High-throughput identification of prefusion-stabilizing mutations in SARS-CoV-2 spike

Timothy J.C. Tan[1], Zongjun Mou[2], Ruipeng Lei[3], Wenhao O. Ouyang[3], Meng Yuan[4], Ge Song[5,6,7], Raiees Andrabi[5,6,7], Ian A. Wilson[4,6,7,8], Collin Kieffer[9], Xinghong Dai[2], Kenneth A. Matreyek[10], Nicholas C. Wu[1,3,11,12,§]

[1] Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2] Department of Physiology and Biophysics, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

[3] Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[4] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

[5] Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA

[6] IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, CA 92037, USA

[7] Consortium for HIV/AIDS Vaccine Development (CHAVD), The Scripps Research Institute, La Jolla, CA 92037, USA

[8] The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

[9] Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[10] Department of Pathology, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

[11] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[12] Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[§] Correspondence: nicwu@illinois.edu (N.C.W.)

ABSTRACT

Designing prefusion-stabilized SARS-CoV-2 spike is critical for the effectiveness of COVID-19 vaccines. All COVID-19 vaccines in the US encode spike with K986P/V987P mutations to stabilize its prefusion conformation. However, contemporary methods on engineering prefusion-stabilized spike immunogens involve tedious experimental work and heavily rely on structural information. Here, we established a systematic and unbiased method of identifying mutations that concomitantly improve expression and stabilize the prefusion conformation of the SARS-CoV-2 spike. Our method integrated a fluorescence-based fusion assay, mammalian cell display technology, and deep mutational scanning. As a proof-of-concept, this method was applied to a region in the S2 domain that includes the first heptad repeat and central helix. Our results revealed that besides K986P and V987P, several mutations simultaneously improved expression and significantly lowered the fusogenicity of the spike. As prefusion stabilization is a common challenge for viral immunogen design, this work will help accelerate vaccine development against different viruses.

INTRODUCTION

SARS-CoV-2 spike (S) glycoprotein, a homotrimeric class I fusion protein, naturally exists in a metastable, prefusion conformation on the virion surface[1]. Once the receptor-binding domain (RBD) of S transitions to an 'up' state and binds to the human angiotensin-converting enzyme II (hACE2) receptor[2-4], a cascade of conformational changes is triggered to promote virus-host membrane fusion, and hence virus entry[1,5-8]. This conformational change, which involves structural rearrangement of the first heptad repeat (HR1) and central helix (CH), as well as the shedding of the S1 subunit, converts S into the postfusion conformation[5-10]. To inhibit virus entry and fusion, neutralizing antibodies target a variety of mainly conformational epitopes on the prefusion conformation of S[11-15]. Many of these conformational epitopes disappear or rearrange in the postfusion conformation, which instead can expose non-neutralizing epitopes that are immunodominant[1]. Consistently, antibody titer to the prefusion conformation has a strong correlation with neutralization potency, whereas that to the postfusion conformation does not[16]. Therefore, effective COVID-19 vaccines require S to be locked in the prefusion conformation to preserve the neutralizing epitopes.

The rapid development of prefusion-stabilized SARS-CoV-2 S during the early phase of COVID-19 pandemic has tremendously benefited from prior studies on prefusion-stabilizing mutations in the S proteins of related betacoronaviruses, namely MERS-CoV[17,18] and SARS-CoV[18]. These studies employed a structure-based approach to identify two prefusion-stabilizing mutations (K986P/V987P, SARS-CoV-2 numbering) at the HR1-CH junction[17-19]. Due to the structural similarities among the S proteins of MERS-CoV, SARS-CoV, and SARS-CoV-2, K986P/V987P were directly applied to engineer the prefusion-stabilized SARS-CoV-2 S during COVID-19 vaccine development. For example, K986P/V987P are included in many nucleic acid- and protein subunit-based COVID-19 vaccines, such as those from Moderna[20], Pfizer-BioNTech[21], Johnson & Johnson-Janssen[22], and Novavax[23]. Subsequent studies, which also used a structure-based approach, identified additional mutations that further improve the expression and prefusion stability of SARS-CoV-2 S[24-27]. Nevertheless, identifying prefusion-stabilizing mutations using structure-based approach is time-consuming and likely not comprehensive, because it relies on low-throughput characterization of individual candidate mutants.

In this study, we developed a method to identify prefusion-stabilizing mutations of SARS-CoV-2 S in a high-throughput and systematic manner, by coupling a fluorescence-based fusion assay, mammalian cell display technology, and deep mutational scanning (DMS). As a proof-of-concept, we screened all possible amino-acid mutations across the entire region spanning HR1 and CH. In addition to the K986P and V987P that are used in current COVID-19 vaccines, we identified several mutations that simultaneously improved expression and stabilized the prefusion conformation of both membrane-bound and soluble S. In this regard, our method circumvents the limitations of using structure-based approaches to engineer prefusion-stabilized S immunogens.

RESULTS

Establishing a high-throughput fusion assay for SARS-CoV-2 S

High-throughput assays for measuring protein mutant expression level in human cells have been developed in previous studies by one of our authors using landing pad cells[28-30], which enable one cell to express one mutant, thereby providing a genotype-phenotype linkage[31,32]. Such assays have also been adopted to study the impact of N-terminal domain (NTD) mutations on SARS-CoV-2 S expression[33]. However, there is no similar assay for measuring fusogenicity. Conventional approaches for quantifying fusogenicity often rely on split fluorescent protein systems[34-39], such as the split GFP system that consists of GFP_{1-10} and GFP_{11}[40]. For example, when cells that express hACE2 and GFP_{1-10} are mixed with cells expressing SARS-CoV-2 S and GFP_{11}, fusion occurs, and the resultant syncytia fluoresce green. In this study, we pioneered an approach by combining this fluorescence-based fusion assay with the use of landing pad cells to establish a high-throughput fusion assay that is compatible with DMS[41].

Specifically, we constructed a DMS library of membrane-bound S that was expressed by HEK293T landing pad cells, such that each cell would encode and express one S mutant. The DMS library contained all possible amino acid mutations from residues 883 to 1034, which covers HR1 (residues 912-984) and CH (residues 985-1034). All S-expressing cells also expressed mNeonGreen2$_{11}$ (mNG2$_{11}$), which belongs to the split monomeric NeonGreen2 system[42]. At the same time, a stable cell line that expressed hACE2 and mNG2$_{1\text{-}10}$ was generated (**Supplementary Fig. 1**). For the rest of the study, unless otherwise stated, HEK293T landing pad cells that expressed S and mNG2$_{11}$ are abbreviated as "S-expressing cells" and those that expressed hACE2 and mNG2$_{1\text{-}10}$ are abbreviated as "hACE2-expressing cells".

When S-expressing cells were mixed with hACE2-expressing cells, S-expressing cells that encoded fusion-competent mutants would fuse with hACE2-expressing cells to form green-fluorescent syncytia (**Fig. 1a,c, see Methods**). In contrast, no fusion would occur with S-expressing cells that encoded fusion-incompetent mutants. Subsequently, fluorescence-activated cell sorting (FACS) was performed to separate the unfused cells and green-fluorescent syncytia, both of which were then analyzed by next-generation sequencing. The fusogenicity of each mutant could be quantified by comparing its frequency between the green-fluorescent syncytia sample and the unfused cell sample. In parallel, the expression level of each mutant was measured in a high-throughput manner as described previously[28,33] (**see Methods**).

Prior to performing the DMS experiments above, the expression of membrane-bound S in HEK293T landing pad cells was verified via flow cytometry analysis using the RBD antibody CC12.3[43] (**Fig. 1b**). Moreover, the formation of green-fluorescent syncytia due to the fusion of S-expressing cells and hACE2-expressing cells was also verified by microscopy and flow cytometry (**Fig. 1c,d, Supplementary Fig. 2a**). We further showed that such fusion can be inhibited by CC40.8, a neutralizing antibody to the stem helix of the S fusion machinery[44], at the highest concentration tested (**Supplementary Fig. 2b**). This result confirmed that the fusion of S-expressing cells and hACE2-expressing cells was mediated by the S protein. We optimized the fusion assay to maximize the formation of green-fluorescent syncytia while minimizing the risk of clogging the cell sorter (**Supplementary Fig. 2c-e**).

Identification of fusion-incompetent S mutations with high expression level

From the DMS results, we computed the fusion score and expression score for each of the 2736 missense mutations, 152 nonsense mutations, and 152 silent mutations (**see Methods**). A higher expression score indicates a higher S expression level. Similarly, a higher fusion score indicates higher fusogenicity. Both expression score and fusion score were normalized such that the average score of silent mutations was 1 and that of nonsense mutations was 0. Three and two biological replicates were performed for the high-throughput expression and fusion assays, respectively. The Pearson correlation coefficient of expression scores among replicates ranged from 0.72 to 0.79, whereas that of fusion scores between replicates was 0.61, confirming the reproducibility of our DMS experiments (**Extended Data Fig. 1a,b**). In addition, the expression score distribution and fusion score distribution of silent mutations were significantly different from those of nonsense mutations (**Extended Data Fig. 1c,d**), indicating that our DMS experiments could distinguish mutants with different expression and fusogenicity levels. The expression score and fusion score for individual mutations are shown in **Extended Data Fig. 2** and **Supplementary Table 1**.

Since our fusion assay measured the fusogenicity at the cell level rather than at the single molecule level, the fusion score would be influenced by the expression level even if the fusogenicity per S molecule remained constant. Consistently, the fusion score positively correlated with the expression score (**Fig. 2a**). To correct for the effect of S expression level on fusogenicity, we computed an adjusted fusion score, which represented the residual of a linear regression model of fusion score on expression score (**Fig. 2b**). Mutations that had a low adjusted fusion score and a high expression score included the well-known prefusion-stabilizing mutations K986P and V987P that were used in current COVID-19 vaccines[45,46] (**Fig. 2b**), substantiating that our method could identify prefusion-stabilizing mutations.

Previous studies have shown that the expression of S with K986P/V987P can be improved by additional mutations[24-27], as exemplified by an S construct known as HexaPro, which contains mutations F817P, A892P, A899P, A942P on top of K986P and V987P. Except for F817P, the other mutations in HexaPro were all present in our DMS library. Consistent with the original report of HexaPro[24], our DMS data showed that A899P had minimal influence on S expression, whereas A892P and A942P noticeably increased S expression (**Fig. 2a,b**). These observations further validated of our DMS data.

Validation and combinations of prefusion-stabilizing mutations

Besides K986P and V987P, we also identified other mutations in HR1 and CH that had a low adjusted fusion score and a high expression score, particularly T961F, D994E, D994Q and Q1005R (**Fig. 2b,c**). Of note, D994E and D994Q were at the same residue position and chemically similar. By expressing these four mutations individually using HEK293T landing pad cells, we validated that they indeed improved the surface expression of S (**Fig. 3a, Extended Data Fig. 3a**) and prevented the formation of syncytia when incubated with hACE2-expressing cells (**Fig. 3d, Extended Data Fig. 4a,b**). Consistent with the DMS data (**Fig. 2**), the effects of T961F, D994E, D994Q and Q1005R on S expression and fusogenicity were comparable to K986P and V987P in the validation experiments. As a control, we also selected two mutations that had a high adjusted fusion score and a high expression score, namely S943H and A944S (**Fig. 2b**), and validated their enhancement in S expression and fusogenicity (**Fig. 3b,e, Extended Data Fig. 3b, Extended Data Fig. 4c,d**).

Subsequently, we combined the validated fusion-incompetent mutations K986P, V987P, D994Q and Q1005R to generate double (K986P/V987P: ‘2P’), triple (K986P/V987P/D994Q: ‘2PQ’, K986P/V987P/Q1005R: ‘2PR’) and quadruple (K986P/ V987P/D994Q/Q1005R: ‘2PQR’) mutants of membrane-bound S. Surface expression of these mutation combinations was higher than that of WT, but comparable with each other (**Fig. 3c, Extended Data Fig. 3c**). As expected, none of these S mutation combinations fused with hACE2-expressing cells (**Fig. 3f, Extended Data Fig. 4e,f**). We further tested the expression of soluble S ectodomain with different mutants. Interestingly, addition of the D994Q to 2P improved expression of soluble S ectodomain by approximately three-fold while the Q1005R drastically reduced expression of soluble S (**Extended Data Fig. 5**). Q1005R seemed to increase the formation of higher order oligomers of soluble S ectodomain, as observed by a peak higher than the expected size of trimeric S ectodomain in size exclusion chromatography of all mutants that contained Q1005R (**Extended Data Fig. 5b**). These observations indicate that certain mutations can improve the expression level of S in membrane-bound form but not soluble ectodomain form.

