives, indicative of a possibly compound effect of S704L.
For a comparative study, the bottom twenty mutations are presented in Supplementary Table 1. In contrast to Table 1, mutations in Supplementary Table 1 only possess one or no significant positive influence. Furthermore, six mutations in Supplementary Table 1 have no significant positive effect, by either computing analysis or literature references. Consequently, top mutations can be more contributive than bottom ones in our ranking results, demonstrating the effectiveness of effect scores.

In terms of related VoCs, most mutations in Table 1 are typical mutations for VoC, except the V213-mutation. On the contrary, over half of the mutations in Supplementary Table 1 are not typical for VoCs. Besides, most mutations in Table 1 have been found in Omicron strains, except for three mutations (V213-, E484K, and A222V). Accordingly, mutations of VoCs, especially those for Omicron variants, usually have high effect scores due to their contributions.

This work further studies mutant RBD proteins by biological experiments. We designed two new RBD sequences (RBD-1 and RBD-2, c.f. Supplementary Table 2) with several key positions replaced by mutations evaluated by our model to be with possible fitness enhancements. Mutant RBDs are expressed, purified, and evaluated for ACE2 binding affinity, compared with the wildtype RBD (RBD-WT). After the successful expression and purification illustrated in Supplementary Fig. 1A-1C, the ACE2 affinity is estimated and presented in Supplementary Table 3, with detailed binding kinetics shown in Supplementary Fig. 1D. Mutant RBDs, especially RBD-1, have stronger affinity to ACE2 than RBD-WT, and thus such mutation combinations are demonstrated to be contributive to the enhancement of RBD-ACE2 binding affinity.

The abovementioned result with validations and experiments demonstrates the effectiveness of this study and the feasibility of further analysis. The effect score of all 107 mutations is provided as a supplementary file.

**Key Mutation Identification On The Spike Protein**

Based on the effect score of mutations, key fitness-increasing mutations can be recognized. Supplementary Fig. 2 illustrates the treemap of Spike mutations by subunits, in which the size represents the effect score, and the color represents the count of mutational occurrences. Generally, the overall size of the S1 subunit is considerably larger than S2, suggesting that the S1 subunit may be more contributive to viral fitness elevation. Moreover, the effect score of mutations is not necessarily correlated with the mutation count. On the one hand, some long-accumulated mutations, such as D614G and T478K, play a significant role in the enhancement of viral fitness. On the other hand, some recently emerged mutations, e.g., V213- and S704L, can still achieve high effect scores by contributions, despite fewer occurrences.

For a structural study on high-scoring mutations, residues of the top ten mutations are visualized in the Spike-ACE2 complex in Fig. 2. Spike structure in the closed conformation (i.e., receptor inaccessible state) is also visualized in Supplementary Fig. 3 for a comparative study. In Fig. 2, four mutations (S477N, T478K, Q498R, and N501Y), occur in the binding interface between Spike and ACE2. Such mutations
locate in the receptor-binding domain (RBD, amino acid position 319–541), possibly engaged in the Spike-ACE2 affinity. Table 1 and references support that these mutations can increase binding affinity (18, 19, 21–23). S371F in the RBD domain of the Spike is reported to increase the Spike stability and ACE2 affinity, and is also involved in immune escape (23, 24). Furthermore, S371 residue may participate in the conformational transition of Spike between the open state (Fig. 2) and closed state (Supplementary Fig. 3), namely the up and down positions of RBD, respectively (25). Two mutations, P26- and V213-, occur in the N-terminal domain (NTD, amino acid position 14–303). NTD can be the target of human monoclonal antibodies (mAbs) (6, 26), hence these mutations may play a part in the immune escape of SARS-CoV-2 (6, 26). For the D614G mutation, besides its influence on Spike stability and viral replications (9, 10), it can participate in the Spike conformation shift toward an ACE2 binding-competent state, before viral membrane fusion with host cells (5, 11). In the subdomain linking S1 to S2, the H655Y mutation gives rise to a less tight loop that wraps the furin cleavage finger, and increases infectivity in the presence of N501Y (27). In terms of S704L, despite the lack of references for functional effects, computing validations in Table 1 verifies its effect on the Spike stability, ACE2 affinity, and immune escape.

