Genome Biology

RESEARCH

Open Access

Real-time identification of epistatic interactions in SARS-CoV-2 from large genome collections

Gabriel Innocenti[1,2,3†], Maureen Obara[4], Bibiana Costa[4], Henning Jacobsen[5,6], Maeva Katzmarzyk[5,6], Luka Cicin-Sain[5,6], Ulrich Kalinke[2,4] and Marco Galardini[1,2*†]

†Gabriel Innocenti and Marco Galardini contributed equally to the work.

*Correspondence: galardini.marco@mh-hannover.de

[1] Institute for Molecular Bacteriology, TWINCORE Centre for Experimental and Clinical Infection Research, a joint venture between the Hannover Medical School (MHH) and the Helmholtz Centre for Infection Research (HZI), Hannover, Germany
Full list of author information is available at the end of the article

Abstract

Background: The emergence of the SARS-CoV-2 virus has highlighted the importance of genomic epidemiology in understanding the evolution of pathogens and guiding public health interventions. The Omicron variant in particular has underscored the role of epistasis in the evolution of lineages with both higher infectivity and immune escape, and therefore the necessity to update surveillance pipelines to detect them early on.

Results: In this study, we apply a method based on mutual information between positions in a multiple sequence alignment, which is capable of scaling up to millions of samples. We show how it can reliably predict known experimentally validated epistatic interactions, even when using as little as 10,000 sequences, which opens the possibility of making it a near real-time prediction system. We test this possibility by modifying the method to account for the sample collection date and apply it retrospectively to multiple sequence alignments for each month between March 2020 and March 2023. We detected a cornerstone epistatic interaction in the Spike protein between codons 498 and 501 as soon as seven samples with a double mutation were present in the dataset, thus demonstrating the method's sensitivity. We test the ability of the method to make inferences about emerging interactions by testing candidates predicted after March 2023, which we validate experimentally.

Conclusions: We show how known epistatic interaction in SARS-CoV-2 can be detected with high sensitivity, and how emerging ones can be quickly prioritized for experimental validation, an approach that could be implemented downstream of pandemic genome sequencing efforts.

Background

The COVID-19 severe respiratory syndrome is caused by the SARS-CoV-2 virus, which emerged in late 2019 in China and quickly escalated to a pandemic. Since then, the virus has differentiated into a number of lineages, some of which have taken over the whole

population in successive sweeps reminiscent of clonal interference [1]. These successful lineages—designated as variants of concern (VOC)—have acquired a number of mutations that have increased their ability to infect new hosts and escape infection and vaccine-induced immunity [2–9]. The unprecedented efforts in the genomic epidemiology of the virus leading to millions of whole sequences being deposited in public and restricted-access databases have accelerated the pace by which emerging lineages are tracked and the impact of genetic mutations is estimated [1, 10–12]. Together with expedited experimental measurements of the impact of single mutations, it has been possible to obtain estimates of the fitness advantage of emerging variants with as little delay as a few weeks [13–15]. Given the successes of SARS-CoV-2 genome sequencing in tracking the emergence and spread of variants and to develop groundbreaking vaccines [16], it is hard to imagine it not becoming a routine tool in handling current and future epidemics [17].

Which other information of relevance can be obtained from large genome sequencing datasets? A growing body of evidence from population genetics and evolutionary studies indicates that the impact of individual genetic variants can be modulated by the presence of other variants in other sites, a phenomenon known as epistasis or genetic background effect [18]. In practice, this means that the measured impact of a genetic variant may be very different when the same variant is present in a different genetic background, which in turn makes sequence and phenotypic evolution unpredictable at medium to long timescales [19]. In the context of SARS-CoV-2 evolution, epistasis might have implications for genomic surveillance of lineages; assumptions about the fitness advantage conferred by a mutation, be it through increased transmissibility or immune escape, might be invalidated and lead to erroneous applications of public health measures or vaccine updates. At a smaller scale, this possibility has been experimentally confirmed for the receptor binding domain (RBD) of the SARS-CoV-2 spike protein, for which mutations have a different impact on antibody escape depending on the overall genetic background (i.e., different VOCs) [20, 21]. Detecting potential epistatic interactions between mutating sites in the SARS-CoV-2 genomes could be used as a sign that the fitness effect of a particular mutation in a particular genetic background might not generalize in another. Furthermore, positions participating in an epistatic interaction might indicate that a particular fitness "peak" can only be reached through another neutral or slightly deleterious mutation [22]. Predicting whether seemingly neutral mutations are enabling further ones that affect transmissibility or immune escape would therefore be a valuable epidemiological tool. Lastly, from the perspective of genomic surveillance, being able to quickly identify sites that participate in an epistatic interaction would help identify virus' variants that could have a fitness advantage and that could therefore quickly spread. The feasibility of predicting epistatic interactions and their potential impact on infection control has been shown for HIV [23, 24], the other recently emerged virus causing a global pandemic.

In order to make the estimation of epistatic interactions useful in the context of a rapidly unfolding pandemic, the method needs to have an appropriate combination of speed and precision/sensitivity. Ideally, it would require modest computational resources and would run in near real-time, so that it could be implemented as part of existing automated genomic epidemiology tools that feed on sequence repositories [1, 10, 25]. Some

of the computational approaches that are able to estimate epistatic interactions would not be suitable for this task for resource considerations, such as pseudolikelihood or phylogenetic methods [26–32]. These methods are generally applicable at the resolution of a single gene for their requirements in terms of computational resources: aggressive subsampling, potentially coupled with batch learning, would make these methods applicable at the whole genome level, but could theoretically reduce their ability to quickly identify new interactions and variants that are starting to emerge, as well as making them more difficult to implement.

Here, we explore the usefulness of a method based on the detection of mutual information (MI) between sites in a multiple sequence alignment [33] as an indirect estimate of mutational epistasis, which is able to handle alignments with more than 10^6 samples. Since MI formally measures the level of correlation in substitutions between two sites, it is liable to incorrectly identify pairs of sites that are correlated through stochastic or population structure effects and not because they participate in an epistatic interaction. Multiple studies have however shown that despite these limitations, MI-based methods perform similarly to pseudolikelihood methods, assuming that a proper control for population structure is applied [34–37]. We applied a MI-based method to all high-quality publicly available SARS-CoV-2 sequences as of April 14th, 2023 ($N = 6{,}644{,}032$). The method detected 474 putative epistatic interactions between different positions, 222 of which with high mutual information. We validated the highest scoring hits using a list of known mutation of interest/concern (MUI and MOC) [38], data from deep mutational scans [20, 21], and different epistasis models [26–28]. We further adapted the method to account for the "age" of each sequence, leveraging the metadata associated with each sequence, and thus showing how interactions gain/loss varies over time and variant emergence, and how this could be used as a near real-time genomic surveillance system. To this extent we demonstrated how our method is highly sensitive, being able to identify a known epistatic interaction in the Omicron variant with as little as seven sequences. Taken together, these results offer yet another vision of the future of pathogen genomic epidemiology, in which the subtle complexities of the evolution of biological sequences are taken into account.

Results

Mutual information methods can estimate epistatic interactions in a large-scale genomic dataset

This analysis was possible thanks to the data available on the GISAID database, from which we downloaded the SARS-CoV2 multialignment file and its relative phylogenetic tree. After 2 steps of filtering, followed by trimming and deduplication, a phylogenetic weighting strategy was applied to each sequence in the alignment in order to account for population structure. Finally, the spydrpick algorithm [33] was run to compute mutual information (MI) between every pair of positions in the genome (pipeline shown in Fig. 1A). Each position pair was then assigned to an outlier level O (O1, O2, O3, O4) according to their score, with O4 values indicating the strongest predicted interactions (see Methods).

The unfiltered dataset counted 15,372,583 sequences (up to 14th April 2023). From these, we picked the public sequences (i.e., also available in the NCBI's

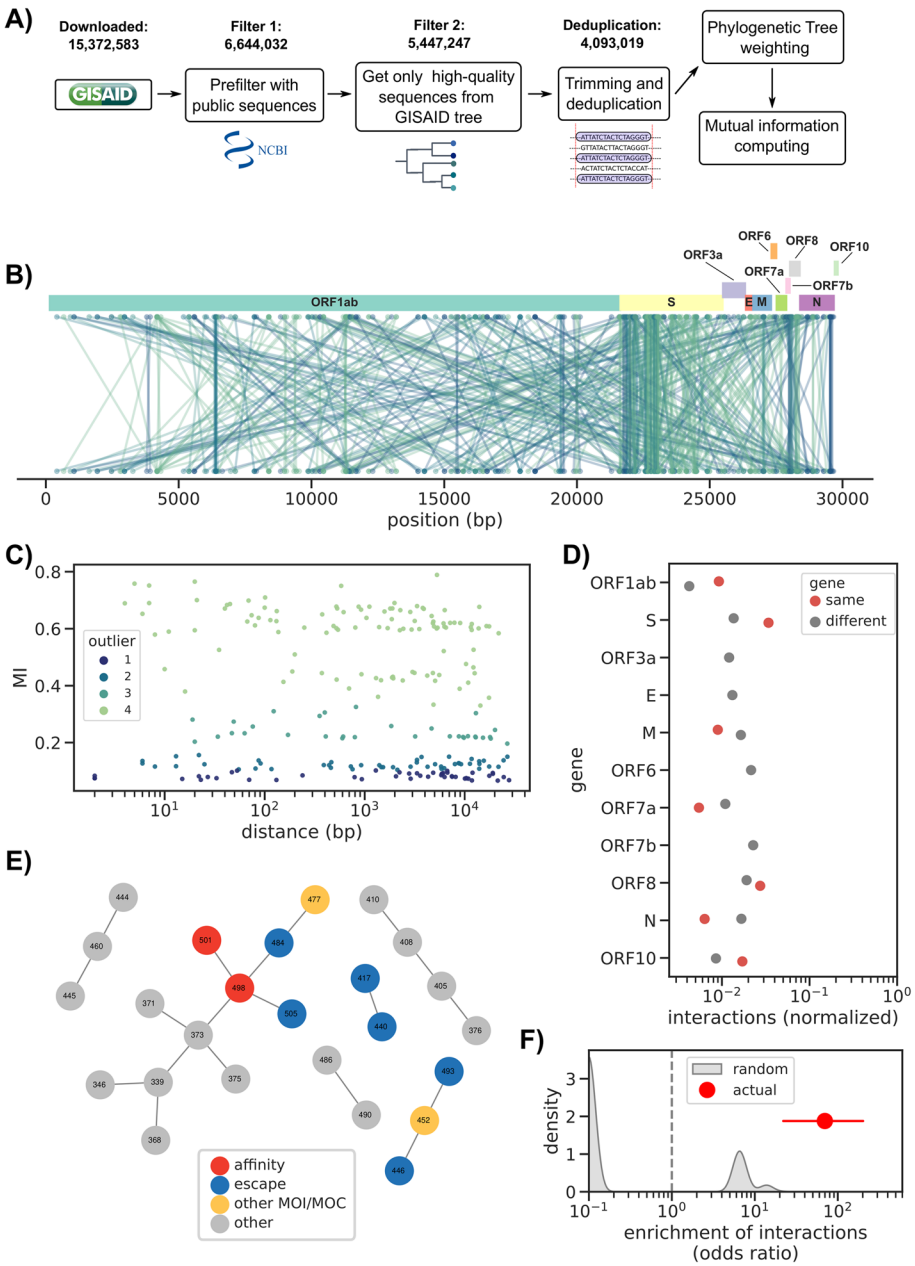

Fig. 1 Estimation of epistatic interactions from > 4 M SARS-CoV-2 sequences. **A** Analysis workflow. **B** Overview of the estimated epistatic interactions across the whole SARS-CoV-2 genome; each interaction is colored according to its outlier level (O1 to O4), using the same color scale as panel **C**. **C** Distance distribution between all interactions. **D** Proportion of interactions within and between genes, normalized by the nucleotide length of each focal gene. **E** Interaction network for the RBD region of the Spike gene; amino acid positions are colored according to publicly available annotations. The category "other" indicates positions which are not known to have an impact on affinity to ACE2, immune escape, or are otherwise flagged as MOI/MOC. **F** Enrichment of interactions between positions known to epistatically interact (red dot) versus a series ($N = 1000$) of random RBD networks with the same number of interactions as the real one (gray distribution). The red horizontal line indicates the 95% confidence interval