Structural and biophysical characterization of 2PQ spike

Due to the improvement of 2PQ over 2P in soluble S ectodomain expression, we proceeded with biophysical characterization of 2PQ to rationalize the prefusion-stabilization mechanism of D994Q. The prefusion conformation of 2PQ was confirmed by low-resolution cryogenic electron microscopy (**Fig. 4a,b, Extended Data Fig. 6a**). We also assessed the thermal stability of 2PQ relative to 2P. Differential scanning fluorimetry revealed that both 2P and 2PQ had an apparent melting temperature at approximately 46.5 °C, similar to the previously reported value for 2P[24].

Nevertheless, 2PQ had another peak at approximately 62 °C, suggesting that the additional D994Q mutation prevents immediate, complete unfolding of S (**Fig. 4c**). This observation is corroborated by *in silico* mutagenesis using Rosetta, which showed that helices are brought together in proximity so that D994Q forms an intraprotomer hydrogen bond with Q758 to stabilize the prefusion conformation (**Fig. 4d, Extended Data Fig. 6b**).

Finally, we tested whether D994Q altered the antigenicity of the S protein. We compared the binding of 2P and 2PQ to various S antibodies, including CC12.3 (RBD)[43], S2M28 (NTD)[47], CC40.8 (S2 stem helix)[44], and COVA1-07 (S2 HR1)[48], using biolayer interferometry (BLI) or flow cytometry. 2P and 2PQ showed similar binding affinity to CC12.3, CC40.8 and S2M28 (**Fig. 4e, Extended Data Fig. 7a,b, Extended Data Fig. 8**). However, when assayed for binding with COVA1-07, a non-neutralizing antibody to the HR1[48], 2PQ had an approximately 10-fold weaker dissociation constant, a slower on-rate and a faster off-rate than 2P (**Fig. 4e, Extended Data Fig. 7c**). Since COVA1-07 only binds efficiently when S is in an open conformation that has transitioned away from the prefusion conformation[48], our result substantiates that D994Q can further enhance the prefusion-stability of 2P, which is known to insufficiently stabilize the prefusion conformation[24,25,49]. Collectively, these data reveal a prefusion-stabilization mechanism of D994Q and demonstrate its minimal impact on S antigenicity.

DISCUSSION

Structure-based design[50] of prefusion-stabilized class I viral fusion proteins has been successfully applied to HIV[51-54], RSV[55], Nipah[56], Lassa[57], Ebola[58], and more recently SARS-CoV-2[24-27]. Although structure-based design is an effective approach for prefusion-stabilization, it requires structural determination and subsequent expression, purification, and characterization of each candidate mutation individually. This laborious experimental process limits the comprehensiveness of using a structure-based approach to identify prefusion-stabilizing mutations. In this study, we established a high-throughput approach to measure the fusogenicity of thousands of mutations in parallel. This approach enables systematic identification of prefusion-stabilizing mutations without relying on structural information. While we only provide a proof-of-concept using the SARS-CoV-2 S protein, our approach can be adopted to fusion proteins of other viruses. Given that prefusion-stabilization is critical for viral immunogen design[50,59], our work here should advance the process of viral vaccine development.

One interesting finding in this study is that improving the expression of membrane-bound (i.e. full-length) S protein does not necessarily improve the expression of soluble S ectodomain, as exemplified by Q1005R. This observation indicates that the ectodomain of the S protein has some long-range interactions with its native transmembrane domain, which is excluded from the S ectodomain construct. As a result, caution is needed when extrapolating the results obtained from full-length S protein to soluble S ectodomain, or vice versa. However, since most COVID-19 vaccines on the market are based on the full-length membrane-bound S protein[60], the results from our high-throughput fusion and expression assays, which are also based on full-length membrane-bound S protein, are directly applicable to COVID-19 vaccine development.

Although most SARS-CoV-2 neutralizing antibodies target RBD[61], recent studies have shown that antibodies to S2 can also neutralize, albeit often at a lower potency[44,62-65]. As a result, understanding the evolutionary constraints of S2 is relevant to SARS-CoV-2 antigenic drift and to design of more universal coronavirus vaccines. While many mutations in HR1 and CH, including those of major SARS-CoV-2 variants (**Extended Data Table 1**), do not negatively impact the expression or fusogenicity of the S protein (**Fig. 2b**), HR1 and CH show high degrees of evolutionary conservation among betacoronaviruses (**Extended Data Fig. 9**). This observation could be due to low levels of positive selection pressure on HR1 and CH, since most neutralizing antibodies are directed towards the RBD[61]. Alternatively, besides S protein expression and fusogenicity, other evolutionary constraints on HR1 and CH may be present *in vivo*. Future studies of the relationship among S protein expression, fusogenicity, and virus replication fitness will provide important biophysical insights into the evolution of SARS-CoV-2.

If the prefusion-stabilizing mutations of betacoronavirus S protein were not reported in late 2010s[17,66], it would not have been possible to develop an effective COVID-19 vaccine at the speed that occurred, even with the availability and utilization of the mRNA vaccine technology. It is unclear whether the next pandemic will be caused by a virus that we have prior knowledge about. Consequently, while the speed of vaccine manufacturing has been revolutionized by the mRNA vaccine technology[67], it is equally important to maximize the speed of immunogen design so that we are fully prepared for the next pandemic. We believe our work here provides an important step in that regard.

METHODS

Cell culture

Human embryonic kidney 293T (HEK293T) landing pad cells were grown and maintained in complete growth medium: Dulbecco's modified Eagle medium (DMEM) with high glucose (Gibco), supplemented with 10% v/v fetal bovine serum (FBS; VWR), 1× non-essential amino acids (Gibco), 100 U/mL penicillin and 100 μg/mL streptomycin (Gibco), 1× GlutaMAX (Gibco) and 2 μg/mL doxycycline (Thermo Scientific) at 37 °C, 5% CO_2 and 95% humidity. Expi293F cells (Gibco) were grown and maintained in Expi293 expression medium (Gibco) at 37 °C, 8% CO_2, 95% humidity and 125 rpm according to the manufacturer's instructions.

Landing pad plasmids

attB plasmids each encoding (hACE2, an internal ribosomal entry site [IRES], and hygromycin resistance: attB-hACE2), (hACE2, an IRES, general control nondepressible 4 [GCN4] leucine zipper fused to $mNG2_{1\text{-}10}$, a (GSG) P2A self-cleaving peptide, and hygromycin resistance: attB-hACE2-mNG2-1-10), and (S with the PRRA motif in the furin cleavage site deleted, an IRES, GCN4 leucine zipper fused to $mNG2_{11}$, a (GSG) P2A self-cleaving peptide, and puromycin resistance: attB-S-mNG2-11) were constructed and assembled via polymerase chain reaction (PCR). The sequence of S used in this study was the ancestral (Wuhan-Hu-1) strain (GenBank accession ID: MN908947.3)[68]. The PRRA motif in the furin cleavage site was deleted to prevent spontaneous fusion of S-expressing cells with each other[69]. For experimental validation, mutants of S were individually constructed using PCR-based site-directed mutagenesis. Pairs of primers used for PCR-based site directed mutagenesis are listed in **Supplementary Table 2**.

Deep mutational scanning library construction

Cassette primers for DMS library construction are listed in **Supplementary Table 3**. Cassette primers were resuspended in MilliQ H_2O such that the final concentration of all primers is 10 μM. Forward cassette primers, named as CassetteX_N (X = 1, 2, …, 19; N = 1, 2, …, 8), that belong to the same cassette (i.e., the same value of X) were mixed in equimolar ratios. Each forward cassette primer also carried unique silent mutations (i.e. synonymous mutations) to help distinguish between sequencing errors and true mutations in downstream sequencing data analysis as described previously[70]. For the first round of PCR, two sets of reactions were set up. The first set had the mixed cassette primers and 5′-ACG ACG TCT CCT TCT CTA GGA AAG TGG GCT TTG C-3′ as forward and reverse primers, respectively. The second set had 5′-TGC TCG TCT CCA AAG TGA CAC TGG CCG ACG CCG G-3′ and CassetteX_Rprimers (X = 1, 2, …, 19) as forward and reverse primers, respectively. Since we had 19 cassettes, there were 19 PCRs for each of the two sets of reactions. For both sets, the template used was attB-S-mNG2-11. Thereafter, products corresponding to the correct size were excised and purified using Monarch DNA Gel Extraction kit (NEB). For the second round of PCR, 10 ng of PCR product from each of the first and second sets in the same cassette were mixed. 5′-ACG ACG TCT CCT TCT CTA GGA AAG TGG GCT TTG C-3′ and 5′-TGC TCG TCT CCA AAG TGA CAC TGG CCG ACG CCG G-3′ were used as the forward and reverse primers, respectively. PCR products corresponding to the correct size were excised and purified using DNA Gel Extraction kit (NEB). 100 ng of each gel-purified PCR products (total of 19) were mixed and digested with BsmBI restriction enzyme (NEB) for 2 h at 55 °C. Then, the product was purified using PureLink PCR Purification kit (Invitrogen) and served as the insert.

To amplify the vector, attB-S-mNG2-11, 5′-CAC TCG TCT CGA GAA GGC GTG TTC GTG TCC AAC G-3′, and 5′-GGC CCG TCT CAC TTT GTT GAA CAG CAG GTC CTC G-3′ were used as template, forward primer, and reverse primer, respectively. The PCR product was digested with DpnI (NEB) for 2 h at 37 °C, purified with PureLink PCR Purification kit (Invitrogen), digested with

BsmBI restriction enzyme (NEB) for 2 h at 55 °C, and purified again using a PureLink PCR Purification kit (Invitrogen). All PCRs were performed using PrimeSTAR Max DNA Polymerase (Takara) according to the manufacturer’s instructions.

BsmBI-digested vector and insert were ligated in a molar ratio of 1:100 to a total of 1 μg using T4 DNA ligase (NEB) for 2 h at room temperature. A control ligation reaction was set up by only having the BsmBI-digested vector (no insert). 1 μL ligation reaction products were transformed into chemically competent DH5α *Escherichia coli* cells and plated onto agar plates with 100 μg/mL ampicillin. The ligation mixture that contained vector and insert had at least 10 times more colonies than the control reaction. Subsequently, the ligation mixture was column-purified using a PureLink PCR Purification kit and eluted in 10 μL of MilliQ H_2O. 1 μL of the purified ligated product was mixed with 30 μL MegaX DH10β T1^R electrocompetent *E. coli* cells (NEB) into an electroporation cuvette with a 1 mm gap (BTX). Electroporation was performed at 2.0 kV, 200 Ω and 25 μF using an ECM 830 square wave electroporation system (BTX). 1 mL of SOC recovery medium (NEB) was added immediately into cells after electroporation. Two electroporation reactions were performed. Cells were recovered for 1 h at 37 °C with shaking at 250 rpm. After recovery, cells were collected via centrifugation, resuspended in 400 μL lysogeny broth (LB), plated onto 150 mm agar plates supplemented with 100 μg/mL ampicillin, and incubated overnight at 37 °C. At least 1×10^6 colonies were scrape-harvested with LB broth and plasmids were extracted using a PureLink Plasmid Midiprep kit (Invitrogen).

Landing pad cell transfection

6.0×10^5 HEK293T landing pad cells in 1.35 mL of complete growth medium were seeded per well of a 6-well plate. 1.7 μg of attB-hACE2-mNG2-1-10 plasmid or attB-S-mNG2-11 plasmid were added into 5 μL FuGENE 6 transfection reagent (Promega) and OptiMEM (Gibco) to a total volume of 240 µL. The transfection mixture was subsequently added dropwise into cells. Transfection was carried out on the same day as seeding. One day post-transfection, 500 µL of complete growth medium was added to cells. Three days post-transfection, medium was discarded, cells were washed with 1× PBS, and incubated in negative selection medium (complete growth medium supplemented with 10 nM AP1903) for one day at 37 °C, 5% CO_2 and 95% humidity. Then, the medium was discarded, cells were washed with 1× PBS, and recovered in complete growth medium for two days at 37 °C, 5% CO_2 and 95% humidity. Cells were then trypsinized and grown in positive selection medium indefinitely: hACE2- and S-expressing cells were maintained in hygromycin medium (complete growth medium supplemented with 100 µg/mL hygromycin B [Invivogen]) and puromycin medium (complete growth medium supplemented with 1 µg/mL puromycin [Invivogen]), respectively.