Regional Study Of The Spike Protein

The Spike protein of SARS-CoV-2 consists of two subunits (Supplementary Fig. 3), S1 and S2, divided by the furin cleavage site (amino acid position 681–685) (3). S1 mainly includes NTD and RBD, and mediates the ACE2 binding to host cells; S2 functionally conducts the membrane fusion with host cells (3, 5). In combination with the regional functions, evaluated mutations can produce a better insight into Spike subunits and domains.

Mutations of positive effect scores are illustrated across the Spike gene in the Manhattan plot in Fig. 3. Among the top twenty mutations, eighteen mutations clustered in the S1 subunit. Hence, the S1 subunit achieves a greater mutational contribution, compared with S2. Specifically, nine mutations occur in RBD, including the top-scoring T478K mutation. Consequently, RBD mutations are vastly important to fitness enhancement, which can be explained by its function of ACE2 binding and immune escape (3, 5, 8). NTD likewise plays a part in viral infection, and contains six high-scoring mutations. Some important mutations, such as H655Y and P681H, locate near the S1-S2 subunit boundary, which may be related to the furin cleavage site (27) and facilitate the conformational shift of Spike (3). Compared with the S1 subunit, mutations in S2 usually have modest effect scores, except for S704L and N764K.

The distribution of effect scores in different subunits/domains is presented in Supplementary Fig. 4, including both positive and negative scores. Despite the approximate length of S1 and S2, the S1 subunit has considerably more mutations, especially high-scoring ones. Thus, S1 is more contributive in fitness elevation. On the contrary, the S2 subunit can be more conserved, with fewer mutations compared with S1 (7). Among the 81 mutations in S1, 46 mutations are concentrated in NTD. Nevertheless, most scores for mutation in NTD are modest. Compared with other regions, RBD generally has a higher distribution of
effect scores. Due to its crucial function, RBD serves as an important domain in the fitness enhancement of SARS-CoV-2.

**Viral Fitness And R0 Prediction**

Based on mutations, the fitness score of SARS-CoV-2 strains can be estimated. The fitness score of one given sequence can be defined by the sum of effect scores of its mutations. Particularly, the original Wuhan strain (wildtype) serves as a baseline for fitness score evaluation.

The fitness score and R0 of wildtype and VoCs is illustrated in Supplementary Fig. 5. The rank by the fitness score is consistent with the rank by R0, indicating the correlation between the two sides. Further, those strains are plotted as points in Fig. 4. All points except subsequent BA.2.12.1 are used to train a polynomial regression (with degree 3). These points are generally close to the regression, located in the 75% confidential interval (75% CI). The regression line after BA.2 is prediction other than training. As a validation, the value of BA.2.12.1 is subsequently plotted, which is close to the regression and well fits the prediction.

Further, the historical fitness score of SARS-CoV-2 is explored. Because of numerous SARS-CoV-2 strains during each period, the historical fitness score is extended to be the overall fitness during a specific period, i.e., the weighted sum of effect scores, with the weight being the mutation frequency during that time period (Fig. 5). A steady increase of the historical viral fitness score during the pandemic has been illustrated in Fig. 5, whose rises synchronize well with the contemporaneous emergence of VoCs. For instance, the D614G strain rose to prominence in Feb 2020 (9), leading to the fitness increase at that time. Similar increases can be found when the Alpha, Delta, and Omicron variant emerged respectively. Over time, the viral fitness increase is accelerated, especially by the emergence of Delta and Omicron. As for Spike regions, the contribution of RBD increases significantly, from the minority in 2020 to the majority since mid-2021. Similarly, the contribution of NTD and other regions increases, though the promotion is less than RBD.