Genbank database, $N = 6{,}644{,}032$) which were then filtered by using only the high-quality sequences present in the GISAID SARS-CoV2 phylogenetic tree and by removing exact duplicates ($N = 4{,}093{,}019$).

We obtained 474 interactions (between 247 unique positions) with many interactions both within and between different genes (Fig. 1B, Additional file 1: Table S1). Most of the observed interactions were concerning the ORF1ab and Spike gene, with 179 and 185 interactions each, respectively. The interactions were assigned to 4 "outliers levels" (O1 to O4) based on their MI value (Methods), with an overrepresentation of O4 interactions ($N_{O4} = 222$) compared to the others which showed similar frequencies ($N_{O1} = 92$, $N_{O2} = 96$, $N_{O3} = 64$). However, they showed a similar distribution in function of the distance between the interacting genome positions (Fig. 1C).

Another important aspect that we considered was if the interactions fell within the same gene or between different ones: ORF3a, the Envelope gene (E), ORF6, and ORF7b presented only interactions between different genes, while the remaining had both kinds. In particular, Spike, ORF1ab, and the Nucleocapsid gene (N) presented the highest ratio between different/same gene interactions using normalized counts (Fig. 1D), showing more interactions in the same gene compared to the others.

To further validate our predictions, we labeled positions in the Spike RBD as either affinity mutations [21], escape mutations [20], or other MOI/MOC [38]. These positions have been shown to engage in epistatic interactions resulting in increased viral fitness through higher infectivity or immune escape. The RBD interaction's network we obtained from our predictions (shown in Fig. 1E) captured all 3 kinds of mutations. Among them, the interaction between affinity mutations 501 and 498 was observed, but also between escape mutations and other mutations of interest including 484, 505, 446, and 477. In addition, other positions with unknown significance (e.g., 373, 390, 486, 376) were highlighted by our method and would need a deeper investigation to understand if they engage in actual epistatic interactions or if they are false positives. We verified that the reconstructed interaction network was significantly different than a random one via a permutation test, which showed strong enrichment in known epistatic interactions in the RBD region of the Spike gene (odds ratio 69.9 [22.0–202.1], Fig. 1F). We observed an even stronger enrichment (odds ratio 104.8 [42.9–336.0]) when restricting interactions with outlier levels O3 and O4. We further used the outlier thresholds O3 and O4 to build a binary classifier for known RBD epistatic interactions (Methods), which had a specificity of 0.33 [0.17–0.71] and 0.50 [0.17–0.71], and a sensitivity of 1 and 0.71, respectively. Lastly, we used the pairwise epistasis coefficients for 15 BA.1 mutations [21] and measured the ability to recover those with a value over 0.4 using O3 and O4 interactions, for which we recorded a specificity of 0.93 [0.86–0.97] and sensitivity of 0.4 for both thresholds. These results showed how this method was able to pick many experimentally verified interactions and possible new ones.

Influence of dataset size on the estimation of epistatic interactions

Even though our implementation of the spydrpick algorithm can be run in reasonable time in a high-performance cluster (i.e., ~ 36 h, each job requiring > 150 Gb of RAM), we reasoned that for it to be of real use it would need to work with a leaner dataset. We therefore created 4 smaller datasets with orders of magnitude less sequences than the complete > 4 M dataset ($N = 1000$, 10,000, 100,000, and 1,000,000, randomly chosen). We observed that the number of predicted interactions was larger with smaller datasets and that the numbers became comparable with the complete dataset when at least

100,000 sequences were considered (Fig. 2A, Additional files 2–5: Table S2–5). Despite the large difference in predicted interactions, we observed a strong correlation ($r > 0.95$) in MI values for those interactions found both in a reduced subset and the complete dataset (Fig. 2B–E). The overlap became larger with subset size, from 57.8% of the 474 interactions from the complete dataset predicted in the $N = 1000$ subset to 92.4% in the $N = 1{,}000{,}000$ subset. We also observed a stronger enrichment in known epistatic interactions in the Spike RBD with increasing sample size, with comparable strength as the complete dataset when using at least 100,000 sequences (odds ratio > 10 for all subsets, Fig. 2F–I). When reducing the interactions to outlier levels O3 and O4 we observed a consistently high level of enrichment across all subsets (odds ratios > 100), which could be explained by the high number of predicted interactions with outlier level O1 and O2 in the smaller 1000 and 10,000 datasets (Additional file 6: Fig. S1). We then used the O3 outlier threshold to build a binary classifier for RBD known epistatic interactions, which yielded the following specificity/sensitivity values: 0.97 (0.95–0.99)/0.19 (0.09–0.28) ($n = 1000$), 0.80 (0.68–0.96)/1.00 ($N = 10{,}000$), 0.46 (0.28–0.84)/1.00 ($N = 100{,}000$), and 0.38 (0.21–0.74)/1.00 ($N = 1{,}000{,}000$). When predicting pairwise interactions with coefficient > 0.15 using O3 interactions we instead recorded the following specificity/sensitivity values: 0.93 (0.86–0.97)/0.40 ($n = 1000$), 0.94 (0.87–0.98)/0.40 (0.00–1.00) ($N = 10{,}000$), 0.94 (0.88–0.98)/0.40 ($N = 100{,}000$), and 0.93 (0.87–0.97)/0.40 ($N = 1{,}000{,}000$). We therefore estimate that mutual information-based estimation of epistatic interactions in SARS-CoV-2 can be performed with datasets with sizes between 10,000 and 100,000 sequences.

Real-time estimation of epistatic interactions

The unprecedented genomic epidemiology effort during the SARS-CoV-2 pandemic offers the opportunity to test the usefulness of computational methods to quickly identify emerging pathogen variants and mutations of concern [15]. In the same spirit, we sought to understand if we could apply the mutual information methodology described here to develop a near real-time system to discover emerging epistatic interactions. We

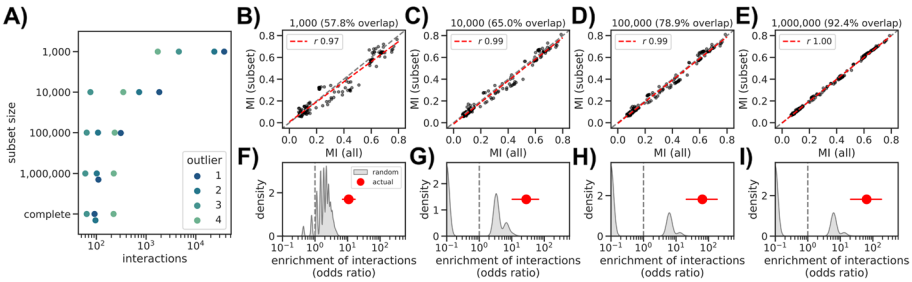

Fig. 2 Influence of dataset size on mutual information estimation. **A** Number of interactions across datasets, divided by outlier level; the "complete" dataset refers to the one computed using 4,093,019 sequences. **B**–**E** Linear regression analysis of mutual information values in the complete dataset (*x*-axis) versus the same interactions in the 1000 (**B**), 10,000 (**C**), 100,000 (**D**), and 1,000,000 (**E**) subsets. The dashed red line shows the linear regression, and the corresponding *r*-value is indicated in the panel legend. **F**–**I** Enrichment of interactions between positions known to epistatically interact (red dots) versus a series ($N = 1000$) of random RBD networks with the same number of interactions as the real one (gray distribution). Subsets are the same as the panel directly above: 1000 (**F**), 10,000 (**G**), 100,000 (**H**), and 1,000,000 (**I**). The red lines indicate the 95% confidence interval

therefore used the sample collection date associated with each viral sequence to divide the complete dataset into 37 subsets, one for each month from March 2020 up to March 2023, and selecting up to 2500 sequences for each month. We then computed the mutual information for each month using the cumulative sequences up until the focal month (Fig. 3A). In order to highlight emerging interactions and remove earlier ones, we introduced a further weight to each sequence based on its distance in time from the focal month. That is, we placed lower importance to older sequences, halving their weight at around 4 months (120 days), using a hill function (Methods, Additional file 6: Fig. S2, and Additional file 6: Fig. S3). We reasoned that this would allow the method to discard older interactions and focus on emerging ones.