To construct the S2 HR1/CH DMS cell line, the above protocol was used with modifications: 3.5 $\times 10^6$ cells in 8 mL of complete growth medium in a T75 flask were transfected with 7.1 µg of the DMS plasmid library and 29 µL of FuGENE6 transfection reagent in 1.4 mL of OptiMEM. For positive selection and regular maintenance, puromycin medium was used.

Flow cytometry

To validate hACE2 surface expression after transfection, landing pad cells were harvested via centrifugation at 300 × *g* for 5 min at 4 °C, resuspended in ice-cold FACS buffer (2% v/v FBS, 50 mM EDTA in DMEM supplemented with high glucose, L-glutamine and HEPES, without phenol red [Gibco]), and incubated with 2 µg/mL of SARS-CoV-2 S RBD-IgG Fc for 1 h at 4 °C. Then, cells were washed once, and resuspended with ice-cold FACS buffer. Cells were incubated with 1 µg/mL of phycoerythrin (PE)-conjugated anti-human IgG Fc (BioLegend). Cells were washed once and resuspended in ice-cold FACS buffer. Cells were analyzed using an Accuri C6 flow cytometer (BD Biosciences).

The above protocol for verification and quantification of S surface expression was used except cells were incubated with 5 µg/mL of CC12.3[43], an RBD antibody, instead of SARS-CoV-2 S RBD-IgG Fc, for 1 h at 4 °C. To quantify fold change in surface expression of S relative to WT based on median fluorescence intensity (MFI), equation (1) was used in the plot of FSC-A against PE:

$$\mathrm{MFI_{FC}} = \frac{\mathrm{MFI_{mutant}} - \mathrm{MFI_{control}}}{\mathrm{MFI_{WT}} - \mathrm{MFI_{control}}} \tag{1}$$

MFI values were obtained after plotting data in FCS Express Flow Cytometry software (De Novo Software). Gating strategy is shown in **Supplementary Fig. 3a**. Raw MFI and $\mathrm{MFI_{FC}}$ values are provided as source data.

To assess fusogenicity of S (WT or mutants), an equal number of hACE2, $\mathrm{mNG2_{1\text{-}10}}$- and S, $\mathrm{mNG2_{11}}$-expressing cells were mixed such that the total cell number is 5.0×10^5 cells per mL of complete growth medium. Cells were co-cultured for 3 h at 37 °C, 5% CO_2 and 95% humidity. Cells were then harvested and resuspended in ice-cold FACS buffer. Cells were analyzed using an Accuri C6 flow cytometer (BD Biosciences). Gating strategy is shown in **Supplementary Fig. 3b**. The percentage of mNG2-positive events of mutants relative to that of WT S was calculated. Percentage of mNG2-positive events and fold change relative to WT are provided as source data.

Expression sorting

Cells expressing the S2 HR1/CH DMS library of S were harvested via centrifugation at 300 × *g* for 5 min at 4 °C. Supernatant was discarded, and cells were resuspended in ice-cold FACS buffer. Cells were incubated with 5 µg/mL of CC12.3 for 1 h at 4 °C. Then, cells were washed once, and resuspended with ice-cold FACS buffer. Cells were incubated with 2 µg/mL of PE anti-human IgG Fc. Cells were washed once, resuspended in ice-cold FACS buffer, and filtered through a 40 μm strainer. Cells were sorted via a four-way sort using a BigFoot spectral cell sorter (Invitrogen) according to PE fluorescence at 4 °C. Cells expressing the highest PE fluorescence were sorted into "bin 3", then the next highest into "bin 2", followed by "bin 1" and then "bin 0". Each bin had ~25% of the singlet population. Gating strategy is shown in **Supplementary Fig. 3c**. Number of cells collected per bin per replicate is shown in **Supplementary Table 4**. Of note, since CC12.3 binds to the RBD[43], an independently folded region of S that is present only in the prefusion but not postfusion conformation[1,71], our sort was based on the expression of prefusion S.

Fusion sorting

Cells expressing the HR1/CH DMS library of S, and cells expressing hACE2 were resuspended in complete growth medium and filtered through a 40 μm cell strainer to obtain single cell suspensions. 2.5×10^6 cells of each were mixed in a T-75 flask and complete growth medium added to a total volume of 10 mL. Six co-cultures were set up, with one of the co-cultures acting as a negative, non-fluorescent control by mixing hACE2- and S-expressing cells that do not have split mNG2. Co-cultures were incubated for 3 h at 37 °C, 5% CO_2 and 95% humidity. Subsequently, cells were harvested and pelleted via centrifugation at 300 × *g* for 5 min at 4 °C. Supernatant was discarded, and cells were resuspended in ice-cold FACS buffer. Cells were sorted via a two-way sort using a BigFoot spectral cell sorter (Invitrogen) according to presence or absence of mNG2 fluorescence at 4 °C. Gating strategy is shown in **Supplementary Fig. 3d**. Number of cells collected per bin per replicate is shown in **Supplementary Table 5**.

Post-sorting genomic DNA extraction

After FACS, cell pellets were obtained via centrifugation at 300 × *g* for 15 min at 4 °C, and the supernatant was discarded. Genomic DNA was extracted using a DNeasy Blood and Tissue Kit (Qiagen) following the manufacturer's instructions with a modification: resuspended cells were incubated and lysed at 56 °C for 30 min instead of 10 min.

Deep sequencing

After genomic DNA extraction, the region of interest was amplified via PCR using 5′-CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT CTA CAT CTG CCC TGC TGG CCG GCA CA-3′ and 5′-GAC TGG AGT TCA GAC GTG TGC TCT TCC GAT CTG CAA AAG TCC ACT CTC TTG CTC TG-3′ as forward and reverse primers, respectively. A maximum of 500 ng of genomic DNA per 50 µL PCR reaction was used as template; 4 µg of genomic DNA per expression or fusion bin, per replicate, was used as template. PCR was performed using KOD DNA polymerase (Takara) with the following settings: 95 °C for 2 min, 25 cycles of (95 °C for 20 s, 56 °C for 15 s, 68 °C for 20 s), 68 °C for 2 min, 12 °C indefinitely. All eight 50 µL reactions per bin per replicate were mixed after PCR. 100 µL of product per bin per replicate was used for purification using a PureLink PCR Purification kit. Subsequently, 10 ng of the purified PCR product per bin per replicate was appended with Illumina deep sequencing barcodes via PCR using KOD DNA polymerase with the following settings: 95 °C for 2 min, 9 cycles of (95 °C for 25 s, 56 °C for 15 s, 68 °C for 20 s), 68 °C for 2 min, 12 °C indefinitely. Barcoded products were mixed and sequenced with a MiSeq PE300 v3 flow cell (Illumina).

Analysis of deep sequencing data

Forward and reverse reads were merged via PEAR[72]. Using custom Python code, forward reads were translated and matched to the corresponding mutant. Counts for expression and fusion bins for each replicate were tabulated. For each replicate, the frequency of each mutant was calculated as the count of that mutant divided by the total number of counts in that bin, as shown in equation (2):

$$F_{\text{mut, binX}} = \frac{C_{\text{mut,binX}}}{\Sigma C_{\text{binX}}} \text{ for X = 0, 1, 2, 3, mNG2}^{-}\text{, mNG2}^{+} \quad (2)$$

For each replicate, the weighted expression score for each mutant (W_{mut}) was calculated using equation (3):

$$W_{\text{mut}} = \frac{(F_{\text{mut, bin0}} \times 0.25) + (F_{\text{mut, bin1}} \times 0.5) + (F_{\text{mut, bin2}} \times 0.75) + (F_{\text{mut, bin3}} \times 1)}{F_{\text{mut, bin0}} + F_{\text{mut. bin1}} + F_{\text{mut, bin2}} + F_{\text{mut, bin3}}} \quad (3)$$

The weighted expression scores were normalized ($W_{\text{mut}}^{\text{norm}}$) such that the average W_{mut} of nonsense mutations equals 0, and the average W_{mut} of silent mutations equals 1 using equation (4):

$$W_{\text{mut}}^{\text{norm}} = \frac{W_{\text{mut}} - W_{\text{nonsense}}^{\text{avg}}}{W_{\text{silent}}^{\text{avg}} - W_{\text{nonsense}}^{\text{avg}}} \quad (4)$$

The final expression score ($W_{\text{mut}}^{\text{avg}}$) for each mutant was calculated using equation (5):

$$W_{\text{mut}}^{\text{avg}} = \frac{1}{3} \times \left(W_{\text{mut}}^{\text{norm,rep1}} + W_{\text{mut}}^{\text{norm,rep2}} + W_{\text{mut}}^{\text{norm,rep3}}\right) \quad (5)$$

Fusion scores (U_{mut}) were calculated for each replicate by the formula shown in equation (6):

$$U_{\text{mut}} = \log_{10}\left(\frac{F_{\text{mut,mNG2}^{+}}}{F_{\text{mut,mNG2}^{-}}}\right) \quad (6)$$

Fusion scores were normalized ($U_{\text{mut}}^{\text{norm}}$) such that the $U_{\text{mut}}^{\text{avg}}$ of silent mutations equals 1, and the $U_{\text{mut}}^{\text{avg}}$ of nonsense mutations equals 0 using equation (7):

$$U_{\text{mut}}^{\text{norm}} = \frac{U_{\text{mut}} - U_{\text{nonsense}}^{\text{avg}}}{U_{\text{WT}}^{\text{avg}} - U_{\text{nonsense}}^{\text{avg}}} \quad (7)$$

Then, the final average score ($U_{\text{mut}}^{\text{avg}}$) for each mutant was calculated using equation (8):

$$U_{\text{mut}}^{\text{avg}} = \frac{1}{2} \times \left(U_{\text{mut}}^{\text{norm,rep1}} + U_{\text{mut}}^{\text{norm,rep2}}\right) \quad (8)$$

Adjusted fusion score of each mutant is equal to the residual of that mutant in a linear regression model of fusion score against expression score. The linear regression model and residuals were calculated using the ‘lm’ and ‘resid’ functions in RStudio, respectively.

Sequence conservation analysis

Sequences were obtained from GenBank or GISAID (**Extended Data Table 1, Supplementary Table 7**). A BLAST database was created, and the reference sequence of the DMS region (residues 883-1034) was used to run tblastn to generate BlastXML files. The reference sequence used was the founder strain of SARS-CoV-2 (GenBank accession number: MN908947.3)[68]. Extracted information was obtained by running ‘XML_Extraction.py’[73]. Multiple alignment using MAFFT was then performed[74]. Sequence conservation was calculated based on the residue conservation at each position relative to the reference sequence. Mean expression score and mean fusion score were calculated by taking the average of the expression scores and fusion scores of all mutants, respectively, at that position.

Fluorescence microscopy

Images were captured with an ECHO Revolve epifluorescence microscope (ECHO) with a UPLANFL N 10×/0.30 NA objective (Olympus) using the FITC channel for mNG2 fluorescence. Brightfield images were also obtained using white light. Fluorescent and brightfield images were then overlaid. Identical exposure and intensity settings were used to capture images. Scale bars correspond to 100 μm for all micrographs.

Cryogenic electron microscopy

To prepare cryoEM grid, an aliquot of 3.5 μL purified protein at ~1 mg/mL concentration was applied to a 300-mesh Quantifoil R1.2/1.3 Cu grid pre-treated with glow-discharge, blotted in a Vitrobot Mark IV machine (force -5, time 3 s), and plunge-frozen in liquid ethane. The grid was loaded in a Titan Krios microscope equipped with Gatan BioQuantum K3 imaging filter and camera. A 10-eV slit was used for the filter. Data collection was done with serialEM[75]. Images were recorded at 130,000× magnification, corresponding to a pixel size of 0.33 Å/pix at super-resolution mode of the camera. A defocus range of -0.8 μm to -1.5 μm was set. A total dose of 50 $e^-/Å^2$ of each exposure was fractionated into 50 frames. The first two frames of the movie stacks were not included in motion-correction. CryoEM data processing was performed on the fly with cryoSPARC Live[76] following regular single-particle procedures.

Rosetta-based mutagenesis

The structure of S was obtained from the Protein Data Bank (PDB ID: 6ZGE). N-acetyl-D-glucosamine and water molecules were removed using PyMOL (Schrödinger), and amino acids were renumbered using pdb-tools[77]. The ‘fixbb’ application in Rosetta (RosettaCommons) was used to generate the D994Q mutation in all protomers. One-hundred poses were obtained, and the lowest scoring pose was used for further processing. A constraint file was generated using the lowest-scoring pose from fixed backbone mutagenesis as input, and the ‘minimize_with_cst’ application in Rosetta. Fast relax was subsequently performed using the ‘relax’ application[78] with the constraint file. The lowest scoring pose out of thirty was used for structural analysis.