**Discussion**

The Spike protein of SARS-CoV-2 is vastly important to viral fitness, especially transmissibility (3, 5, 8), and is one of the most investigated proteins of SARS-CoV-2. This manuscript concentrates on Spike mutations for mutational contributions to viral fitness. Despite studies on individual mutations such as D614G (9–11) and N501Y (18, 19, 21), extensive assessment of mutational contributions in the context of large-scale genomes is still challenging. Meanwhile, causal inference is one of the most promising methods in data science. It produces an unbiased estimation of the treatment effect on outcomes, as a function of observable characteristics of samples (13–15, 28). Benefitted from causal inference, Spike mutations can be evaluated and identified for further analysis.
To employ causal inference models, a quantitative phenotype is required for representing viral fitness as the outcome variable, which is set to be R0 in this paper. On the one hand, R0 clearly reveals viral transmissibility, as one of the most important quantitative phenotypes for viral fitness; on the other hand, the R0 estimation of SARS-CoV-2 variants is extensively studied and widely recognized (16, 17). Therefore, R0 is considerably representative and qualified for viral fitness. Likewise, other phenotypes of SARS-CoV-2 are qualified for causal inference models as long as quantitated. Hence, this work can be transferred and applied to other quantitative phenotypes, and even to other viral genomic data.

The validation of this study includes computing methods, whose results are consistent with literature references. This consistency provides a solid cornerstone for the validation and explanation of this research and further demonstrates the effectiveness of effect scores.

Although this paper includes three important mutational effects, there are still other aspects of effects, such as viral replication and viral pathogenicity. Thus, a mutation with no validated influence in Table 1 does not necessarily imply no positive effect at all. Accordingly, literature references also serve as supplementary validations. For instance, the R408S mutation shows no remarkable influence by computing validations in Table 1 but may facilitate the opening of RBD by previous research (29).

Another noticeable mutation is S704L, which is positively influential in Table 1 by computing results. To date, S704L has not been thoroughly investigated, which may be an interesting direction for future research.

Another interesting phenomenon is that the effect of Spike mutations can be correlated with the region. For instance, RBD functionally conducts ACE2 binding (5) and can be the target of antibodies (8), hence many affinity-enhancing and immune-escape mutations occur in RBD, making RBD a region of high-scoring mutations. Another domain, NTD, has maximum mutations but mostly with modest average scores, which may be not vastly noticeable. Although some mutations in NTD, e.g., T95I, V213G, and A222V, may be involved with immune escape (30, 31), the specific function of NTD and its mutations still remain to be elucidated. Besides, some important mutations locate at/near the furin cleavage site. As the furin cleavage site is crucial to SARS-CoV-2 (3, 32), those mutations may functionally correlate with it, like P681R (33) and H655Y (27).

Particularly, the S1 and S2 subunits can be rather different in immune escape. As shown in Fig. 3, high-scoring mutations are concentrated in S1, indicating that S1 can be more mutational. The S1 subunit is usually the target of antibodies (6, 8, 26) and holds many mutations, so S1 may be vastly important to viral immune escape. On the contrary, S2 is more conserved than S1, with fewer mutations. This characteristic may make S2 a potential target of medicine and general vaccine development (7).

In the present study, contributions of mutations are learned from R0, and conversely, reveal the relative viral fitness. During the model training, however, BA.2.12.1 strains are not particularly distinguished but recognized as ordinary BA.2 strains. Even so, BA.2.12.1 still achieves a higher fitness score than BA.2, hence the fitness score can effectively reveal viral fitness. Similarly, the regression in Fig. 4 is not trained by BA.2.12.1 strains, but BA.2.12.1 well fits the regression, which validates the effectiveness of the model.
Therefore, for an unknown strain of SARS-CoV-2, its relative fitness, as well as $R_0$, can be computationally predicted, solely based on its viral sequence, by computing its fitness score by mutations and further estimating $R_0$ according to the regression.