We first verified that the final subset yielded comparable results as larger datasets and as a partially orthogonal method [26–28]. For this, we used the last subset (March 2023, $N = 84{,}923$) and avoided the time weighting step to allow for a fair comparison. We observed that MI values had a high correlation with the complete dataset, with 64.1% of the interactions found in the complete dataset recovered in the final

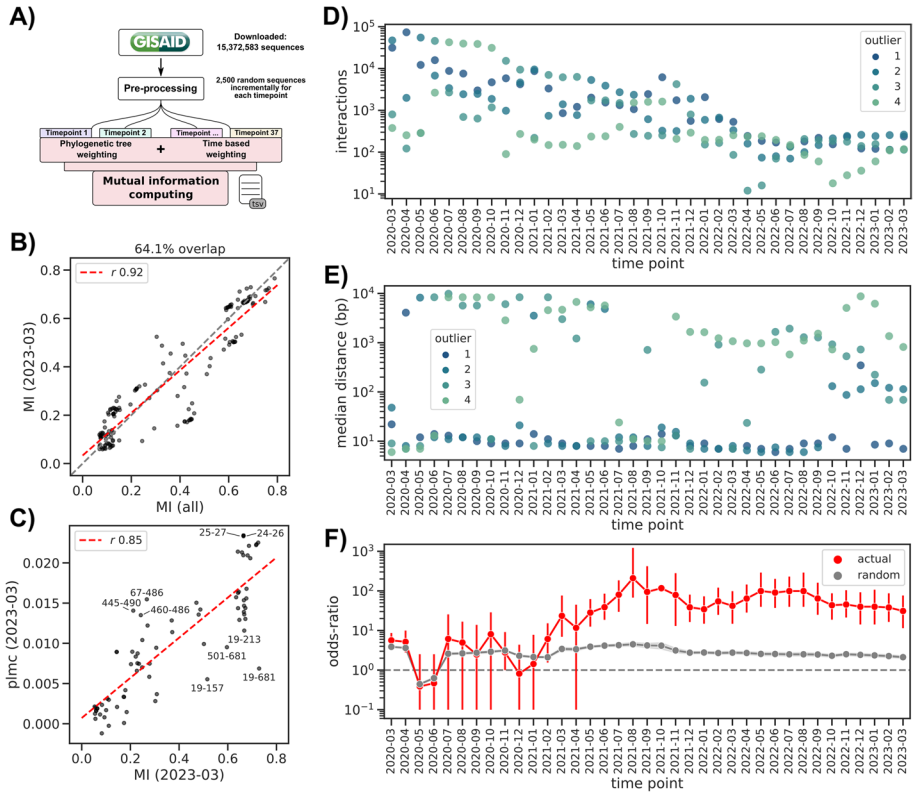

Fig. 3 Time-based mutual information estimation. **A** Analysis workflow. **B** Comparison between MI values in the total dataset versus those in the last subset (March 2023, without the time weighting correction). **C** Comparison of MI values in the Spike gene between the March 2023 subset and the output of the plmc pseudo-likelihood method; the 10 interactions with the highest absolute residuals with respect to the linear regression are labeled with the amino acid position in the Spike gene. **D** Number of interactions for each outlier category across all the time subsets. **E** Median distance between interacting positions, divided by outlier level and across all the time subsets. **F** Enrichment of interactions between positions known to epistatically interact (red dots) versus a series ($N = 1000$) of random RBD networks with the same number of interactions as the real one (gray line, shaded area represents the standard deviation, vertical red lines indicate the 95% confidence interval)

subset and with a linear regression *r*-value of 0.92 (Fig. 3B, Additional file 7: Table S6). We also compared the correlation between the MI values against estimates using a pseudo-likelihood method [26–28] (Methods). For this analysis, we focused only on the Spike gene as the plmc implementation does not scale well to genome-scale nucleotide alignments. We again observed a very high correlation (linear regression *r*-value of 0.85, Fig. 3C, Additional file 8: Table S7). We singled out the 10 interactions with the highest residual: two of them (amino acid positions 460/486 and 445/490) belonged to the RBD region of the spike and had a higher interaction score when using the pseudolikelihood method. Conversely, the interaction between the mutations of concern 501/681 was scored higher when using the MI-based method. Even when applying our time-based weighting we observed a high correlation between the two methods, with a Pearson *r*-value higher than 0.4 for all but three timepoints (Additional file 6: Fig. S4). Perhaps more importantly, we found that our method was able to identify known epistatic interactions in the Spike RBD [21] earlier and with more consistency over timepoints than the pseudolikelihood method (Additional file 6: Fig. S5), which is a desirable property (other than speed and ease of implementation) for a real-time surveillance system.

We next computed MI values for all subsets with the time correction; the number of estimated interactions decreased steadily over time, from a total of 79,586 for the March 2020 subset to 730 in the final March 2023 subset. The number of O4 interactions was more stable over time, especially starting from 2021: the median number of O4 interactions in 2020 was 1513 versus 196 in the 2021 to 2023 period (Fig. 3D, Additional file 9: Table S8). The median distance between interacting nucleotide positions was more stable over time, with O3/O4 interactions being more distant from each other than lower confidence ones: the median distance was 284 and 1232 bases over all subsets, respectively, while O1/O2 interactions' median distance was 9 and 12 bases, respectively (Fig. 3E). A lower distance between interactions could be a sign of a lower confidence interaction, at least in the general case.

In line with the fact that variants of concern with both higher infectivity and immune escape capabilities did not emerge before the spring of 2021 (e.g., the Delta variant), we did not observe an appreciably high enrichment of known interactions between affinity and escape Spike mutations before at least July 2021 (odds ratio 80.2 [29.5–184.2]), and we observed a consistently high enrichment level until the last time point (odds ratio > 30, Fig. 3F, Additional file 10: Table S9). We measured the ability of this method to recover known RBD interactions by assessing the F1 score, specificity, and sensitivity of binary classifiers built using the outlier thresholds (Additional file 6: Fig. S6). The O4 binary classifier had the highest F1-score in the December 2021 dataset (value 0.8 [0.61–0.91]), consistent with the emergence of the Omicron variant, and no predictive power before then. When we used the predicted RBD interaction networks before applying the ARACNE filtering algorithm [39], we observed a similar trend for both enrichment and binary classification than with the filtered datasets, albeit with lower odds ratio and F1 scores (Additional file 6: Fig. S7). Lastly, using O4 interactions to predict BA.1 epistatic interactions we measured the highest F1 score for the December 2021 dataset (0.35 [0.02–0.67]), again consistent with the rise of this variant at the same time (Additional file 6: Fig. S8). This validation

using experimentally determined epistatic interactions suggests that efficient computation of MI values could be used to implement a near real-time surveillance system for epistatic interactions.

In order to put this idea of a near real-time surveillance system for epistatic interactions to the test, we focused on determining how early known epistatic interactions would be highlighted by the MI-based method. An ideal interaction pair is between positions 498 and 501 in the Spike protein, which has been shown experimentally to increase the Spike's affinity to the ACE2 receptor both compared to the wild type and the single mutants alone, and thus the virus' infectivity [20, 40]. While mutations at position 501 were already observed in the Alpha variant, the double mutation was not observed before the Omicron variant appeared around November 2021 (Fig. 4A). Consistent with this observation, we first predicted an O3 interaction between these two positions in the November 2021 dataset, which contained only 7 viral sequences out of 2500 that were annotated as lineage 21 K (Omicron BA.1, Fig. 4B and D). We note that our weighting scheme inflated the effective number of sequences belonging to this lineage to 86 (Additional file 6: Fig. S3). In the following month the frequency of Omicron viral sequences already increased to 56.4% and as expected the interaction strength between the two positions increased to reach the O4 level. We observed that

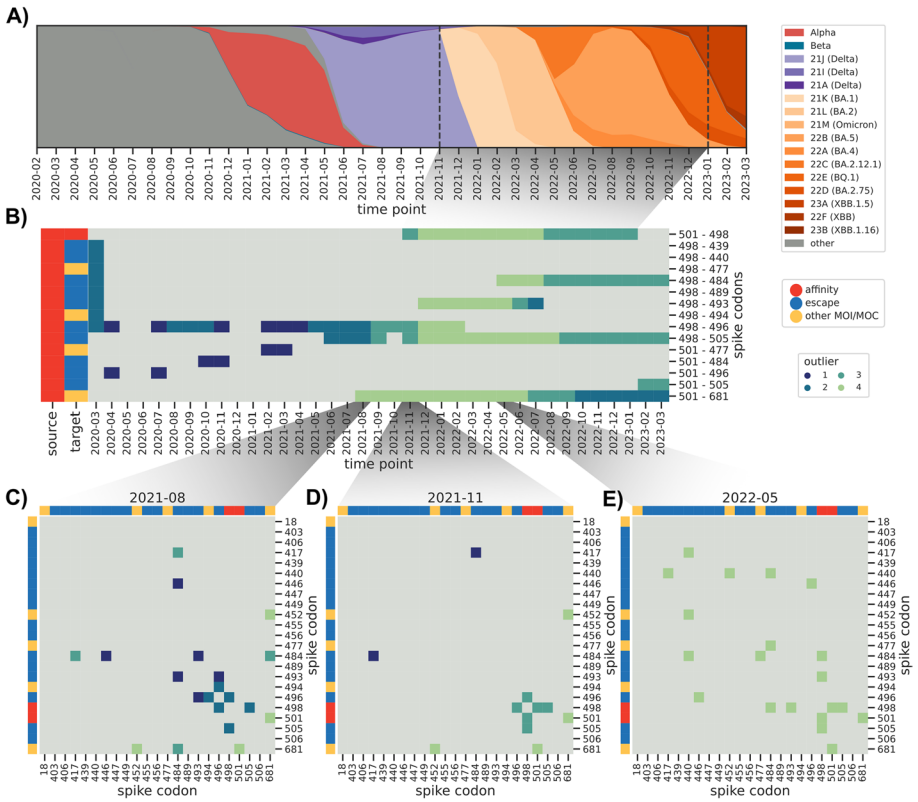

Fig. 4 Predicted epistatic interaction in the Spike gene as a function of time. **A** Muller plot indicating the relative frequency of each lineage as a function of time. **B** Presence/absence matrix of predicted interactions between Spike codons 498/501 and those labeled either as escape variants or other mutation of interest/concern. Gray indicates no predicted interaction. **C**–**E** Interaction heatmaps between selected Spike gene codons at particular time points

this interaction faded with time and eventually disappeared in February 2023, consistent with mutations having reached fixation in the population and thus bearing no further mutual information between them. While the time point datasets are cumulative and thus contain all sequences up until the focal time point, our sequence age-based correction greatly reduces the contribution of sequences older than a year in the calculation of the MI values (Additional file 6: Fig. S2 and Additional file 6: Fig. S3). This seems a desirable property for a real-time surveillance system that is focused on highlighting new epistatic interactions as they appear, even when the number of sequences bearing a double mutation is very low (e.g., $N=7$).

We observed a similar pattern for the interaction between either affinity-related mutation (Spike codon 498 or 501) and escape or other mutations of concern, such as 501/484 (appeared for the first time in the Beta B.1.135 variant) already emerged in October 2020. In many cases the interactions were first estimated when the Omicron variant first appeared (around November/December 2021), consistent with the studies that have characterized mutations at these positions to interact epistatically and to modulate infectivity and immune escape [20, 21].

Another particularly interesting interaction we singled out was between spike codons 501 and 681, the latter having been described as conferring enhanced resistance to innate immunity [41]. Position 681 in particular has accumulated different amino acid substitutions in different variants: P681H in lineage 20I (Alpha B.1.1.7) and all Omicron subvariants, and P681R in lineage 21A (Delta B.1.617.2). Both positions have mutations that were first observed together when the Alpha variant emerged (around November 2020) and we therefore expected to see a predicted interaction with a high MI value. We however first estimated the 501/681 interaction in August 2021 (Fig. 4C), which corresponded with the virtual extinction of the Alpha variant. We then checked the raw MI interactions, which contain indirect ones (Methods), and as expected the 501/681 interaction was first observed in November 2020 with outlier level O3, which we observed every month until the last time point (Additional file 6: Fig. S9). We posit that this apparent failure of the method to single out a known combination of mutations enhancing viral fitness was due to the relatively low number of sequences and their diversity until Summer 2021, which we have shown results in a large number of estimated interactions, many of which are likely false positives (Fig. 3D).