Antibody expression and purification

Codon-optimized oligonucleotides encoding the heavy chain and light chain of the indicated antibodies were cloned into phCMV3 plasmids in an IgG1 Fc format with a mouse immunoglobulin kappa signal peptide. Plasmids encoding the heavy chain and light chain of antibodies were transfected into Expi293F cells using an Expifectamine 293 transfection kit (Gibco) in a 2:1 mass ratio following the manufacturer’s protocol. Supernatant was harvested 6 days post-transfection and centrifuged at 4000 × *g* for 30 min at 4 °C to remove cells and debris. The supernatant was subsequently clarified using a polyethersulfone membrane filter with a 0.22 µm pore size (Millipore).

CaptureSelect CH1-XL beads (Thermo Scientific) were washed with MilliQ H_2O thrice and resuspended in 1× PBS. The clarified supernatant was incubated with washed beads overnight at 4 °C with gentle rocking. Then, flowthrough was collected, and beads washed once with 1× PBS. Beads were incubated in 60 mM sodium acetate, pH 3.7 for 10 min at 4 °C. The eluate containing antibody was buffer-exchanged into 1× PBS using a centrifugal filter unit with a 30 kDa molecular weight cut-off (Millipore) four times. Antibodies were stored at 4 °C.

Soluble S protein expression and purification

SARS-CoV-2 S ectodomain (residues 1-1213, which includes the native signal peptide) with the PRRA motif in the furin cleavage site deleted, C-terminal SGGGG linker, biotinylation site, thrombin cleavage site, Foldon trimerization sequence, and 6×His-tag were all cloned in-frame into a phCMV3 vector via PCR. Site-directed mutagenesis via PCR was performed to generate the indicated mutants of soluble S protein.

Expi293F cells were transfected with vectors encoding the indicated soluble spike protein mutant using an Expifectamine 293 transfection kit following the manufacturer's protocol. Cells were harvested six days post-transfection. The supernatant was collected via centrifugation at 4000 × *g* for 30 min at 4 °C, and further clarified using a polyethersulfone membrane with a 0.22 µm pore size (Millipore). The clarified supernatant was incubated with washed Ni sepharose excel His-tagged protein purification resin (Cytiva) with gentle rocking overnight at 4 °C. Flow-through was collected. Beads were washed once with 20 mM imidazole in 1× PBS, then washed once with 40 mM imidazole in 1× PBS, and finally eluted with 300 mM imidazole in 1× PBS thrice. Wash and elution fractions were subjected to denaturing sodium dodecyl sulfate-polyacrylamide gel electrophoresis (**Extended Data Fig. 5a**). All elution fractions were combined and concentrated using a centrifugal filter unit with a 30 kDa molecular weight cut-off (Millipore) via centrifugation at 4000 × *g* and 4 °C for 15 min. The concentrated protein mixture was passed through a Superdex 200 XK 16/100 column in 20 mM Tris-HCl pH 8.0 and 150 mM NaCl for size-exclusion chromatography (**Extended Data Fig. 5b-c**). Fractions corresponding to ~540 kDa were pooled and concentrated using a centrifugal filter unit with a 30 kDa molecular weight cut-off (Millipore) via centrifugation at 4000 × *g* and 4 °C for 15 min.

Differential scanning fluorimetry

200 ng/µL of purified S protein and 5× SYPRO orange (Thermo Fisher Scientific) were added into 20 mM Tris-HCl pH 8.0, 150 mM NaCl in optically clear tubes. SYPRO orange fluorescence intensity in relative fluorescence units (RFU) was measured over temperatures ranging from 10 °C to 95 °C using a CFX Connect Real-Time PCR Detection System (Bio-Rad). Melting temperature (T_m) was calculated as the temperature at which the first derivative of fluorescence intensity with respect to temperature, $-\frac{d(RFU)}{dT}$, was minimum.

Biolayer interferometry

Anti-His-tag (HIS1K) biosensors (Sartorius) were incubated in BLI buffer (0.002% v/v Tween-20 in 1× PBS) at room temperature for 30 min. Then, BLI was performed on an Octet 96e system (Sartorius) at room temperature with shaking at 1000 rpm as follows: baseline in BLI buffer for 60 s, loading with 500 nM 6×His-tagged soluble 2P or 2PQ for 15 min, baseline in BLI buffer for 60 s, association with the indicated antibody and concentration for 2 min, dissociation in BLI buffer for 5 min. Data were collected with Octet Data Acquisition software (Sartorius) and analyzed with

Octet Data Analysis software (Sartorius). A 1:1 model was used for curve fitting and estimating the dissociation constants, on-rates and off-rates.

DATA AVAILABILITY

Structures from the following identifiers from the Protein Data Bank (PDB) were used in this study: 6VXX and 6VYB. The cryoEM map of 2PQ spike will be deposited to Electron Microscopy Data Bank (EMDB) prior to publication. Raw sequencing data have been submitted to the NIH Short Read Archive under accession number: PRJNA826665. Biological materials including plasmids, and the S2 HR1/CH DMS and hACE2 stable cell lines can be obtained by contacting N.C.W. Source data are available for this paper.

CODE AVAILABILITY

Custom codes to analyze deep mutational scanning, thermal stability, flow cytometry and BLI data have been deposited to https://github.com/nicwulab/SARS2_S_fusogenicity_DMS.

ACKNOWLEDGEMENTS

We thank the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign for assistance with fluorescence-activated cell sorting and deep sequencing. We thank the cryogenic-electron microscopy core facility at the Case Western Reserve University School of Medicine. This work was supported by National Institutes of Health (NIH) R01 AI167910 (N.C.W.), DP2 AT011966 (N.C.W.), R35 GM142886 (K.A.M.), the Michelson Prizes for Human Immunology and Vaccine Research (N.C.W.), the Searle Scholars Program (N.C.W.), and the Bill and Melinda Gates Foundation INV-004923 (I.A.W.).

AUTHOR CONTRIBUTIONS

T.J.C.T and N.C.W. conceived and designed the study. TJ.C.T. established the fusion assay and performed the deep mutational scanning experiments. T.J.C.T and N.C.W. analyzed the deep mutational scanning data. T.J.C.T., R.L. and W.O.O. expressed and purified recombinant proteins. Z.M. and X.D. performed cryo-EM analysis. K.A.M. provided the landing pad cells and helped establish the fusion assay. M.Y. and I.A.W. provided the CC12.3 antibody; G.S. and R.A. provided the CC40.8 antibody. T.J.C.T. and C.K. performed the microscopy analysis. T.J.C.T. and N.C.W. wrote the paper and all authors reviewed and/or edited the paper.

COMPETING INTERESTS

N.C.W., K.A.M. and T.J.C.T. have filed a provisional patent application with the University of Illinois covering the method described in this article. N.C.W. serves as a consultant for HeliXon. The authors declare no other competing interests.

FIGURES

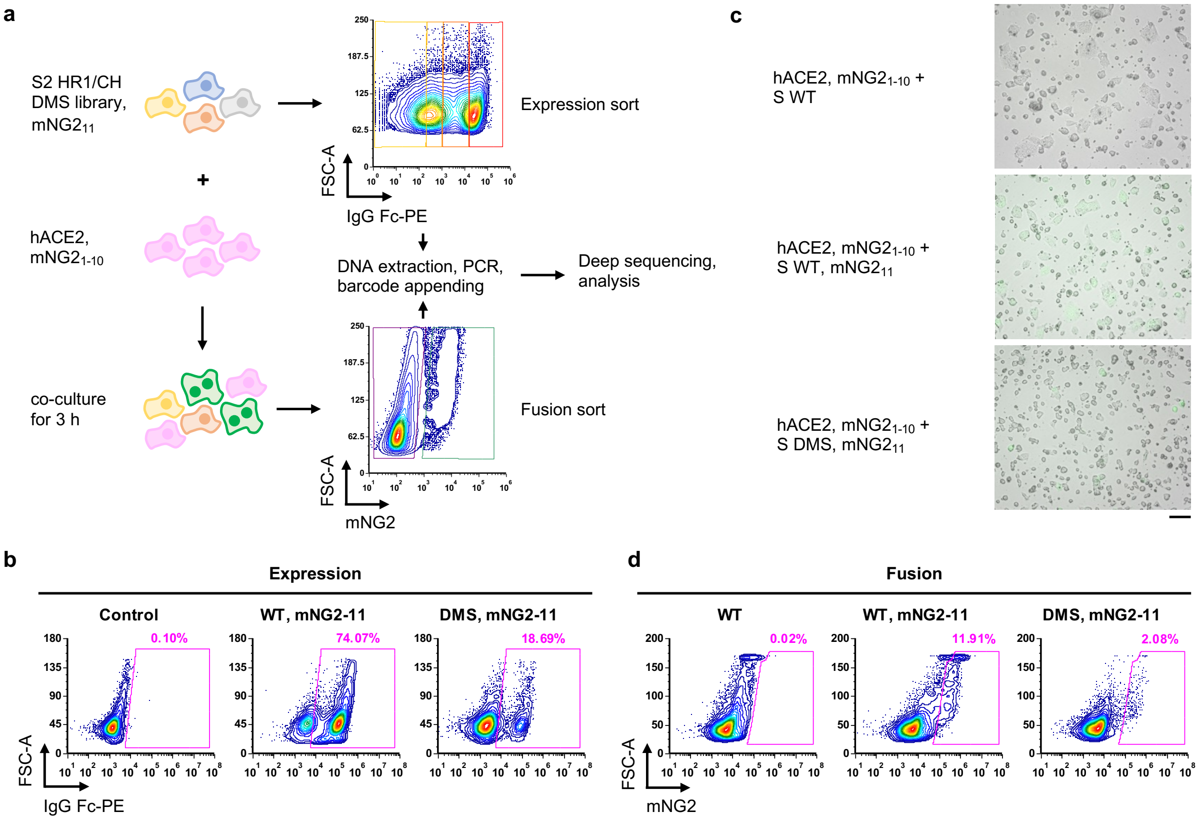


Fig. 1 | Measuring protein expression and fusogenicity of SARS-CoV-2 S mutations using deep mutational scanning. a, Schematic of high-throughput expression and fusion assays for S mutants. **b,** Flow cytometry analysis of S protein expression in HEK293T landing pad cells that encoded WT S or the DMS library. **c,** Fluorescent micrographs of co-culturing S-expression cells with hACE2-expressing cells. Scale bar: 100 μm. **d,** Flow cytometry analysis of fusion activity of co-culturing hACE2-expressing cells with HEK293T landing pad cells that encoded WT S or the DMS library. Components of split mNG2 are indicated where present.

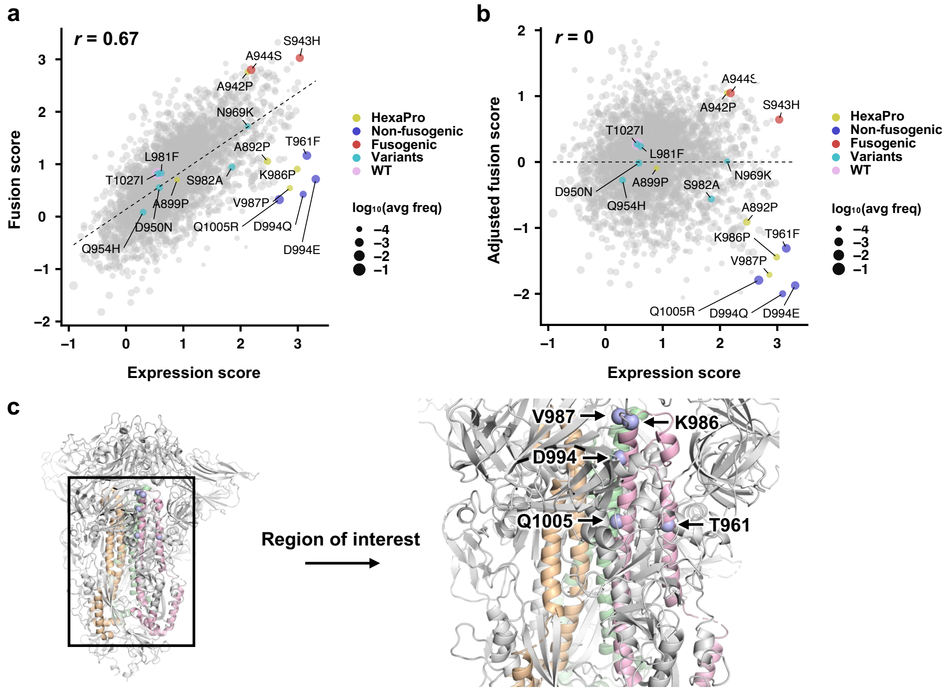


Fig. 2 | Expression and fusion scores of individual mutations in the DMS library. a, Plot of fusion score against expression score for each mutant is shown. WT is indicated in pink. Mutations used in HexaPro[24] are in yellow. Representative fusion-incompetent mutations identified in this study are in purple (non-fusogenic). Representative mutations that enhance S fusogenicity are in red (fusogenic). Mutations found in major SARS-CoV-2 variants (**Extended Data Table 1**) are in teal (variants). Each data point represents one mutation in the DMS library. Individual data points are sized according to average frequency of the corresponding mutations. **b,** Plot of adjusted fusion score against expression score for each mutant is shown. Pearson correlation coefficient, *r*, is shown in **a,b**. **c,** Locations of fusion-incompetent mutations are indicated by light blue spheres. Regions that are mutated in the DMS library are colored wheat, green and pink for each monomer. Other regions on the S are colored in grey. Source data are available as **Supplementary Table 1**.