In terms of the historical fitness in Fig. 5, one interesting phenomenon is the synchronization between the fitness increase and the emergence of variants. The driving forces behind may be related to the selective sweep of SARS-CoV-2 (34), in which previous predominant strains are swept and replaced by new ones. The prevalence of novel variants usually implies possible adaptive advantages compared with previous ones, leading to a higher fitness score. According to the upward trend of the regression line in Fig. 4, new strains with higher scores may enjoy significantly enhanced virological fitness such as transmissibility and immune escape, which may further prolong the pandemic. Accordingly, epidemiological surveillance of new SARS-CoV-2 variants is supposed to be strengthened.

The interpretability of causal inference model can produce secondary information. The model of each mutation includes other mutations as covariates, which can be interpreted by SHAP values (35). Supplementary Fig. 6 interprets the model of the top mutations, which represents the unidirectional influence of other mutations on the object mutation. For mutations that are mutually top influential to each other, their bidirectional influences may reveal possible mutation interactions. These interactions are categorized into positive and negative ones, according to the mutational co-occurrence and exclusion, respectively. Supplementary Fig. 7 illustrates the possible interactions discovered in this study, which remains to be further investigated and confirmed.

**Methodology**

This work is schematically depicted in Fig. 1, consisting of Data Preprocessing, Effect Estimation, Validation and Application, etc.

**Data Preprocessing**

**Data source and data curation.** From GISAID Website (12), 7,699,174 high-quality SARS-CoV-2 complete genome sequences as of 11 May 2022 are retrieved (c.f. Supplementary_FastaID.csv for more information). For each genome, amino acid mutations on the Spike gene are identified in alignment with the reference sequence Wuhan-Hu-1 (GenBank accession number NC_045512). Spike mutations with global occurrence of less than 1,500 are considered infrequent and then discarded.

**Feature extraction.** The feature matrix is generated to depict Spike sequences by the mutation combination and $R_0$. Each Spike sequence is mapped into a mutation combination, and then is represented by a bool vector of Spike mutations, as a row vector in the feature matrix, along with $R_0$ according to the variant type (16, 17). Mutation combinations with either global occurrence less than 1500 or inestimable $R_0$ are abandoned. Since replicate rows in the feature matrix have no influence on the unbiased estimation of causal inference, redundant rows are merged. For a sound estimation,
mutations are supposed to be observed in at least two mutation combinations, otherwise will be discarded. Consequently, 107 major amino acid Spike mutations are retained for further studies.

**Effect Estimation**

Mutations are modelled and evaluated successively. For each mutation, a causal inference model is utilized for an estimation of its effect across lineages. For a given mutation $M_i$ as treatment $T$, its effect score is estimated by $\theta_T$, namely the average treatment effect (ATE) on outcomes, with $R_0$ as the observed outcome $Y$ and other mutations as observable characteristics (covariates) $X$ on samples. The estimation is based on all the observed i.i.d. samples from the feature matrix, with the $j$-th row being the sample $(Y_j, T_j, X_j)$. Specifically, Linear Doubly Robust Learner (Linear DRL) \( (36, 37) \) is employed, with an assumption of linear form of treatment effect \( (36, 37) \).

\[
\theta_T = E \left[ Y^{(T=1)} - Y^{(T=0)} \mid X \right]
\]

**Validation**

For validations of effect scores, computing methods are utilized for functional influences of mutations, including the Spike protein stability, ACE2 binding affinity, and potential for immune escape.