Apart from recapitulating known epistatic interactions, we sought to determine whether MI-based predictions have real predictive power. To answer this question, we devised a "blind" validation experiment. We extended our dataset to include viral sequences until August 2023 (Fig. 5A) and selected O4 interactions that fulfilled the following criteria: (i) both sites needed to be in the Spike RBD, (ii) one of the sites had to be previously characterized, and (iii) the interaction between the two sites had to have emerged at later timepoints. Based on these criteria, we selected three interactions (Fig. 5C): 446–460, 446–486, and 452–490 (Fig. 5B). We then analyzed the changes in the frequency of non-synonymous variants at these sites and selected six for testing, combined in four pairs. Of those, L452R/F490S and G446S/F486V were anti-correlated in their frequency and virtually never occurred together in our dataset. We observed the G446S/F486S pair at a low frequency (87 samples after 2023–05), and almost never observed the G446S/N460H pair (5 samples). We chose the

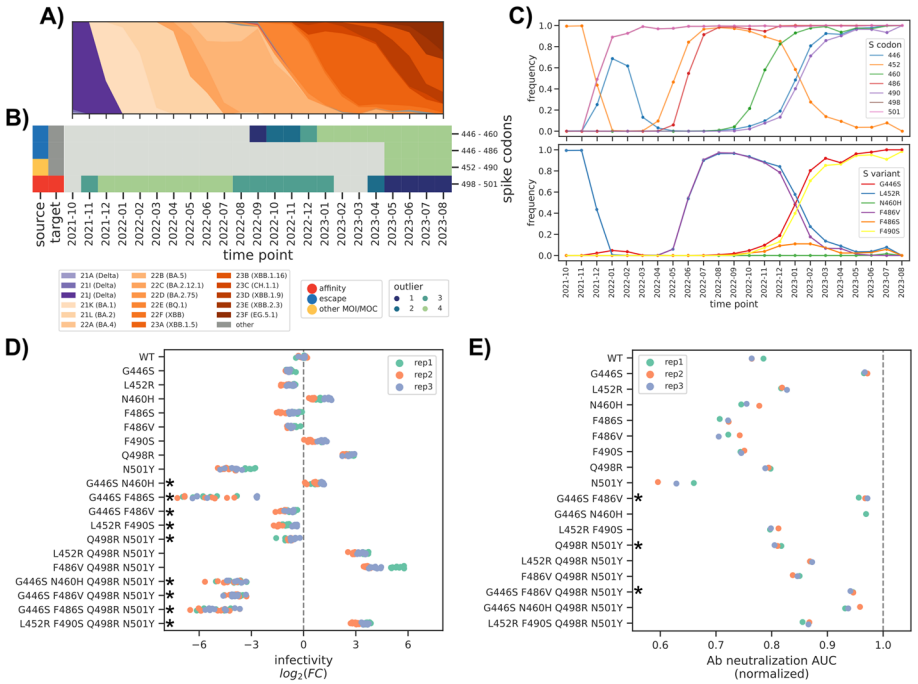

Fig. 5 Experimental validation of emerging interactions. **A** Muller plot indicating the relative frequency of each lineage as a function of time, extending the dataset until August 2023. **B** Presence/absence matrix of the top predicted interactions between in the Spike RBD, plus codons 498/501. Gray indicates no predicted interaction. **C** Frequency of variants at each Spike codon shown in **B** (top) and of six selected non-synonymous variants (bottom). **D** Impact of single and double mutants on viral infectivity, in the WT and Q498R/N501Y backgrounds. **E** Impact of single and double mutants on viral immune escape, using the area under the curve (AUC) of the neutralization assay (Additional file 6: Fig. S10) as a metric. The gray dashed vertical line indicates the maximum detectable neutralization (i.e., a flat neutralization curve). Bold asterisks in both the **D** and **E** panels indicate a pair with a significant epistatic interaction (p-value < 0.05)

anti-correlated pairs to elucidate why these variants "avoid" each other, and the low-frequency pairs as a way to test an early warning system for mutation signatures that might rise in the future. We then tested each variant separately and as a pair in both the WT and Q498R/N501Y backgrounds, using the latter as a crude model of the Omicron variant. We used a pseudovirus system to test the impact of these variants on both infectivity (i.e., the ability to infect Vero-B4 cells, Fig. 5D, Additional file 11: Table S10) and immune escape (i.e., lower neutralization to the Imdevimab monoclonal antibody, Fig. 5E, Additional file 6: Fig. S10, Additional file 11: Table S10). We then used a simple model of epistasis to validate interaction effects (Methods).

We first confirmed the previously described epistatic interaction for the Q498R/N501H pair [21], both for its effect on infectivity (p-value 10^{-7}) and immune escape (p-value 10^{-9}). For both anti-correlated pairs, we found a significant interaction for either the infectivity assay (L452R/F490S and G446S/F486V with p-value $< < 10^{-10}$ in the WT background) and immune escape (G446S/F486V p-value 10^{-4}). In particular, we found that on the one hand, the G446S/F486V pair induced a large drop in infectivity in the Q498R/N501H background (p-value 10^{-4}) while the double mutant was fairly similar to the immune escape profile of the single G446S variant, thus compensating for the loss of escape shown by the F486V variant alone. For infectivity, we

observed the opposite effect for the L452R/F490S pair, observing a large increase in infectivity in the Q498R/N501H background, a significant effect (p-value 10^{-3}). The double mutant had a slightly better immune escape profile than the single mutants, although not significant. From these observations we can hypothesize that the G446S and F486V variants are anticorrelated for their strong defect in infectivity; we cannot however apply the same reasoning for the L452R/F490S pair, whose absence from circulating variants could be ascribed to stochasticity in population dynamics or interactions with other variants. We observed a similar impact on infectivity for the G446S/F486S (p-value $< < 10^{-10}$ for both backgrounds) and G446S/N460H (p-value $< < 10^{-10}$ and 0.02 for the WT and Q498R/N501Y backgrounds, respectively) pairs as G446S/F486V; based on these results we could estimate that variants carrying these pairs might have a fitness disadvantage, which indicates that they might have a low chance of increasing in frequency. Through this relatively small "blind" validation assay we demonstrated how interactions with high MI values indicate sites that interact epistatically, and thus the usefulness of the method to prioritize mutations for further validation.

The SARS-CoV-2 pandemic accelerated the usual pace of infection biology research: deep-mutational scanning and large-scale antibody escape assays have been developed and released at unprecedented speed [42, 43]. These are however necessarily limited to a single region of the genome of the virus: specifically the RBD region of the Spike gene. This necessarily excludes longer-range interactions; not only within the Spike gene, as the 501/681 interaction, but also between different genes. Given the difficulty in testing those interactions in a laboratory assay, we focused on those for which we had the highest confidence to provide the community with a list of potentially interesting interactions. Overall the number of predicted interactions between genes is much lower than those within each gene, with the lowest number being 102 interactions in May 2022, and similar to the median number for the last 12 time points (137, Fig. 6A). To focus on the highest-confidence predictions, we selected the inter-gene interactions that had one position in the Spike gene and had a large MI value (outlier level O4) in at least 9 time points; this resulted in 7 interactions (Fig. 6B). Interestingly we observed three intergene interactions involving the notable 681 Spike codon, which is known to increase viral fitness. Among them, we found codon 203 in the Nucleocapsid gene (N) [44], which is known to increase infectivity. The other two positions with interactions with the 681 Spike codon were codon 26 in the ORF3a gene and codon 82 in the M protease gene; both are not known to influence viral fitness, and indeed they both are defining mutations for the Delta variant, their frequency closely following that of this variant (Fig. 6C). We therefore suspect that these two interactions might be false positives, thus suggesting that despite the population structure correction, the method is sometimes still susceptible to the impact of the strong clonal interference observed with this virus and almost complete lack of recombination. It is however also possible that these mutations do have a yet to be determined impact on viral fitness.

Discussion

The tragic public health toll of the COVID-19 pandemic has coincided with an unprecedented pace of scientific discovery and application in developing diagnostics and treatment solutions. Large-scale sequencing of viral samples has played an important part

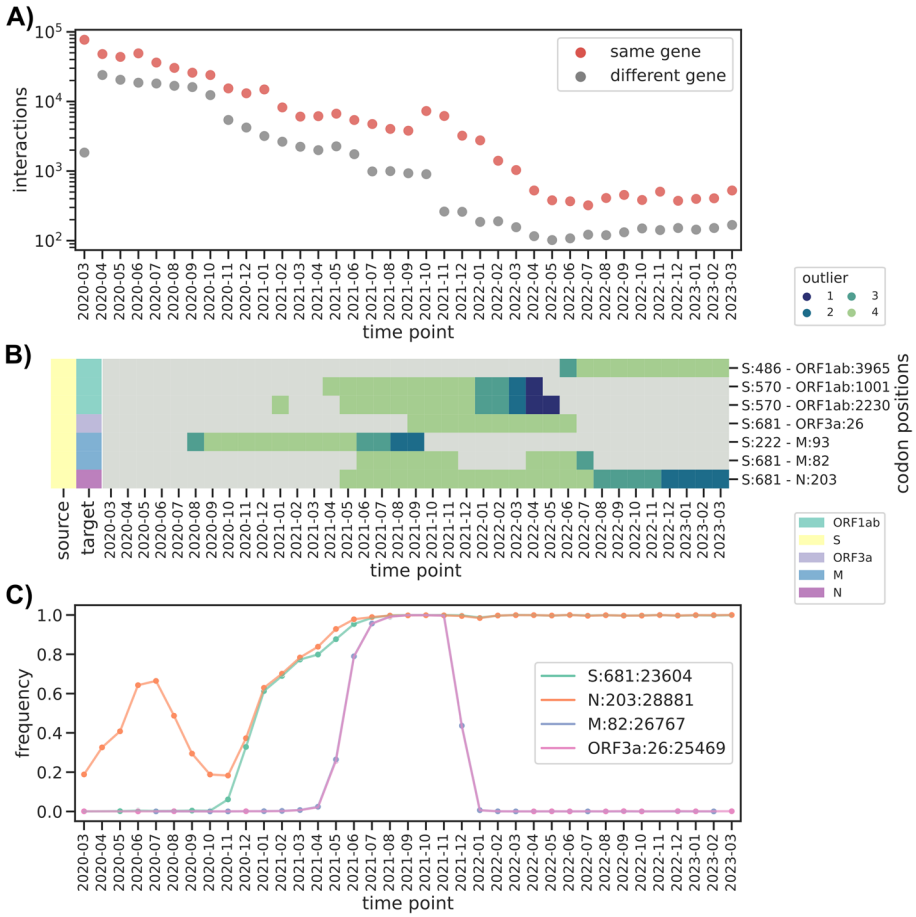

Fig. 6 Between gene interactions. **A** Change in within and between gene interactions across time points. **B** Presence/absence matrix of predicted interactions between Spike codons and positions in other genes. Gray indicates no predicted interaction. **C** Frequency of nucleotide substitutions at specific positions interacting with Spike codon 681. Labels are encoded with format gene:codon:nucleotide position. Purple and pink lines overlap perfectly

in guiding public health interventions and more recently vaccine updates. The unprecedented scale of these genomic epidemiology efforts offer an ideal testbed for new computational approaches aimed at tackling future epidemics, for which genome sequencing will surely continue to play an important part [17].