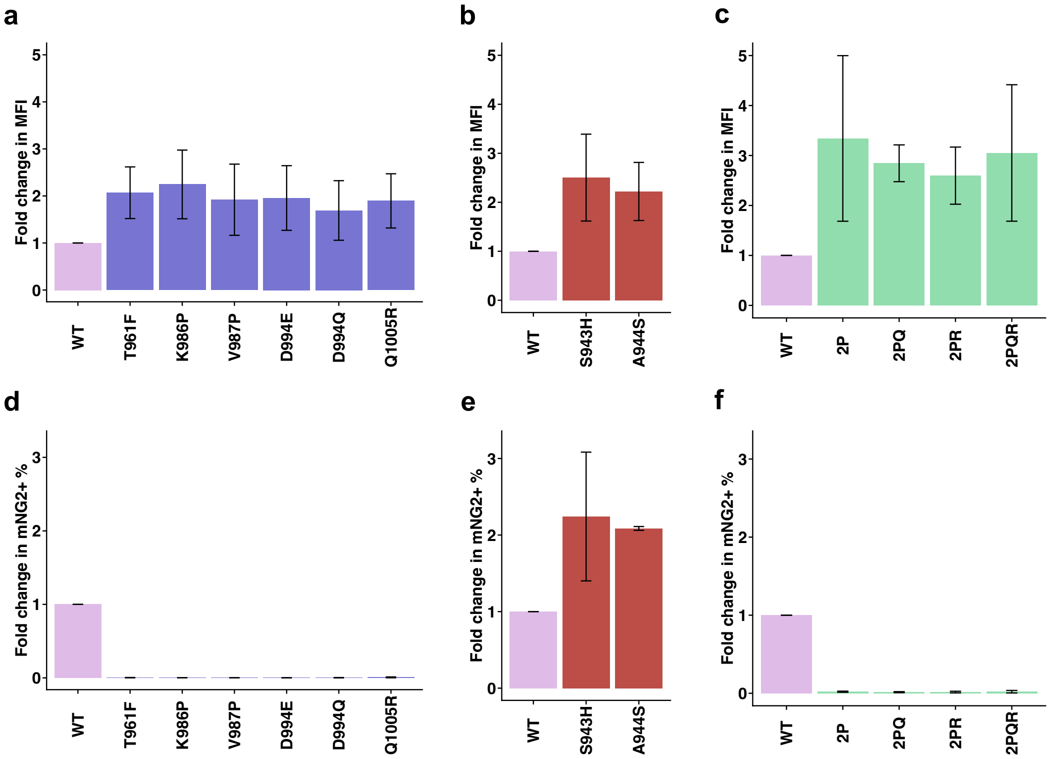
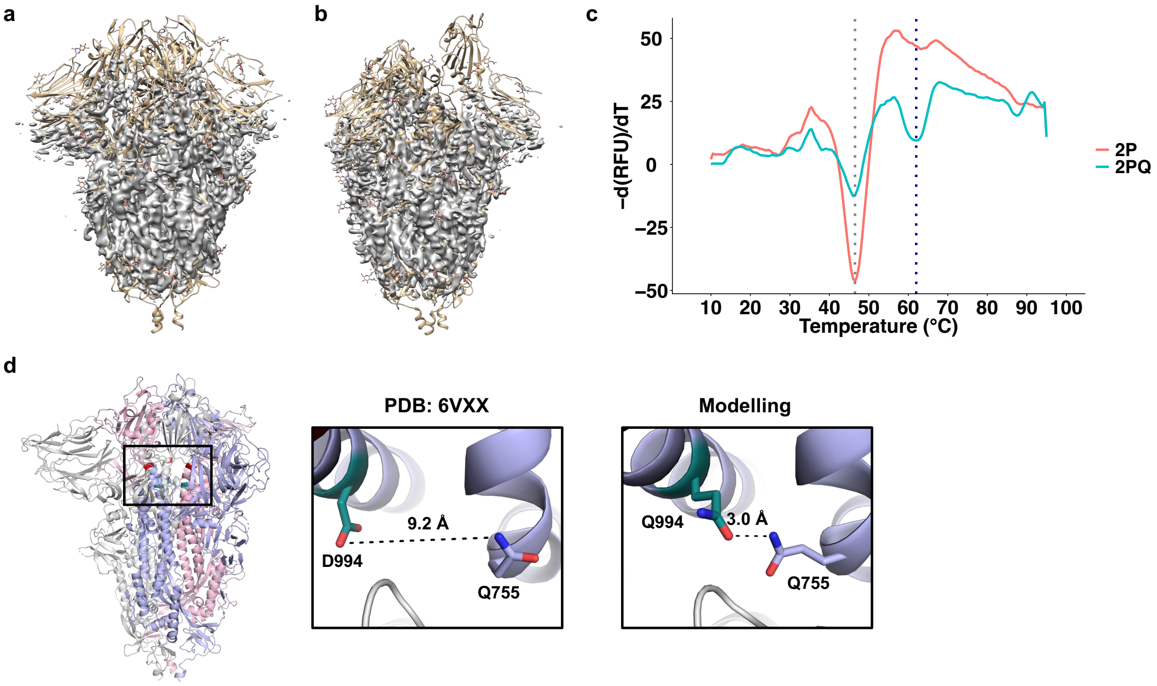


Fig. 3 | Validation of candidate prefusion-stabilizing mutations. a-c, Expression of prefusion-stabilizing mutations **(a)**, fusion-enhancing mutations **(b)**, and combinations of candidate prefusion-stabilizing mutations of S **(c)** relative to WT. Of note, the numerical values of fold change in median fluorescence intensity (MFI) indicate relative and not absolute fold changes in surface expression levels of S. **d-f,** Fold change in fusion activity of candidate prefusion-stabilizing mutations **(d)**, fusion-enhancing mutations **(e)**, and combinations of candidate prefusion-stabilizing mutations of S **(f)** relative to WT at 3 hours post-mixing with hACE2-expressing cells. Abbreviations for combinatorial mutations are as follows: 2P, K986P/V987P; 2PQ, K986P/V987P/D994Q; 2PR, K986P/V987P/Q1005R; 2PQR, K986P/V987P/D994Q/Q1005R. Fold changes are shown as mean ± range. Data are from *n* = 2 independent replicates. Source data are available.

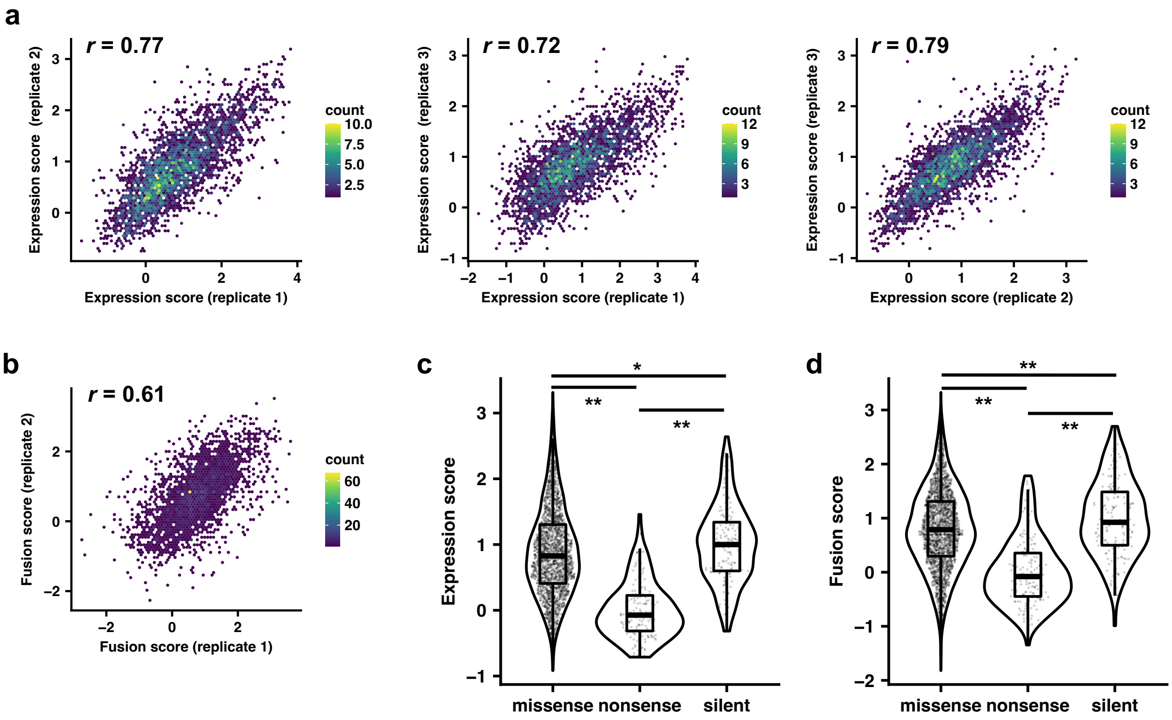


e

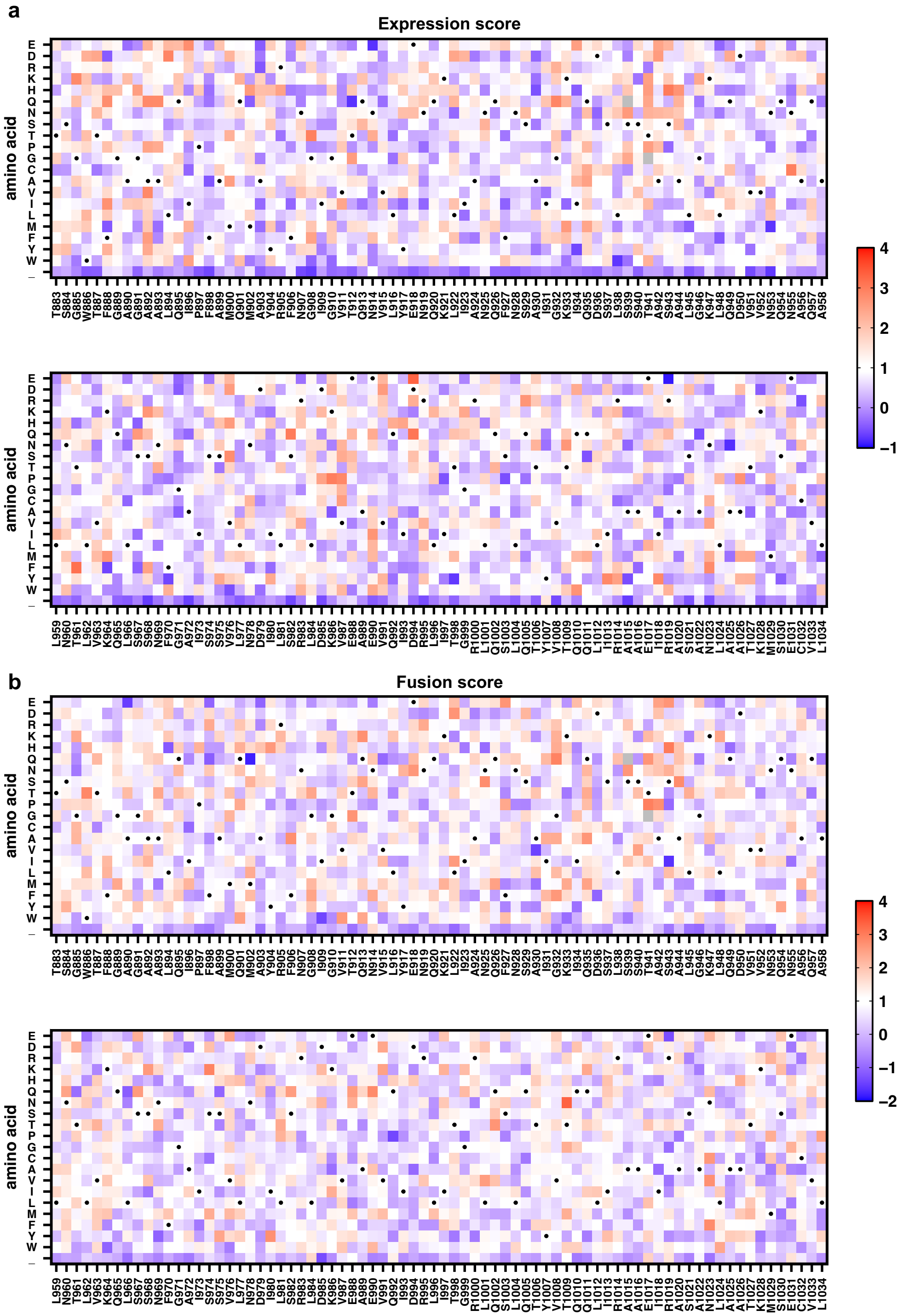
Antibody	Epitope	K_D (nM)		k_{on} (× 10^6 M^{-1} s^{-1})		k_{off} (× 10^{-3} s^{-1})	
		2P	2PQ	2P	2PQ	2P	2PQ
CC12.3	RBD	1.81 ± 0.06	1.87 ± 0.07	2.11 ± 0.07	2.18 ± 0.08	3.82 ± 0.04	4.08 ± 0.05
CC40.8	S2 stem helix	1.83 ± 0.05	2.39 ± 0.06	1.47 ± 0.04	1.12 ± 0.03	2.69 ± 0.03	2.67 ± 0.03
COVA1-07	S2 HR1	2.75 ± 0.11	31.73 ± 5.16	0.83 ± 0.03	0.21 ± 0.03	2.27 ± 0.04	6.67 ± 0.27