**Spike protein stability.** FoldX5 \( (38) \) is performed to estimate mutational effects on the stability of Spike protein in closed conformation (PDB: 7DDD) \( (39) \). Specifically, FoldX5 evaluates quantitative changes in the Gibbs energy of protein folding caused by mutations (\( \Delta \Delta G \), unit: kcal/mol) \( (38) \). Mutation effects on \( \Delta \Delta G \) include highly positive effect \( (\Delta \Delta G < -1.0) \), potential positive effect \( (-1.0 < \Delta \Delta G < 0) \), and no significant positive effect \( (\Delta \Delta G > 0) \). Indels (insertions and deletions) and mutations at unmodeled residues of the protein are inapplicable to FoldX5 and are labelled as NA (not applicable).

**ACE2 binding affinity.** FoldX5 \( (38) \) is performed to estimate mutational effects on the Spike-ACE2 complex (PDB: 7A94) \( (40, 41) \). FoldX5 evaluates quantitative changes in the Gibbs energy caused by mutations (\( \Delta \Delta G \), unit: kcal/mol) \( (38) \). Mutation effects on \( \Delta \Delta G \) include highly positive effect \( (\Delta \Delta G < -1.0) \), potential positive effect \( (-1.0 < \Delta \Delta G < 0) \), and no significant positive effect \( (\Delta \Delta G > 0) \). Indels (insertions and deletions) and mutations at unmodeled residues of the protein are inapplicable to FoldX5 and are labelled as NA (not applicable).

**Immune escape.** A system named Constrained Semantic Change Search (CSCS) \( (42) \) is utilized to estimate semantic changes (\( \Delta s \)) of SARS-CoV-2 Spike sequence for the potential for immune escape caused by mutations \( (42) \). Mutation effects on semantic changes include highly positive effect \( (\Delta s > 0.9) \), potential positive effect \( (0.75 < \Delta s < 0.9) \), and no significant positive effect \( (\Delta s < 0.75) \).

**Rbd-ace2 Affinity**
For biological experimental investigations, mutant RBD proteins based on RBD mutations in Table 1 are expressed, purified, and then evaluated for ACE2 binding affinity, compared with the wildtype RBD.

**RBD Expression and Purification.** The wildtype and mutant RBD proteins are expressed and purified by the method in prior works (43). Mutant RBDs are designed by mutation sites into pPICZαA-RBD-WT, according to the mutation in Supplementary Table 2. The plasmids of pPICZαA-RBD are linearized by BglII and transformed into the glycoengineered yeast (44). Positive clones of RBD are screened by western blot analysis. After the shake-flask culture, the product is centrifuged at 8500× g rpm for 15 min. The harvested supernatant is purified as described previously (43). Purified proteins are analyzed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE).

**RBD-ACE2 Affinity.** The binding kinetics of RBDs to His-tagged human angiotensin-converting enzyme 2 (ACE2) is assayed and evaluated by the ForteBio Octet™ QKe System (Pall ForteBio Corporation) (45). RBDs and ACE2 are diluted to 400 nM with HBS-EP (Cytiva), and an additional well with only HBS-EP is set up as a control. ACE2 is bound to the probe capturing the His tag. After the stabilization of RBD-ACE2 binding, the dissociation is performed in HBS-EP. The dissociation constant (Kd) is calculated by Data Analysis Software 7.0 (Pall ForteBio Corporation).

**Conclusions**

This manuscript proposes and formulates a principled framework of an unbiased approach for evaluating Spike mutations of SARS-CoV-2 by causal inference models, in the context of large-scale genomes. Based on 7.7 million viral genome sequences, mutations are evaluated on their contributions to viral fitness across lineages. As validated, high-scoring mutations possess one or more positive functional influences, which demonstrates the effectiveness of this research. Based on the effect score, key fitness-enhancing mutations and protein regions are identified. Particularly, RBD mutations play an important role in the fitness elevation of SARS-CoV-2. Besides, the fitness and R0 of unknown SARS-CoV-2 strains can be predicted, solely based on the viral sequence. This approach provides reliable guidance about mutations of interest, including some high-scoring but less-studied mutations like S704L. Moreover, the present work can be transferred to other quantitative phenotypes of SARS-CoV-2 for evaluating specific mutational effects, e.g., immune escape. Altogether, this approach produces an innovative and systematic insight into SARS-CoV-2 mutations, which may contribute to the evolutionary characterization of SARS-CoV-2 and the Spike-targeted medicines and vaccines against SARS-CoV-2. As the first to apply causal inference to mutational analysis on SARS-CoV-2 genomes, this work may inspire more related applications and promote the development of interdisciplinary fields.