While multiple approaches to rapidly estimate the fitness effects of individual genetic variants have been proposed [13–15], very few attempts have been made to estimate epistatic interactions between pairs of mutations. The interest in predicting them is not purely academic, as best exemplified by the appearance of the Omicron subvariants, each characterized by higher infectivity and immune evasion thanks to epistatic interactions, many of which have been experimentally verified [20, 21]. Of note is the combination of amino acid changes at Spike codons 498 and 501, each of which alone results in a modest increase of the Spike protein for the ACE2 receptor, while together they have been shown to result in a > 350-fold increase in ACE2 affinity, a clear example of positive epistasis. This large increase in fitness in turn likely allowed for the accumulation of slightly deleterious immune escape variants. Being able to quickly discover these

interactions from the large number of viral sequences being routinely generated could complement predictors for single variants to form a reliable early warning system.

The few approaches that have so far been described in the literature make them relatively unsuitable for a near real-time system for predicting epistatic interactions. Neverov and colleagues recently proposed a method that relies on ancestral sequence reconstruction over a time-calibrated phylogeny to infer epistatic interactions based on mutations appearing one after the other over a relatively short time span [29]. This method is interesting in its similarity to what we presented here because it also explicitly includes a time component to aid prediction; it is however most certainly not able to scale to a large number of sequences and be efficient, as it relies on constructing a timed phylogeny and reconstruct the ancestral states for each branch of the tree. In this study, we used a pseudolikelihood method, also termed direct coupling analysis (DCA) to validate the approach based on mutual information. Multiple studies have applied DCA to predict epistatic interactions in SARS-CoV-2, although at different times in the pandemic and with differences in the input alignments. Zeng and colleagues used DCA in the early stage of the pandemic (summer 2020) and reported a small number of predicted interactions, which is expected given the low diversity of the virus at that time [45]. While DCA can be considered a more accurate model to predict epistasis, available implementations of the model tend to require more computational resources and more time to complete (~4.5 h on 100 k Spike sequences using 16 cores), making them less suitable for a real-time system. Two interesting exceptions are the study by Rodriguez-Rivas and colleagues, in which they used DCA over multiple sequence alignments spanning a longer evolutionary timescale [46], which could be run as soon as the first SARS-CoV-2 genome sequence was available and required no subsequent update. The second study used a similar approach to what we have described in this study, that is, dividing sequences according to their collection date, but used an implementation of DCA that was able to scale up sufficiently [47]. We did not benchmark this implementation, but we note how selecting an appropriate threshold for "significant" interactions is as challenging as the MI-based approach we have followed. Filtering out positions with low diversity from the multiple sequence alignment might be a relatively simple strategy to reduce the computational requirement of DCA-based methods, although they could reduce their sensitivity, as emerging variants may be filtered out. Lastly, protein structure modeling has been used as a more direct way to measure the impact of combinations of mutations on the binding of the Spike protein to the ACE2 receptor [46], a method that is likely unable to scale to thousands of sequences.

We chose to use a method based on computing mutual information between pairs of positions for its balance between speed, ease of implementation, and precision/sensitivity. While MI-based methods are generally acknowledged to suffer from a number of biases such as those caused by uneven sampling or linkage between sites not actually participating in an epistatic interaction, they have been shown to be able to identify bona-fide interactions with comparable accuracy to pseudolikelihood methods, both in simulations [34, 37] and real sequence datasets. We have in fact shown how reasonably accurate reconstruction of known interactions can be obtained with as little as 10,000 sequences with a short computation time (~2 h to process 10,000 sequences). These characteristics make it an attractive approach to be integrated into automated systems

that feed on central sequence repositories. We introduced a sequence weighting strategy based on the age of each sequence so that a higher score would be given to emerging interactions and old ones be gradually removed. We showed how we were able to predict the known Spike 498/501 interaction as early as seven sequences encoding the double mutations were present in the dataset, thus demonstrating its potential use as an early warning system. We have in fact validated the predictive potential of this approach by extending our dataset beyond March 2023 and selecting three interactions that had emerged since the original cutoff. For those interactions, we chose six non-synonymous variants and tested their impact on infectivity and immune escape for the single and double mutants in two backgrounds, validating all interactions. Since we used an unsupervised method to identify interactions, we are not able to predict which particular phenotype might be affected and in which direction (e.g., higher or lower infectivity). However, reducing the number of variants to be tested would still be a very valuable tool, which could be combined with recent advances in high-throughput molecular assays [48] to further speed-up the characterization of emerging virus variants. We also note that interacting pairs estimated with a method based on mutual information could then be further filtered using a DCA-based method, which could reduce the false positive rate and simplify the dense parts of the interaction graph; the theoretical and technical details of such a two-step approach are however still to be properly defined.

One challenge of this genome-wide method is the interpretation of interactions predicted between genes [18]. The mechanism behind interactions within a single gene is intuitive; in the case of the Spike protein, many immune escape mutations destabilize the protein structure while providing a fitness advantage. In more general terms, changes in amino acid sequences might result in protein structure changes that might need to be compensated by changes in residues that are close to each other in the protein structure. In the case of an interaction between protein-coding genes, unless these proteins share an interaction interface, two possible explanations are left to explain the prediction of an interaction: a functional interaction or a false positive. We found it compelling that 3 between-gene interactions out of 7 that we flagged as high quality involved the known 681 Spike codon, one of which with the known 203 codon in the Nucleocapsid gene, whose mutations resulted in higher infectivity [44]. Given the recent report on the importance of intersegment (i.e., between genetic backgrounds) epistatic interactions in modulating the evolution of the hemagglutinin gene in the flu virus [49], we believe that these longer-range interactions are worth reporting and explored in further detail.

Conclusions

With this study, we have tested the applicability of a relatively simple computational method to estimate epistatic interactions from a large collection of viral sequences. By leveraging the metadata associated with each sample, notably the collection date, we could track the dynamics of these interactions, which can be used to filter out spurious predictions and prioritize other ones that appear over multiple subsequent time points. This approach is however heavily reliant on the quality of the metadata; as we have shown we could identify the interaction between the 498/501 spike codons with as little as 7 samples bearing the double mutation (equivalent to 86 effective sequences following our weighting scheme, Additional file 6: Fig. S3). During our earlier analysis, we noticed how we could predict this interaction

even earlier (December 2020, Additional file 6: Fig. S11), which puzzled us because at the time the Alpha variant, which only had mutations at the 501 codon was increasing in frequency. Upon careful inspection we discovered that 2 Omicron sequences out of the 2500 used in the December 2020 timepoint were erroneously dated, thus generating a modest signal. The analysis presented in the main text has been performed after a further round of filtering out sequences that are known to have incorrect metadata. This accident speaks in favor of the high sensitivity of the method, but also of the need for a well-curated sequence repository, which requires proper funding, infrastructure, and a community engaged in curating it [17]. The potential of such repositories in enabling real-time surveillance and interventions in the face of rapidly unfolding epidemics is well worth the effort.

Materials and methods

Dataset and estimation of epistatic interactions

We used a mutual information (MI) based method to estimate epistasis across multiple sequence alignments (MSAs) of SARS-CoV-2 sequences. This method is based on a previously developed software (spydrpick [33]). Our implementation, derived from that provided in the panaroo software [50], was optimized to process a very high number of sequences in a short amount of time.

We selected the whole pan-genome alignment available on the GISAID platform, counting 15,372,583 sequences by mid-April 2023. We only extracted the high-quality sequences according to a curated GISAID tree [51] which were also available in the public NCBI database (N=5,447,247). Also, since the non-coding $5'$ and $3'$ ends of the SARS-CoV-2 genomes are generally of lower quality, we resized them, considering only the positions from 266 to 29,768 (98.7% of the alignment length). Sequences were then deduplicated (N=4,093,019) and stored in a matrix file. Before computing the mutual information between positions, sequences were weighted according to their distance to the root of the phylogenetic tree, normalized by the number of leaves at each internal node. This in turn ensures that the contribution of a large number of very similar sequences to the final prediction is reduced.

The actual MI value was computed in the same way as the original implementation [33]. Briefly, mutual information is an extension of Shannon's entropy to two random variables, which are the two sites (X and Y) in the MSA being considered. Its value can be better understood as the reduction in uncertainty for one site (e.g., X) when given the value at the other site (e.g., Y). Mutual information can be computed between each pair of residues in a multiple sequence alignment, which can take one of 5 values for each sample, the four nucleotides plus a character for gaps or undetermined bases. The main equation is as in the original implementation, and as follows:

$$MI(X,Y) = \sum_{x \epsilon val(X)} \sum_{y \epsilon val(Y)} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

in which X and Y indicate the two positions in the multiple sequence alignment, $val(X)$ and $val(Y)$ the five possible discrete states at both positions, $p(x,y)$ the joint probability of $X = x$ and $Y = y$, and $p(x)$ and $p(y)$ the marginal probabilities. As described in the

original publication, the joint probability $\mathrm{p}(x, y)$ has to be estimated from the data itself, with the following equation:

$$p(x, y) = \frac{n_{eff}(x, y) + 0.5}{n_{eff} + r_X r_Y \cdot 0.5}$$

in which $n_{eff}(x, y)$ indicates the effective count of occurrences of the x and y pair (see below), n the number of samples, and $r_X = val(X)$, $r_Y = val(Y)$. Marginal probabilities are computed with equations:

$$p(x) = \sum\nolimits_{y \epsilon val(Y)} p(x, y)$$

and.

$$p(y) = \sum\nolimits_{x \epsilon val(X)} p(x, y)$$

Thus described, this mutual information strategy will likely be influenced by sampling biases, which could for instance arise when a particular lineage dominates the population. To overcome this problem, and as indicated above, we introduced a sample weighting strategy similar to that used in the implementation that is present in the panaroo software package [50], and based on the phylogenetic tree of all samples. For each sample i, we computed the weight w_i as the distance from the root, with a normalization step such that the length of each branch is normalized by the number of leaves downstream of each branch:

$$w_i = \sum\nolimits_{b=1}^{t} \frac{len(b)}{n_b}$$

In which b is the internal branch over the total number t, $len(b)$ the length of the branch, and n_b the number of downstream leaves. The weights are used to derive $n_{eff}(x, y)$, which is the count of occurrences of the x and y pair multiplied by the weights.