Fig. 4 | Biophysical characterization of 2PQ spike. a,b, Electron density map (colored grey) of 2PQ fitted on S with all-down RBD (PDB: 6VXX) **(a)**, and one-up RBD (PDB: 6VYB) **(b)**. **c,** The first differential curves for the relative fluorescence unit (RFU) from differential scanning fluorimetry with respect to temperature are shown for soluble 2PQ and 2P. Grey dotted line indicates the first apparent melting temperature (T_m) of 2P and 2PQ at approximately 46.5 °C; blue dotted line indicates the second apparent T_m of 2PQ at approximately 62 °C. **d,** D994Q allows formation of an additional intraprotomer hydrogen bond as shown by structural modelling. Distinct protomers are in grey, light blue and pink. The Q758 and Q994 side chains are shown as sticks representation. Hydrogen bond is indicated in black dashed line with the distance indicated. **e,** Summary of binding kinetics data from biolayer interferometry (BLI) experiments. Dissociation constants (K_D), on-rates (k_{on}) and off-rates (k_{off}) are shown as mean ± standard error. Source data are available.

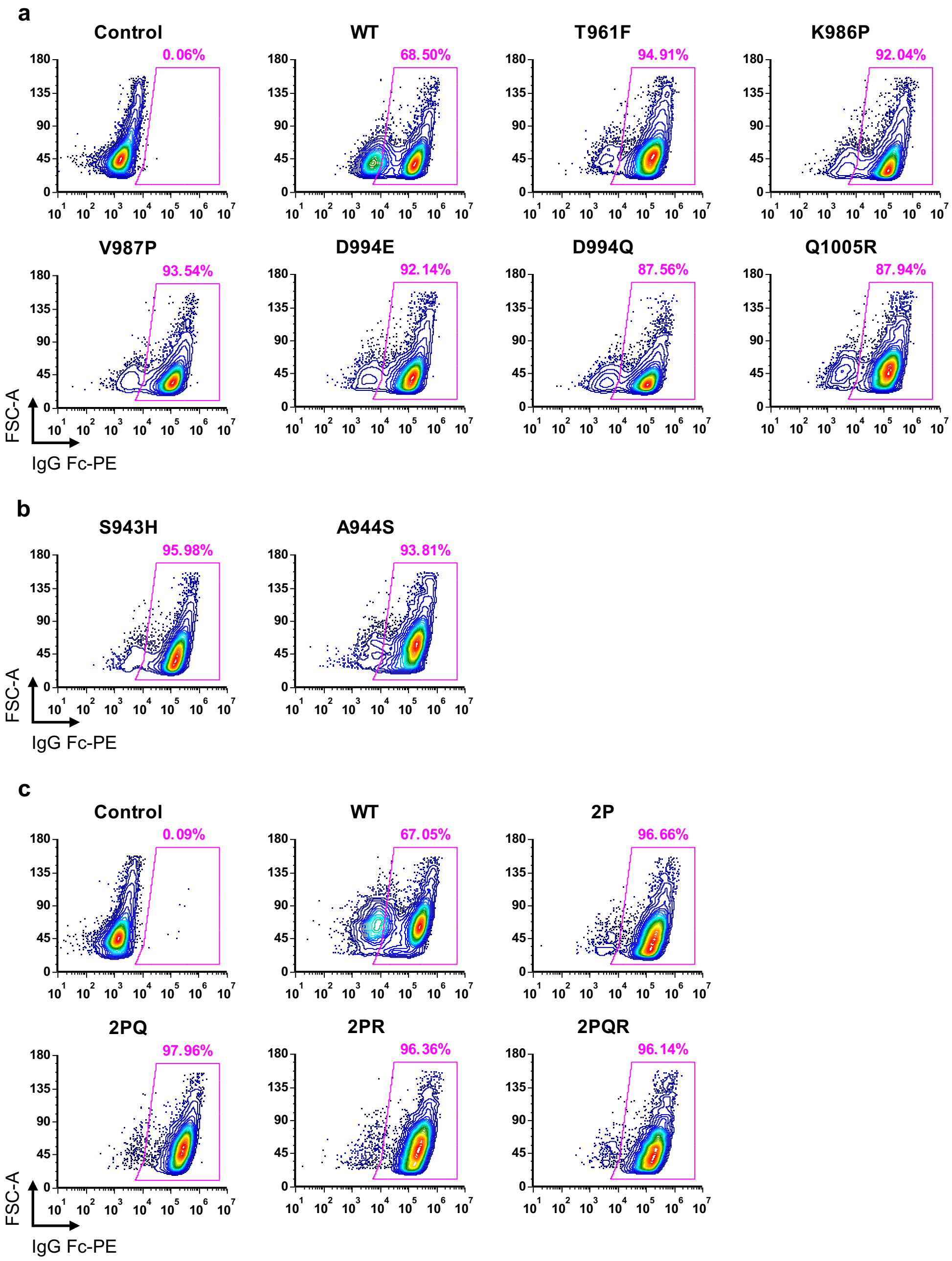
EXTENDED DATA FIGURES



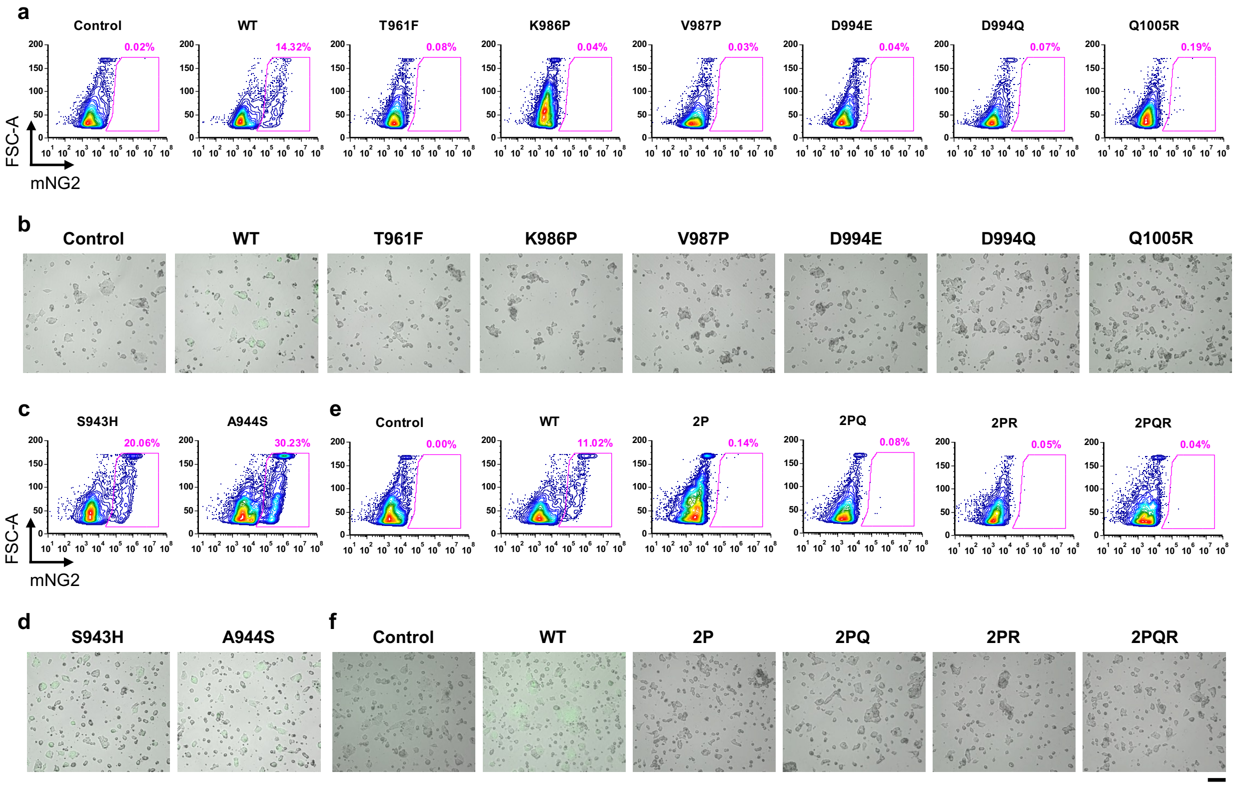
Extended Data Fig. 1 | Additional analyses for expression and fusion assays. a,b, Correlation of expression scores **(a)** and fusion scores **(b)** between replicates. Pearson correlation coefficient, *r*, is shown for each plot. **c,d,** Violin plots of expression scores **(c)** and fusion scores **(d)** of missense, nonsense, and silent mutations are compared. Box and whisker plots are also shown. *, $p < 0.05$; **, $p < 0.001$. Exact *p*-values from two-sided *t* test are shown in **Supplementary Table 6**. Source data are available as **Supplementary Table 1**.



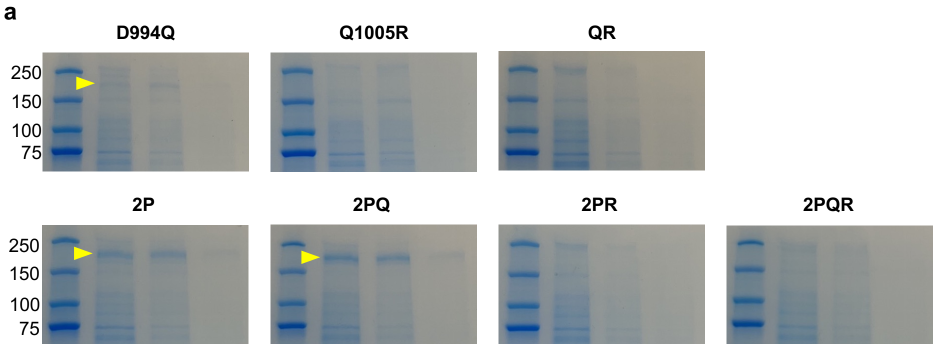
Extended Data Fig. 2 | Heatmaps of expression and fusion scores. a-b, Expression scores **(a)**, and fusion scores **(b)** of all amino-acid mutations in the DMS library. Underscore (_) in amino acid represents a nonsense mutation. Dots indicate WT. Grey squares indicate mutations that were not observed. Expression and fusion scores shown are the mean of *n* = 3 and *n* = 2 independent replicates, respectively. Lower expression score implies lower surface expression of the S protein, whereas lower fusion score suggests fusogenicity is impaired. Abbreviations of amino acid residues are as follows: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine. Source data are available as **Supplementary Table 1**.