**Declarations**

Data Availability
SARS-CoV-2 genome sequences as of May 11, 2022, are retrieved from GISAID Website (12). Only high-quality complete sequences are retained; hence 7,699,174 genome sequences of SARS-CoV-2 are obtained. Please c.f. Supplementary_FastaID.csv for detailed information of genome sequences.

Please c.f. Supplementary_EffectScore.csv for effect scores and validations of all 107 estimated mutations.

**Code Availability**

Source code of ATE estimation is available at: [https://github.com/Dywangxin/spikemut](https://github.com/Dywangxin/spikemut).

FoldX5 for validation is available at: [https://foldxsuite.crg.eu/](https://foldxsuite.crg.eu/) (38).

CSCS for validation is available at: [https://github.com/brianhie/viral-mutation](https://github.com/brianhie/viral-mutation) (42).

**Declaration of Competing Interests**

None.

**Acknowledgments**

This work was supported by The National Natural Science Foundation of China (Grant Nos. 62206309, 32070025 and 31800136). The authors gratefully acknowledge authors and laboratories that shared SARS-CoV-2 genome sequences via GISAID, on which this research is based.

**References**


16. Liu Y, Rocklöv J. The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. Journal of travel medicine. 2021;28(7).


Table 1

Table 1. Effect score of the top twenty mutations, with validations and references. The effect is validated in the Spike protein stability, ACE2 binding affinity, the potential for immune escape, and supporting references, abbreviated as Stability, Affinity, Escape, and References, respectively. Symbol representations: , highly positive effect; , potential positive effect; , no significant positive effect; NA, not applicable. VoC variants related to each mutation are represented in Greek letters. The mutation with no validated positive effect is in red.

<table>
<thead>
<tr>
<th>Mutation</th>
<th>VoC strains</th>
<th>Effect Score</th>
<th>Validations</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stability</td>
<td>Affinity</td>
</tr>
<tr>
<td>T478K</td>
<td>δ, o</td>
<td>1.9298</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D614G</td>
<td>α, β, γ, δ, o</td>
<td>0.6786</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S704L</td>
<td>o</td>
<td>0.5466</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H655Y</td>
<td>γ, o</td>
<td>0.2628</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N501Y</td>
<td>α, β, γ, o</td>
<td>0.2578</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V213-</td>
<td>None</td>
<td>0.2332</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>S477N</td>
<td>o</td>
<td>0.1614</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P26-</td>
<td>o</td>
<td>0.1316</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Q498R</td>
<td>o</td>
<td>0.1263</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S371F</td>
<td>o</td>
<td>0.1120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R408S</td>
<td>o</td>
<td>0.1114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T95I</td>
<td>o</td>
<td>0.0899</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L24-</td>
<td>o</td>
<td>0.0795</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>E484K</td>
<td>β, γ</td>
<td>0.0755</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T376A</td>
<td>o</td>
<td>0.0736</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V213G</td>
<td>o</td>
<td>0.0722</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P681H</td>
<td>α, o</td>
<td>0.0705</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>A222V</td>
<td>δ</td>
<td>0.0661</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N764K</td>
<td>o</td>
<td>0.0656</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D405N</td>
<td>o</td>
<td>0.0602</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figures