We ran our implementation of the spydrpick algorithm on a random sample of 100 positions in the MSA to extract a MI threshold based on the 90th percentile of the computed MI values. The resulting interactions were then filtered by an algorithm implemented for inferring gene expression networks (ARACNE [39]) and used to retain direct interactions, discarding indirect ones with a lower MI value. The output is generated as a tab-separated values (tsv) file. We defined 4 thresholds to indicate our confidence in the computed MI values based on the Tukey method as follows:

$$O_n = Q3 + n * (Q3 - Q1)$$

Q3 and Q1 respectively indicate the upper and lower quartiles of the MI values, and n are four different coefficients ($n = \{1.5, 3, 6, 12\}$) to be multiplied to the interquartile range, one for each outlier level. The resulting output with MI values was then annotated using the SARS-CoV2 GFF file (RefSeq NC_045512.2) and each pair of positions was associated to the respective gene, codon number, and gene relative codon number. We excluded interactions within the same codon and between adjacent codons, as well as interactions between different genes whose nucleotide distance was < 2. For the

overall analysis (Fig. 1, Additional file 1: Table S1) we ran the pipeline on the whole dataset of public sequences up to March 2023. A total number of 4,093,019 deduplicated and weighted sequences were processed. Smaller subsets of 1000, 10,000, 100,000, and 1,000,000 sequences were generated by drawing random sequences from the complete dataset.

In silico validation of predictions

We validated the interactions using a set of experimentally validated epistatic interactions and notable positions in the RBD of the Spike protein. We divided them into three sets, based on their known impact on viral fitness; we termed codons 498 and 501 as "Affinity" mutations for their positive epistatic interaction resulting in increased affinity to the ACE2 receptor [21]. We termed codons 406, 417, 446, 447, 449, 484, 493, 496, 505, and 506 as "Escape" mutations for their contribution to immune escape, especially in the 498/501 genetic background [20, 21]. We added to this category those Spike codons with a relatively high escape score (>0.1) as computed by the Escape Calculator [52]. Lastly, we added other notable Spike codons based on their designation as mutation of interest or mutations of concern if they were not already included in the other two sets: 18, 439, 452, 477, 494, and 681 [38].

We used interactions between the codons in all three sets if they were in the RBD region ($318 < \text{codon} < 541$) and three different methods to validate our predicted interactions. The first method is based on a hypergeometric test (Fisher's test) for the enrichment of observed known interactions over all possible RBD interactions. We defined the universe of all possible interactions as those between residues mutated in at least one sequence of those used to build the predictions. We then compared the odds ratio thus computed against that from 1000 permutations of the actual RBD network, avoiding self links (a position interacting with itself) and those between adjacent codons. For the second validation method we built a binary classifier to indicate whether the interactions passing the first MI value threshold (90th percentile of MI values computed over 100 random positions) were the known ones or not. For each outlier level (O1, O2, O3, O4) we used its threshold value to classify the interactions and computed the F1 score using the known interactions as a truth set. For the third validation, we used the inferred pairwise epistatic coefficients between 15 BA.1 mutations and tested the ability of our predicted interactions to predict interactions with a coefficient >0.15, using the outlier levels as thresholds. For all three validation methods, we measured the 95% confidence interval of each indicator (i.e., odds ratio, specificity, sensitivity, F1 score) through bootstrapping with $N=1000$.

Time-resolved analysis

We also tested the accuracy of our method on different time-points along a roughly 3-year period (December 2019 to March 2023). To this purpose, we divided our dataset into different subsets of sequences for every month as follows. Firstly, we binned the sequences for each month, collapsing the period between December 2019 and February 2020, since the number of sequences collected in the databases was very low at the very beginning of the pandemic ($N=592$), obtaining a total of 38 bins. Furthermore, we randomly selected 2500 sequences for each month filtering out those known to have an

erroneous sample collection date [53]. For each subset, we then created a MSA with the selected sequences at that specific time point plus all the previously selected sequences. Moreover, beyond the phylogenetic weights, a second time-based weighting system was applied to the sequences in order to maximize the importance of emerging interactions. This has been made to reduce the MI values of less recent interactions. For this purpose, we developed a function of exponential decay that follows a Hill curve function for the weighting of each sequence. We used a Hill coefficient of 3 and a weight of 0.5 at 120 days (Additional file 6: Fig. S2). Both phylogenetic and time-based weights were then combined via multiplication to generate a final weight w_i for each sequence in each subset (Additional file 6: Fig. S3). We used Nextclade [54] v2.14.0 to generate the list of amino acid substitutions in each sequence as well as their lineage, and used the pyfish package [55] v1.0.3 to generate a Muller plot to represent the changes in lineage relative proportions across all subsets.

Comparison with direct coupling analysis (DCA)

To further validate our data, we also used a pseudolikelihood method called DCA (Direct Coupling Analysis), and implemented in the plmc tool [26] that calculates the covariation and coevolution of biological sequences by inferring undirected graphical models, and applies a pseudo-likelihood approximation (Potts model) to impute the interaction strength between all pairwise positions in a given sequence. Since this implementation scales poorly with large MSA, both in terms of length and depth (i.e., number of sequences), we used the spike portion of the MSA to compare values computed using plmc against MI values for all the time-resolved datasets. We used the following command line arguments when running plmc (commit 18c9e55): "–fast -m 20 -le 20.0 -lh 0.01 -a -AGCT". To have a more direct comparison for Fig. 3C, we did not apply the time-based weighting for MI values but only that based on phylogenetic distances. For Additional file 6: Fig. S4, we used the weighting scheme including the time-based correction. In order to compare the ability of the DCA-based method to identify known Spike RBD interactions [21], we computed outlier levels for the output of the plmc implementation in a different manner from the MI-based method. We first selected the 1000 top scoring interactions, and then computed the four O_n levels using the following quartiles of the score distribution: $n = \{0,25,50,75\}$. We followed this approach as it allowed us to obtain a comparable number of predicted interactions, which would allow for a fair comparison between the two methods.

Plasmid cloning of selected SARS-CoV-2 point mutations

For the generation of the expression plasmids for SARS-CoV-2 point mutations, site-directed mutagenesis was performed on a previously described full-length human-codon-optimized plasmid pCG1_SARS-2-Sdel18 encoding the spike protein of the Wuhan-Hu-1 SARS-CoV-2 (18 amino acid truncation at the C-terminus) [56]. Plasmids encoding the selected spike protein mutations were integrated using the Q5 Site-Directed Mutagenesis Kit (NEB #E0554S) according to the manufacturer's instructions. Successful introduction of the respective mutation was verified by sequence analysis using a commercial sequencing service (Microsynth Seqlab). Mutagenic primers listed in Additional file 12: Table S11 were designed to flank sequences of the respective target

regions to introduce up to two selected mutations per reaction. Primers were designed using the "NEBase Changer" tool (https://nebasechanger.neb.com/) which also provided primer-specific annealing temperatures. PCR amplification was performed using Q5 High-Fidelity DNA Polymerase. The PCR program included denaturation at 98 °C for 30 s, annealing at primer-specific temperatures for 30 s, and extension at 72 °C for 2 min. The PCR was followed by DpnI treatment at room temperature for 15 min and heat inactivation of the enzyme at 80 °C for 20 min to remove the parental template. The mutagenized PCR products were then transformed into *Escherichia coli* DH5 alpha cells in the presence of ampicillin. Positive colonies were identified and successful introduction of the respective mutation was verified by sequence analysis using a commercial sequencing service (Microsynth Seqlab).

Generation of VSV-SARS-CoV-2 mutant pseudotypes

Vesicular stomatitis virus (VSV, Indiana strain) pseudotypes harboring the SARS-CoV-2 WT spike or the spike protein-coding for the different point mutations were generated as previously described [57]. Briefly, HEK-293 T cells maintained in DMEM (Capricorn Scientific DMEM-HXA) supplemented with 10% FBS (Capricorn Scientific FBS-11A) and 2 mM L-glutamine (Thermo Fisher 25,030,081) were transfected with different SARS-CoV-2 mutant spike protein expression plasmids using polyethylenimine Max (polysciences 24,765). Twenty-eight hours post-transfection, cells were transduced with a replication-deficient VSV expressing the enhanced green fluorescent protein (eGFP) cassette, in lieu of the VSV glycoprotein G open reading frame (ORF) (VSV*ΔG-GFP) at a multiplicity of infection (MOI) of 3 at 37 °C and 5% CO_2 for 2 h. Cells were then washed with PBS, and a fresh medium with an anti-VSV-G antibody was added to neutralize residual VSV*ΔG-GFP. The supernatant containing the WT VSV-SARS-CoV-2-S or the mutant pseudotypes was collected 18 h post-transduction, pooled, and centrifuged at $300 \times g$ for 6 min to remove cellular debris. Supernatant was then concentrated using 100 kD Amicon® Ultra-15 centrifugal filters (Millipore UFC910024), by centrifugation at $2000 \times g$ for 10 min. Concentrated pseudotypes were aliquoted and stored at − 80 °C until further use. Pseudo-virus was titrated on Vero-B4 cells to quantify the amount of infectious virus to normalize viral input of the generated mutants between assays.

VSV-SARS-CoV-2-S pseudovirus titration

To determine the concentration of infectious VSV-SARS-CoV-2-S pseudovirus particles, the virus stock was initially tenfold diluted in MEM (5% FBS, 1% Glutamax) in a 96-well plate. This was followed by a threefold serial dilution for a total of 11 dilutions each in 3 replicates. A control well without pseudovirus was also included. The dilutions were then added to a separate 96-well plate containing a monolayer of Vero B4 cells and incubated at 37 °C and 5% CO_2 for 24 h. Infected cells were identified and quantified for positive GFP signal by full-well widefield microscopy using the Olympus FV3000 microscope CellSens software. The number of GFP-positive cells for each dilution was quantified by ImageJ2 version 2.9.0. The virus titer was calculated by the mean number of the replicates based on the dilution factor and expressed as focal forming units per milliliter (ffu/ml). SARS-CoV-2 pseudovirus infectivity was measured by threefold serial

dilution of each pseudo-virus in triplicates starting from 6000 ffu/ml over 6 dilution steps in MEM (5% FBS, 1% Glutamax) in a 96-well plate. The respective pseudovirus dilutions were then added to Vero-B4 cell monolayer and incubated at 37 °C and 5% CO_2 for 24 h. Infected cells were identified and quantified for positive GFP signal by full-well widefield microscopy using the Olympus FV3000 microscope CellSens software. GFP-positive cells were quantified by ImageJ2 [58] version 2.9.0. Three consecutive dilution columns with GFP-positive cells ranging between 50 and 2000 were selected and averaged to quantify the mutant pseudovirus infectivity (ffu/ml). The pseudo-virus titer was then calculated by the mean number of ffu/ml from all three replicates of each pseudovirus dilution.