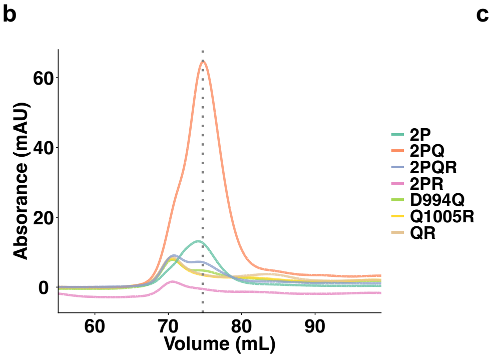


Extended Data Fig. 3 | Validation of surface expression of mutants of interest. a-c, Flow cytometry plots showing surface expression of candidate prefusion-stabilizing mutations **(a)**, fusion-enhancing mutations **(b)**, and combinations of candidate prefusion-stabilizing mutations **(c)**. Plots are representative of *n* = 2 independent replicates. Abbreviations for combinatorial mutations are as follows: 2P, K986P/V987P; 2PQ, K986P/V987P/D994Q; 2PR, K986P/V987P/Q1005R; 2PQR, K986P/V987P/D994Q/Q1005R.



Extended Data Fig. 4 | Fusion activity of mutants of interest. a,b, Flow cytometry plots **(a)** and micrographs **(b)** of S-expressing cells with fusion-incompetent mutations at 3 hours post-mixing with hACE2-expressing cells. **c,d,** Flow cytometry plots **(c)** and micrographs **(d)** of S-expression cells with fusion-enhancing mutations at 3 post-mixing with hACE2-expressing cells. **e,f,** Flow cytometry plots **(e)** and micrographs **(f)** of S-expression cells with combinations of fusion-incompetent mutations at 3 hours post-mixing with hACE2-expressing cells. Plots and micrographs are representative of *n* = 2 independent replicates. Abbreviations of combinatorial mutations are as follows: 2P, K986P/V987P; 2PQ, K986P/V987P/D994Q; 2PR, K986P/V987P/Q1005R; 2PQR, K986P/V987P/D994Q/Q1005R. Scale bar: 100 µm.

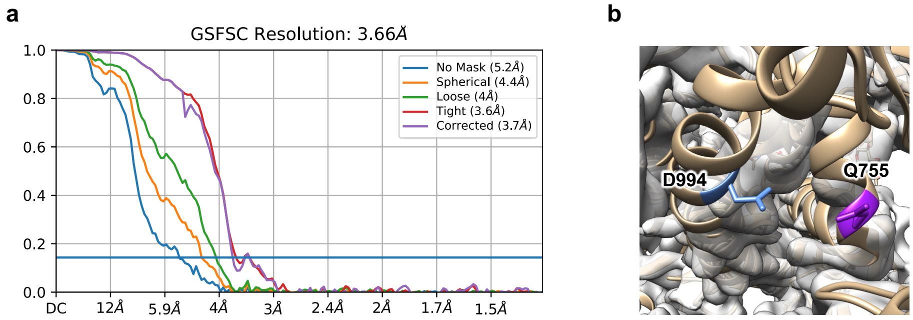




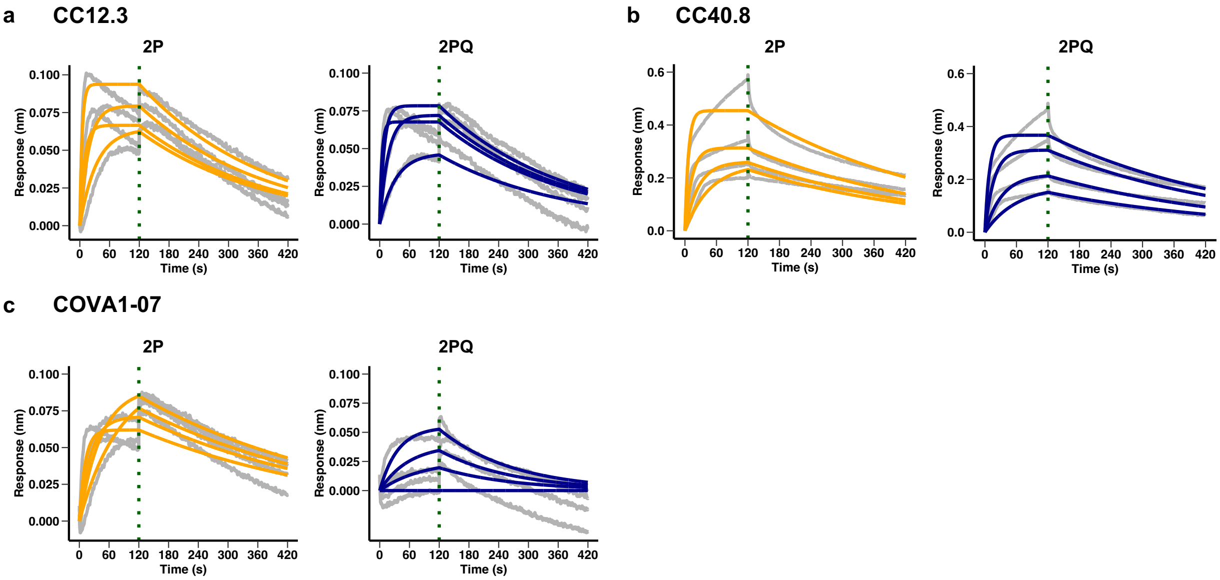
c

Mutant	Culture volume (mL)	Yield (µg per mL of culture)
D994Q	25	0.69
Q1005R	25	Undetectable
QR	25	Undetectable
2P	100	0.66
2PQ	100	2.24
2PR	25	0.05
2PQR	200	0.18

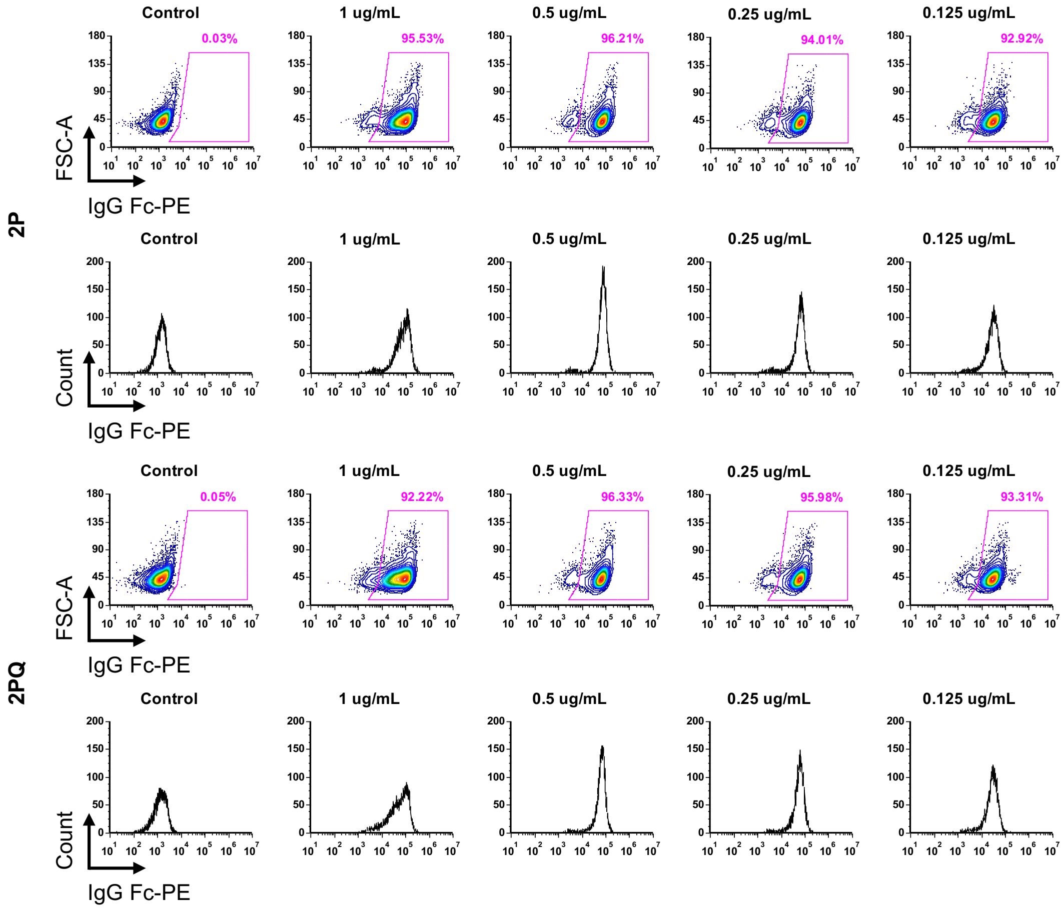
Extended Data Fig. 5 | Purification of S ectodomain mutants. a, SDS-PAGE gels of soluble spike proteins after His-tag affinity purification. Molecular weights (in kDa) of protein standards are shown. Arrowheads point to the band corresponding to the size of a glycosylated S monomer (~180 kDa). **b,** Chromatograms of soluble spike ectodomain mutants for biophysical characterization. Dotted line indicates the peak volume (~74 mL) corresponding to the size of a glycosylated spike trimer (~540 kDa). Fractions from 73 mL to 79 mL were collected and concentrated. **c,** Culture volume and yield for each mutant in **b**. Abbreviations for combinatorial mutations are as follows: QR, D994Q/Q1005R; 2P, K986P/V987P; 2PQ, K986P/V987P/D994Q; 2PR, K986P/V987P/Q1005R; 2PQR, K986P/V987P/D994Q/Q1005R. Source data are available.



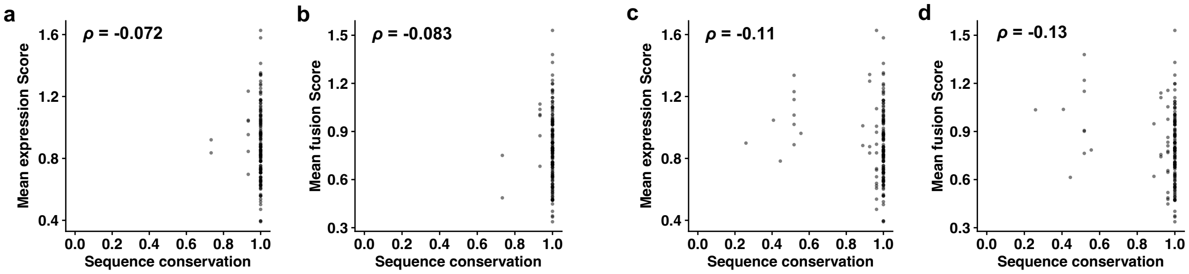
Extended Data Fig. 6 | Cryo-EM structure validation and analysis. a, Gold-standard Fourier shell correlation curve of soluble 2PQ ectodomain. **b,** Fit of cryo-EM density map of S with 2PQ mutations (in grey) to resolved structure of S with 2P mutations (PDB: 6VXX). Helices in 2PQ as indicated by the electron density map are closer compared to those in 2P. Contour level of density map was set to 0.2 in UCSF Chimera.



Extended Data Fig. 7 | Binding kinetics of SARS-CoV-2 S antibodies to soluble 2P and 2PQ. **a-c,** Sensorgrams for biolayer interferometry experiments to quantify binding kinetics of soluble 2P and 2PQ with CC12.3, an RBD antibody **(a)**, CC40.8, an S2 stem helix antibody **(b)**, and COVA1-07, an S2 HR1 antibody **(c)**. Grey lines correspond to raw data. Orange (2P) and blue (2PQ) lines correspond to fitted curves. The green vertical dotted line marks the end of association step and the beginning of dissociation step. Source data are available.



Extended Data Fig. 8 | Binding of S2M28 to membrane-bound 2P and 2PQ. Flow cytometry plots and histograms showing binding of S2M28, an NTD antibody, to membrane-bound 2P or 2PQ. Varying concentrations of S2M28 are indicated above each plot. Plots are representative of *n* = 2 independent replicates.



Extended Data Fig. 9 | Sequence conservation analysis of S2 HR1 and CH mutations. a,b, Plots of mean expression score **(a)** and mean fusion score **(b)** against natural frequency in HR1 and CH of major SARS-CoV-2 variants (**Extended Data Table 1**). **c,d,** Plots of mean expression score **(c)** and mean fusion score **(d)** against sequence conservation of the S2 HR1 and CH regions of related betacoronaviruses listed in **Supplementary Table 6**. Spearman correlation coefficients, *ρ*, are shown in **a-d**. Source data are available.

EXTENDED DATA TABLES

Extended Data Table 1. Mutations that are included in our DMS library (residues 883 to 1034) and also found in SARS-CoV-2 variants of concern and variants of interest.