Data Preprocessing

- **Data Source**
  - SARS-CoV-2 Sequences from GISAID

- **Data Curation**
  - Sequences and mutations

- **Feature Extraction**
  - \( M = \begin{pmatrix}
  M_1 & M_2 & M_3 & \cdots & M_n & R_0 \\
  1 & 1 & 1 & \cdots & 0 & R_{01} \\
  0 & 1 & 1 & \cdots & 1 & R_{02} \\
  1 & 0 & 0 & \cdots & 1 & R_{03} \\
  \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  1 & 0 & 1 & \cdots & 1 & R_{0m}
  \end{pmatrix} \)

Effect Estimation

- **Model Training**
  - For mutation \( M_i \):
    \[
    \theta_i = E[Y^{(T-1)} - Y^{(T-0)}|X]
    \]
  - where \( T = \begin{pmatrix}
    M_1 \\
    0 \\
    0 \\
    \vdots \\
    1
  \end{pmatrix} \)
  - \( Y = \begin{pmatrix}
    R_0 \\
    R_{01} \\
    R_{02} \\
    \vdots \\
    R_{0m}
  \end{pmatrix} \)

  \[
  X = \begin{pmatrix}
    M_1 & M_2 & M_3 & \cdots & M_{n-1} & M_n \\
    1 & 1 & 1 & \cdots & 1 & 0 \\
    0 & 1 & 1 & \cdots & 0 & 1 \\
    1 & 0 & 0 & \cdots & 1 & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
    1 & 0 & 1 & \cdots & 0 & 1 \\
  \end{pmatrix}
  \]

Validation & Application

- **Mutation Influence**
  - 1. Spike protein stability
  - 2. ACE2 binding affinity
  - 3. Immune escape

- **Results**
  - 1. Effect score (ATE) of mutations
  - 2. Secondary results

- **Downstream Analysis**
  - 1. Key mutation identification
  - 2. Regional study on Spike
  - 3. Viral fitness and R0 prediction
  - 4. Other analysis

Figure 1

Schematic representation for the framework of this study, including the Data Preprocessing, the Effect Estimation, and the Validation and Application. Genome sequences of SARS-CoV-2 are retrieved and integrated for data curation and feature extraction. In data curation, Spike sequences are aligned to the reference sequence for mutation detection; in feature extraction, each sequence is mapped into a mutation combination with the corresponding R0, as a row vector in the feature matrix. For each mutation, the causal inference model is utilized to estimate its average treatment effect (ATE) on the outcome R0, with other mutations serving as observable covariates. Estimated ATE serves as the effect score of each mutation, which is validated by computing methods for detailed influences, including the Spike protein stability, the human angiotensin-converting enzyme 2 (ACE2) binding affinity, and immune escape. Effect scores can support downstream analysis, including key mutation identification, protein region study, viral fitness and R0 prediction, etc.
Figure 2

A. Residues of the top ten mutations represented by red spheres, in the Spike-ACE2 complex (PDB: 7A94) (40, 41), visualized by Visual Molecular Dynamics (VMD) (50, 51). ACE2 is colored in cyan. The Spike monomer binding with ACE2 is colored, with the other monomers in grey. B. Close-up view of the binding interface between Spike and ACE2.

Figure 3
Manhattan plot of mutations with positive effect scores across the Spike gene. The top twenty mutations are explicitly labelled. The vertical dashed line represents the S1-S2 subunit boundary (amino acid position 685). The horizontal dashed line represents the lower limit of the top twenty mutations. The blue and yellow rectangles represent the region of the N-terminal domain (NTD) and the receptor-binding domain (RBD), respectively.

Figure 4

Polynomial regression of fitness score and R0. The regression is trained by all points except BA.2.12.1, with the shaded area representing the 75% confidence interval. The point of BA.2.12.1 is subsequently illustrated as a validation.
Figure 5

Monthly historical fitness score of SARS-CoV-2. The line represents the historical fitness score. Stacked bars show the contribution of different Spike regions. The emergence of strains is indicated: the D614G strain, Alpha, Delta, and Omicron.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFiguresTables.docx
- SupplementaryFastalD.7z
- SupplementaryEffectScore.csv