VSV-SARS-CoV-2-S pseudotypes spike dependent viral entry

To determine virus entry into the cells, SARS-CoV-2 pseudovirus infectivity was measured by threefold serial dilution of each pseudo-virus in triplicates starting from an initial concentration of 6000 ffu/ml over 6 dilution steps in MEM (5% FBS, 1% Glutamax) in a 96-well plate. The respective pseudovirus dilutions were then added to the Vero-B4 cell monolayer and incubated at 37 °C and 5% CO_2 for 24 h. Infected cells were identified under full-well widefield microscopy using the Olympus FV3000 microscope CellSens software to quantify GFP expression as a marker of viral entry. GFP-positive cells were quantified by ImageJ2 [58] version 2.9.0. Three consecutive dilution columns with GFP-positive cells ranging between 50 and 2000 were selected and averaged to quantify the mutant pseudovirus infectivity (ffu/ml).

VSV-SARS-CoV-2-S pseudovirus neutralization assay

The commercially available monoclonal antibody (mAb) against SARS-CoV-2, Imdevimab (REGN10987) was threefold serially diluted in MEM (Capricorn Scientific MEM-XA) supplemented with 5% FBS and 2 mM L-glutamine in triplicates starting from 4 µg/ml over 11 dilution steps. An equal volume of pseudovirus was then added to the antibody dilutions at a final concentration of 300 ffu/well. Wells containing pseudovirus in the absence of the monoclonal antibody were included as a control. After 1 h of virus-antibody incubation at 37 °C, 100 µl of the pseudo-virus/antibody mixture was transferred to the Vero-B4 cell monolayer and incubated at 37 °C for 24 h. GFP-positive infected cells were identified by full-well widefield microscopy and the number of GFP-positive infected cells per well was quantified by ImageJ2. Triplicate values of the GFP-positive infected cells were averaged and normalized relative to the control wells containing the virus in the absence of mAb.

Identification of epistatic interactions from pseudovirus assays

We transformed the viral entry data into normalized viral particle counts and computed the fold-change with respect to the wild-type pseudovirus for all tested variants, which we used as a measure of the impact of each variant set on viral infectivity. We similarly converted the antibodies neutralization curves to a normalized area under the curve (AUC) value, with 1 representing a complete failure in viral neutralization, which we used as a measure of the impact of each variant set on immune

escape. In order to formally test for epistatic interactions based on the data from the pseudovirus assays, we fitted two ordinary least squares models for each tested pair:

$$Y = V_a + V_b + V_a \cdot V_b + \epsilon \text{ and } Y = V_a + V_b + \epsilon$$

where Y represents the value of the target phenotype (i.e., infectivity log2 fold-changes or immune escape AUC), V_a and V_b the presence/absence vector of the target variant at site a and b, respectively, $V_a \cdot V_b$ the interaction term between the two sites, and ϵ the biological replicate. We then used a likelihood ratio test to compare the two models and derive a p-value. We applied the same approach for the variants tested in the Q498R/N501Y background with two slightly modified ordinary least squares models:

$$Y = V_a + V_b + V_a \cdot V_b + V_{Q498R} + V_{N501Y} + V_{Q498R} \cdot V_{N501Y} + \epsilon$$

$$Y = V_a + V_b + V_{Q498R} + V_{N501Y} + V_{Q498R} \cdot V_{N501Y} + \epsilon$$

where V_{Q498R} and V_{N501Y} represent the presence/absence vector of the target variant at sites 498 and 501, respectively, and $V_{Q498R} \cdot V_{N501Y}$ the interaction term between the two sites.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03355-y.

Additional file 1. Table S1: predicted interactions from the complete dataset.

Additional file 2. Table S2: predicted interactions for the subsets with size 1,000. Columns are the same as those in Supplementary Table 1.

Additional file 3. Table S3: predicted interactions for the subsets with size 10,000. Columns are the same as those in Supplementary Table 1.

Additional file 4. Table S4: predicted interactions for the subsets with size 100,000. Columns are the same as those in Supplementary Table 1.

Additional file 5. Table S5: predicted interactions for the subsets with size 1,000,000. Columns are the same as those in Supplementary Table 1.

Additional file 6. Supplementary figures S1-11.

Additional file 7. Table S6: predicted interactions from the March 2023 dataset, without the time weighting correction.

Additional file 8. Table S7: predicted interactions for the spike gene using the plmc implementation of the pseudo-likelihood method; positions are relative to the first base of the spike codon, and all pairwise interactions are scored.

Additional file 9. Table S8: predicted interactions for the time subsets, from March 2020 until March 2023; columns are the same as those in Table S7.

Additional file 10. Table S9: enrichment of known interactions in each time subset. For each month between March 2020 and March 2023 the enrichment value for the actual dataset is reported, as well as the average values for 1,000 randomized RBD interaction networks.

Additional file 11. Table S10: infectivity and antibody neutralization data for a selected number of mutants.

Additional file 12. Table S11: list of mutants and primers used in this study.

Additional file 13. Review history.

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We thank Theo Sanderson and Chris Lauber for their feedback on the manuscript draft.

Review history

This article was first peer-reviewed at Review Commons and reviewer reports are available online [59, 60] with the authors' response also available [61]. The rest of the review history containing additional reviewer comments and further authors' responses is available as Additional File 13

Peer review information
Andrew Cosgrove was the primary editor of this article at Genome Biology and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
MG conceived the study, GI and MG wrote the analysis code, performed the analysis, assembled figures, and wrote the manuscript. HJ and MK generated the plasmids encoding for the Spike RBD mutants. MO and BC tested the Spike RBD mutants and analyzed the results. MG, LCS, and UK obtained funding to support the work. All authors contributed edits to the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. MG and GI were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2155—project number 390874280. GI was further supported by a FEMS Research and Training Grant (FEMS-GO-2020–258). UK was funded by the German Centre for Infection Research (DZIF)—Project ID 8003901903. MO was supported by the Hannover Biomedical Research School (HBRS) and the Center for Infection Biology (ZIB). This research was funded by the Ministry of Science and Culture of Lower Saxony, through the grant COFONI, flex fund project 10FF22 (PREMUS). Funders had no role in the study design and in the presentation of the results.

Availability of data and materials
All code used to compute MI values and run all the other analyses is available in the following code repository released under a permissive open-source license (MIT): https://github.com/microbial-pangenomes-lab/2022_sarscov2_epistasis. The version of the code used to generate all the results presented here has been deposited in Zenodo [62]. The code is based on bash scripts running a series of python scripts, using the following libraries: numpy [63] v1.24.4, scipy [64] v1.11.1, pandas [65] v2.0.3, numba [66] v0.57.1, treeswift [67] v1.11.37, dendropy [68] v4.6.1, networkx [69] v3.1, scikit-learn [70] v1.3.0, statsmodels v0.13.2, matplotlib [71] v3.7.2, seaborn [72] v0.12.2, and jupyter-lab [73] v4.0.4.
We downloaded the input MSA and relative metadata and phylogenetic tree from the GISAID database [51] on April 14th. We have restricted our analysis strictly to those sequences that were also deposited in public databases, as indicated at the following URL: https://hgwdev.gi.ucsc.edu/~angie/epiToPublicAndDate.latest [74]. As such we believe we have respected the stringent end user agreement imposed by GISAID. Filtering of samples with an erroneous sampling date was performed using a publicly available list of problematic GISAID IDs, available as part of the covariants repository: https://github.com/hodcroftlab/covariants [75]. Escape mutations in the Spike's RBD were partly derived from the SARS-CoV-2 escape calculator: https://github.com/jbloomlab/SARS2-RBD-escape-calc [52].

Declarations

Ethics approval and consent to participate
No ethical approval was required for this work, and since no human subject was included, there was also no need to implement a consent to participate procedure.

Consent for publication
All authors have read the final version of the manuscript and have given their explicit consent for publication.

Competing interests
All authors have no competing interests to report.

Author details
[1]Institute for Molecular Bacteriology, TWINCORE Centre for Experimental and Clinical Infection Research, a joint venture between the Hannover Medical School (MHH) and the Helmholtz Centre for Infection Research (HZI), Hannover, Germany. [2]Cluster of Excellence RESIST (EXC 2155), Hannover Medical School (MHH), Hannover, Germany. [3]Center for Cancer Research, Medical University of Vienna, Vienna, Austria. [4]Institute for Experimental Infection Research, TWINCORE Centre for Experimental and Clinical Infection Research, a joint venture between the Hannover Medical School (MHH) and the Helmholtz Centre for Infection Research (HZI), Hannover, Germany. [5]Helmholtz Centre for Infection Research, Department of Viral Immunology (VIRI), Brunswick, Germany. [6]Centre for Individualized Infection Medicine (CiiM) a Joint Venture of Helmholtz Centre for Infection Research and Hannover Medical School, Hannover, Germany.

Received: 2 May 2024 Accepted: 26 July 2024
Published online: 22 August 2024

References

1. Vöhringer HS, Sanderson T, Sinnott M, De Maio N, Nguyen T, Goater R, et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. Nature. 2021;600:506–11. https://doi.org/10.1038/s41586-021-04069-y.
2. Kevadiya BD, Machhi J, Herskovitz J, Oleynikov MD, Blomberg WR, Bajwa N, et al. Diagnostics for SARS-CoV-2 infections. Nat Mater. 2021;20:593–605. https://doi.org/10.1038/s41563-020-00906-z.
3. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, et al. Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A. 2020;117:11727–34. https://doi.org/10.1073/pnas.2003138117.
4. Fu Y, Cheng Y, Wu Y. Understanding SARS-CoV-2-mediated inflammatory responses: from mechanisms to potential therapeutic tools. Virol Sin. 2020;35:266–71. https://doi.org/10.1007/s12250-020-00207-4.

5. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell. 2020;182:1284–1294.e9. https://doi.org/10.1016/j.cell.2020.07.012.
6. Chen J, Wang R, Wang M, Wei G-W. Mutations strengthened SARS-CoV-2 infectivity. J Mol Biol. 2020;432:5212–26. https://doi.org/10.1016/j.jmb.2020.07.009.
7. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021;19:409–24. https://doi.org/10.1038/s41579-021-00573-0.
8. Gobeil SM-C, Janowska K, McDowell S, Mansouri K, Parks R, Stalls V, et al. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. Science. 2021;373:eabi6226. https://doi.org/10.1126/science.abi6226.
9. Chakraborty S. Evolutionary and structural analysis elucidates mutations on SARS-CoV2 spike protein with altered human ACE2 binding affinity. Biochem Biophys Res Commun. 2021;534:374–80. https://doi.org/10.1016/j.bbrc.2020.11.075.
10. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121–3. https://doi.org/10.1093/bioinformatics/bty407.
11. Singh J, Rahman SA, Ehtesham NZ, Hira S, Hasnain SE. SARS-CoV-2 variants of concern are emerging in India. Nat Med. 2021;27:1131–3. https://doi.org/10.1038/s41591-021-01397-4.
12. Greaney AJ, Starr TN, Barnes CO, Weisblum Y, Schmidt F, Caskey M, et al. Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. Nat Commun. 2021;12:4196. https://doi.org/10.1038/s41467-021-24435-8.
13. Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science. 2022;376:1327–32. https://doi.org/10.1126/science.abm1208.
14. Bloom JD, Neher RA. Fitness effects of mutations to SARS-CoV-2 proteins 2023:2023.01.30.526314. https://doi.org/10.1101/2023.01.30.526314.
15. Beguir K, Skwark MJ, Fu Y, Pierrot T, Carranza NL, Laterre A, et al. Early computational detection of potential high-risk SARS-CoV-2 variants. Comput Biol Med. 2023;155:106618. https://doi.org/10.1016/j.compbiomed.2023.106618.
16. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. N Engl J Med. 2020;383:2603–15. https://doi.org/10.1056/NEJMoa2034577.
17. Ladner JT, Sahl JW. Towards a post-pandemic future for global pathogen genome sequencing. Plos Biol. 2023;21:e3002225. https://doi.org/10.1371/journal.pbio.3002225.
18. Lehner B. Molecular mechanisms of epistasis within and between genes. Trends Genet TIG. 2011;27:323–31. https://doi.org/10.1016/j.tig.2011.05.007.
19. Park Y, Metzger BPH, Thornton JW. Epistatic drift causes gradual decay of predictability in protein evolution. Science. 2022;376:823–30. https://doi.org/10.1126/science.abn6895.
20. Starr TN, Greaney AJ, Hannon WW, Loes AN, Hauser K, Dillen JR, et al. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. Science. 2022;377:420–4. https://doi.org/10.1126/science.abo7896.
21. Moulana A, Dupic T, Phillips AM, Chang J, Nieves S, Roffler AA, et al. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA1. Nat Commun. 2022;13:7011. https://doi.org/10.1038/s41467-022-34506-z.
22. Diaz-Colunga J, Skwara A, Gowda K, Diaz-Uriarte R, Tikhonov M, Bajic D, et al. Global epistasis on fitness landscapes. Philos Trans R Soc B Biol Sci. 2023;378:20220053. https://doi.org/10.1098/rstb.2022.0053.
23. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc Natl Acad Sci. 2011;108:11530–5. https://doi.org/10.1073/pnas.1105315108.
24. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat Commun. 2016;7:11660. https://doi.org/10.1038/ncomms11660.
25. Sanderson T. Taxonium, a web-based tool for exploring large phylogenetic trees. eLife. 2022;11:e82392. https://doi.org/10.7554/eLife.82392.
26. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35:128. https://doi.org/10.1038/nbt.3769.
27. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Phys Rev E. 2013;87:012707. https://doi.org/10.1103/PhysRevE.87.012707.
28. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families. Proteins Struct Funct Bioinforma. 2011;79:1061–78. https://doi.org/10.1002/prot.22934.
29. Neverov AD, Fedonin G, Popova A, Bykova D, Bazykin G. Coordinated evolution at amino acid sites of SARS-CoV-2 spike. eLife. 2023;12:e82516. https://doi.org/10.7554/eLife.82516.
30. Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. Prevalence of epistasis in the evolution of influenza a surface proteins. Plos Genet. 2011;7:e1001301. https://doi.org/10.1371/journal.pgen.1001301.
31. Neverov AD, Popova AV, Fedonin GG, Cheremukhin EA, Klink GV, Bazykin GA. Episodic evolution of coadapted sets of amino acid sites in mitochondrial proteins. Plos Genet. 2021;17:e1008711. https://doi.org/10.1371/journal.pgen.1008711.
32. Neverov AD, Kryazhimskiy S, Plotkin JB, Bazykin GA. Coordinated evolution of influenza a surface proteins. Plos Genet. 2015;11:e1005404. https://doi.org/10.1371/journal.pgen.1005404.
33. Pensar J, Puranen S, Arnold B, MacAlasdair N, Kuronen J, Tonkin-Hill G, et al. Genome-wide epistasis and co-selection study using mutual information. Nucleic Acids Res. 2019;47:e112. https://doi.org/10.1093/nar/gkz656.
34. Pensar J, Xu Y, Puranen S, Pesonen M, Kabashima Y, Corander J. High-dimensional structure learning of binary pairwise Markov networks: a comparative numerical study. Comput Stat Data Anal. 2020;141:62–76. https://doi.org/10.1016/j.csda.2019.06.012.
35. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics. 2005;21:4116–24. https://doi.org/10.1093/bioinformatics/bti671.

36. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24:333–40. https://doi.org/10.1093/bioinformatics/btm604.
37. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY, et al. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. Plos Genet. 2017;13:e1006508. https://doi.org/10.1371/journal.pgen.1006508.
38. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. Nat Methods. 2023;20:512–22. https://doi.org/10.1038/s41592-023-01769-3.
39. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7:S7. https://doi.org/10.1186/1471-2105-7-S1-S7.
40. Han P, Li L, Liu S, Wang Q, Zhang D, Xu Z, et al. Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2. Cell. 2022;185:630–640.e10. https://doi.org/10.1016/j.cell.2022.01.001.
41. Lista MJ, Winstone H, Wilson HD, Dyer A, Pickering S, Galao RP, et al. The P681H mutation in the spike glycoprotein of the alpha variant of SARS-CoV-2 escapes IFITM restriction and is necessary for type I interferon resistance. J Virol. 2022;96:e0125022. https://doi.org/10.1128/jvi.01250-22.
42. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 2020;182:1295–1310.e20. https://doi.org/10.1016/j.cell.2020.08.012.
43. Dadonaite B, Crawford KHD, Radford CE, Farrell AG, Yu TC, Hannon WW, et al. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. Cell. 2023;186:1263-1278.e20. https://doi.org/10.1016/j.cell.2023.02.001.
44. Wu H, Xing N, Meng K, Fu B, Xue W, Dong P, et al. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell Host Microbe. 2021;29:1788–1801.e6. https://doi.org/10.1016/j.chom.2021.11.005.
45. Zeng H-L, Dichio V, Rodríguez Horta E, Thorell K, Aurell E. Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes. Proc Natl Acad Sci. 2020;117:31519–26. https://doi.org/10.1073/pnas.2012331117.
46. Rodriguez-Rivas J, Croce G, Muscat M, Weigt M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. Proc Natl Acad Sci. 2022;119:e2113118119. https://doi.org/10.1073/pnas.2113118119.
47. Zeng H-L, Liu Y, Dichio V, Aurell E. Temporal epistasis inference from more than 3 500 000 SARS-CoV-2 genomic sequences. Phys Rev E. 2022;106:044409. https://doi.org/10.1103/PhysRevE.106.044409.
48. Loes AN, Tarabi RAL, Huddleston J, Touyon L, Wong SS, Cheng SMS, et al. High-throughput sequencing-based neutralization assay reveals how repeated vaccinations impact titers to recent human H1N1 influenza strains 2024:2024.03.08.584176. https://doi.org/10.1101/2024.03.08.584176.
49. Liu T, Wang Y, Tan TJC, Wu NC, Brooke CB. The evolutionary potential of the influenza A virus hemagglutinin is highly constrained by intersegment epistasis 2022:2022.05.19.492711. https://doi.org/10.1101/2022.05.19.492711.
50. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21:180. https://doi.org/10.1186/s13059-020-02090-4.
51. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAID's role in pandemic response. China CDC Wkly. 2021;3:1049–51. https://doi.org/10.46234/ccdcw2021.255.
52. Greaney AJ, Starr TN, Bloom JD. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. Virus Evol. 2022;8:veac021. https://doi.org/10.1093/ve/veac021.
53. Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. 2021.
54. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6:3773. https://doi.org/10.21105/joss.03773.
55. Streck A, Kaufmann TL, Schwarz RF. SMITH: spatially constrained stochastic model for simulation of intra-tumour heterogeneity. Bioinforma Oxf Engl. 2023;39:btad102. https://doi.org/10.1093/bioinformatics/btad102.
56. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell. 2020;181:271–280.e8. https://doi.org/10.1016/j.cell.2020.02.052.
57. Rentsch MB, Zimmer G. A vesicular stomatitis virus replicon-based bioassay for the rapid and sensitive determination of multi-species type I interferon. Plos One. 2011;6:e25858. https://doi.org/10.1371/journal.pone.0025858.
58. Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET, et al. Image J2: ImageJ for the next generation of scientific image data. BMC Bioinformatics. 2017;18:529. https://doi.org/10.1186/s12859-017-1934-z.
59. Review Commons Report 1. Early Evidence Base. 2024. https://doi.org/10.15252/rc.2024921543
60. Review Commons Report 2. Early Evidence Base. 2024. https://doi.org/10.15252/rc.2024745734
61. Review Commons Response. Early Evidence Base. 2024. https://doi.org/10.15252/rc.2024066833
62. Galardini M, Innocenti G. microbial-pangenomes-lab/2022_sarscov2_epistasis: Manuscript version 2024. https://doi.org/10.5281/zenodo.12731178.
63. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.
64. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
65. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference 2010:56–61. https://doi.org/10.25080/Majora-92bf1922-00a.
66. Lam SK, Pitrou A, Seibert S. Numba: a LLVM-based Python JIT compiler. Proc. Second Workshop LLVM Compil. Infrastruct. HPC, New York, NY, USA: Association for Computing Machinery; 2015, p. 1–6. https://doi.org/10.1145/2833157.2833162.

67. Moshiri N. TreeSwift: a massively scalable python tree package. SoftwareX. 2020;11:100436. https://doi.org/10.1016/j.softx.2020.100436.
68. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 2010;26:1569–71.
69. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference, Pasadena, CA USA: 2008, p. 11–5.
70. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
71. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9:99–104. https://doi.org/10.1109/MCSE.2007.55.
72. Waskom ML. seaborn: statistical data visualization. J Open Source Softw. 2021;6:3021. https://doi.org/10.21105/joss.03021.
73. Perez F, Granger BE. IPython: a system for interactive scientific computing. Comput Sci Eng. 2007;9:21–9.
74. Hinrichs AS. Epi to Public and Date table 2024. https://hgwdev.gi.ucsc.edu/~angie/epiToPublicAndDate.latest Accessed 12 July 2024.
75. Hodcroft E. covariants. GitHub 2024. https://github.com/hodcroftlab/covariants Accessed 12 July 2024.

Publisher's Note