Variants of concern		
PANGO Lineage	**Mutations**	**GenBank Reference**
Alpha (B.1.1.7)	S982A	MZ344997.1
Beta (B.1.351)	-	MW598419.1
Gamma (P.1)	T1027I	MZ169911.1
Delta (B.1.617.2)	D950N	MZ359841.1
Omicron (B.1.1.529: BA.1)	Q954H, N969K, L981F	OL672836.1
Omicron (B.1.1.529: BA.2, BA.4, BA.5)	Q954H, N969K	OM685375.1 (BA.2) ON373214.1 (BA.4) ON249995.1 (BA.5)
Variants of interest		
PANGO Lineage	**Mutations**	**GenBank Reference**
Epsilon (B.1.427, B.1.429)	-	MW453103.1
Zeta (P.2)	-	MW523796.1
Eta (B.1.525)	F888L	MW560924.1
Iota (B.1.526)	-	MW643362.1
Kappa (B.1.617.1)	-	MW966601.1
Lambda (C.37)	-	MW850639.1
Mu (B.1.621)	D950N	EPI_ISL_1220045 (GISAID)

Extended Data Table 2. Cryo-EM data collection statistics.

	2PQ spike (EMDB-xxxx)
Data collection and processing	
Magnification	130,000
Voltage (kV)	300
Electron exposure (e$^-$/Å^2)	50
Defocus range (μm)	-0.8 to -1.5
Pixel size (Å)	0.66
Symmetry imposed	C1
Initial particle images (no.)	238,524
Final particle images (no.)	140,183
Map resolution (Å)	3.66
FSC threshold	0.143
Map resolution range (Å)	N/A

REFERENCES

1. Cai, Y. *et al.* Distinct conformational states of SARS-CoV-2 spike protein. *Science* **369**, 1586-1592 (2020).

2. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).

3. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562-569 (2020).

4. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215-220 (2020).

5. Chiliveri, S. C., Louis, J. M., Ghirlando, R. & Bax, A. Transient lipid-bound states of spike protein heptad repeats provide insights into SARS-CoV-2 membrane fusion. *Sci. Adv.* **7**, eabk2226 (2021).

6. Marcink, T. C. *et al.* Intermediates in SARS-CoV-2 spike–mediated cell entry. *Sci. Adv.* **8**, eabo3153 (2022).

7. Dodero-Rojas, E., Onuchic, J. N. & Whitford, P. C. Sterically confined rearrangements of SARS-CoV-2 Spike protein control cell invasion. *eLife* **10**, e70362 (2021).

8. Walls, A. C. *et al.* Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc. Natl. Acad. Sci. USA* **114**, 11157-11162 (2017).

9. Gorgun, D., Lihan, M., Kapoor, K. & Tajkhorshid, E. Binding mode of SARS-CoV-2 fusion peptide to human cellular membrane. *Biophys. J.* **120**, 2914-2926 (2021).

10. Koppisetti, R. K., Fulcher, Y. G. & Van Doren, S. R. Fusion peptide of SARS-CoV-2 spike rearranges into a wedge inserted in bilayered micelles. *J. Am. Chem. Soc.* **143**, 13205-13211 (2021).

11. Cerutti, G. *et al.* Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host Microbe* **29**, 819-833.e7 (2021).

12. Qing, E. *et al.* Inter-domain communication in SARS-CoV-2 spike proteins controls protease-triggered cell entry. *Cell Rep.* **39**, 110786 (2022).

13. Barnes, C. O. *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682-687 (2020).

14. Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450-456 (2020).

15. Yuan, M., Liu, H., Wu, N. C. & Wilson, I. A. Recognition of the SARS-CoV-2 receptor binding domain by neutralizing antibodies. *Biochem. Biophys. Res. Commun.* **538**, 192-203 (2021).

16. Bowen, J. E. *et al.* SARS-CoV-2 spike conformation determines plasma neutralizing activity. *bioRxiv*, doi:10.1101/2021.12.19.473391 (2021).

17. Pallesen, J. *et al.* Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc. Natl. Acad. Sci. USA* **114**, E7348-E7357 (2017).

18. Yuan, Y. *et al.* Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat. Commun.* **8**, 15092 (2017).

19. Kirchdoerfer, R. N. *et al.* Pre-fusion structure of a human coronavirus spike protein. *Nature* **531**, 118-121 (2016).

20. Gilbert, P. B. *et al.* Immune correlates analysis of the mRNA-1273 COVID-19 vaccine efficacy clinical trial. *Science* **375**, 43-50 (2022).

21. Skowronski, D. M. & De Serres, G. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New Engl. J. Med.* **384**, 1576-1577 (2021).

22. Sadoff, J. *et al.* Safety and efficacy of single-dose Ad26.COV2.S vaccine against Covid-19. *New Engl. J. Med.* **384**, 2187-2201 (2021).

23. Heath, P. T. *et al.* Safety and efficacy of NVX-CoV2373 Covid-19 vaccine. *New Engl. J. Med.* **385**, 1172-1183 (2021).

24. Hsieh, C.-L. *et al.* Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **369**, 1501-1505 (2020).

25. Juraszek, J. *et al.* Stabilizing the closed SARS-CoV-2 spike trimer. *Nat. Commun.* **12**, 244 (2021).

26. Riley, T. P. *et al.* Enhancing the prefusion conformational stability of SARS-CoV-2 spike protein through structure-guided design. *Front. Immunol.* **12**, 660198 (2021).

27. Olmedillas, E. *et al.* Structure-based design of a highly stable, covalently-linked SARS-CoV-2 spike trimer with improved structural properties and immunogenicity. *bioRxiv*, doi:10.1101/2021.05.06.441046 (2021).

28. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874-882 (2018).

29. Suiter, C. C. *et al.* Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc. Natl. Acad. Sci. USA* **117**, 5394-5401 (2020).

30. Chiasson, M. A. *et al.* Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife* **9**, e58026 (2020).

31. Matreyek, K. A., Stephany, J. J., Chiasson, M. A., Hasle, N. & Fowler, D. M. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Res.* **48**, e1 (2020).

32. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45**, e102 (2017).

33. Ouyang, W. O. *et al.* Probing the biophysical constraints of SARS-CoV-2 spike N-terminal domain using deep mutational scanning. *bioRxiv*, doi:10.1101/2022.06.20.496903 (2022).

34. Kondo, N., Miyauchi, K., Meng, F., Iwamoto, A. & Matsuda, Z. Conformational changes of the HIV-1 envelope protein during membrane fusion are inhibited by the replacement of its membrane-spanning domain. *J. Biol. Chem.* **285**, 14681-14688 (2010).

35. Baviskar, P. S., Hotard, A. L., Moore, M. L. & Oomens, A. G. The respiratory syncytial virus fusion protein targets to the perimeter of inclusion bodies and facilitates filament formation by a cytoplasmic tail-dependent mechanism. *J. Virol.* **87**, 10730-10741 (2013).

36. Atanasiu, D. *et al.* Dual split protein-based fusion assay reveals that mutations to herpes simplex virus (HSV) glycoprotein gB alter the kinetics of cell-cell fusion induced by HSV entry glycoproteins. *J. Virol.* **87**, 11332-11345 (2013).

37. Meng, B. *et al.* SARS-CoV-2 Spike N-Terminal Domain modulates TMPRSS2-dependent viral entry and fusogenicity. *Cell Rep.* **40**, 111220 (2022).

38. Meng, B. *et al.* Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity. *Nature* **603**, 706-714 (2022).

39. Bradel-Tretheway, B. G. *et al.* Nipah and Hendra virus glycoproteins induce comparable homologous but distinct heterologous fusion phenotypes. *J. Virol.* **93**, e00577-19 (2019).

40. Cabantous, S., Terwilliger, T. C. & Waldo, G. S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* **23**, 102-107 (2005).

41. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801-807 (2014).

42. Feng, S. *et al.* Improved split fluorescent proteins for endogenous protein labeling. *Nat. Commun.* **8**, 370 (2017).

43. Yuan, M. *et al.* Structural basis of a shared antibody response to SARS-CoV-2. *Science* **369**, 1119-1123 (2020).

44. Zhou, P. *et al.* A human antibody reveals a conserved site on beta-coronavirus spike proteins and confers protection against SARS-CoV-2 infection. *Sci. Transl. Med.* **14**, eabi9215 (2022).

45. Krammer, F. SARS-CoV-2 vaccines in development. *Nature* **586**, 516-527 (2020).

46. Klasse, P. J., Nixon, D. F. & Moore, J. P. Immunogenicity of clinically relevant SARS-CoV-2 vaccines in nonhuman primates and humans. *Sci. Adv.* **7**, eabe8065 (2021).

47. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332-2347.e16 (2021).

48. Claireaux, M. *et al.* A public antibody class recognizes an S2 epitope exposed on open conformations of SARS-CoV-2 spike. *Nat. Commun.* **13**, 4539 (2022).

49. Henderson, R. *et al.* Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nat. Struct. Mol. Biol.* **27**, 925-933 (2020).

50. Sanders, R. W. & Moore, J. P. Virus vaccines: proteins prefer prolines. *Cell Host Microbe* **29**, 327-333 (2021).

51. Binley, J. M. *et al.* A recombinant human immunodeficiency virus type 1 envelope glycoprotein complex stabilized by an intermolecular disulfide bond between the gp120 and gp41 subunits is an antigenic mimic of the trimeric virion-associated structure. *J. Virol.* **74**, 627-643 (2000).

52. Sanders, R. W. *et al.* Stabilization of the soluble, cleaved, trimeric form of the envelope glycoprotein complex of human immunodeficiency virus type 1. *J. Virol.* **76**, 8875-8889 (2002).

53. Kong, L. *et al.* Uncleaved prefusion-optimized gp140 trimers derived from analysis of HIV-1 envelope metastability. *Nat. Commun.* **7**, 12040 (2016).

54. Sanders, R. W. *et al.* A Next-Generation Cleaved, Soluble HIV-1 Env Trimer, BG505 SOSIP.664 gp140, Expresses Multiple Epitopes for Broadly Neutralizing but Not Non-Neutralizing Antibodies. *PLoS Path.* **9**, e1003618 (2013).

55. McLellan, J. S. *et al.* Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody. *Science* **340**, 1113-1117 (2013).

56. Loomis, R. J. *et al.* Structure-based design of Nipah virus vaccines: a generalizable approach to paramyxovirus immunogen development. *Front. Immunol.* **11**, 842 (2020).

57. Hastie, K. M. *et al.* Structural basis for antibody-mediated neutralization of Lassa virus. *Science* **356**, 923-928 (2017).

58. Rutten, L. *et al.* Structure-based design of prefusion-stabilized filovirus glycoprotein trimers. *Cell Rep.* **30**, 4540-4550.e3 (2020).

59. Caradonna, T. M. & Schmidt, A. G. Protein engineering strategies for rational immunogen design. *npj Vaccines* **6**, 154 (2021).

60. Heinz, F. X. & Stiasny, K. Distinguishing features of current COVID-19 vaccines: knowns and unknowns of antigen presentation and modes of action. *npj Vaccines* **6**, 104 (2021).

61. Premkumar, L. *et al.* The receptor-binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. *Sci. Immunol.* **5**, eabc8413 (2020).

62. Wang, C. *et al.* A conserved immunogenic and vulnerable site on the coronavirus spike protein delineated by cross-reactive monoclonal antibodies. *Nat. Commun.* **12**, 1715 (2021).

63. Pinto, D. *et al.* Broad betacoronavirus neutralization by a stem helix-specific human antibody. *Science* **373**, 1109-1116 (2021).

64. Hurlburt, N. K. *et al.* Structural definition of a pan-sarbecovirus neutralizing epitope on the spike S2 subunit. *Commun. Biol.* **5**, 342 (2022).

65. Zhou, P. *et al.* Broadly neutralizing anti-S2 antibodies protect against all three human betacoronaviruses that cause severe disease. *bioRxiv*, doi:10.1101/2022.03.04.479488 (2022).

66. Kirchdoerfer, R. N. *et al.* Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci. Rep.* **8**, 15701 (2018).

67. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines—a new era in vaccinology. *Nat. Rev. Drug Discov.* **17**, 261-279 (2018).

68. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).

69. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell* **78**, 779-784.e5 (2020).

70. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643-2651 (2014).

71. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444-1448 (2020).

72. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620 (2014).

73. Ream, D. & Kiss, A. J. NCBI/GenBank BLAST Output XML Parser Tool. (2013).

74. Katoh, K., Misawa, K., Kuma, K. i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066 (2002).

75. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36-51 (2005).

76. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290-296 (2017).

77. Rodrigues, J. P., Teixeira, J. M., Trellet, M. & Bonvin, A. M. pdb-tools: a swiss army knife for molecular structures. *F1000Research* **7**, 1961 (2018).

78. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47-55 (2014).