

1 **A cell-free antibody engineering platform rapidly generates SARS-CoV-2**
2 **neutralizing antibodies**

3 Xun Chen^{1,#}, Matteo Gentili², Nir Hacohen^{2,3,4}, Aviv Regev^{1,5,6,7,#}

4

5 ¹ Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA

6 ² Broad Institute of MIT and Harvard, Cambridge, MA, USA

7 ³ Department of Medicine, Harvard Medical School, Boston, MA, USA

8 ⁴ Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA

9 ⁵ Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, USA

10 ⁶ Howard Hughes Medical Institute, Chevy Chase, MD, USA

11 ⁷ Current address: Genentech, 1 DNA Way, South San Francisco, CA, USA

12 [#] To whom correspondence should be addressed: xun@broadinstitute.org (X.C.), aregev@broadinstitute.org (A.R.)

13

14

15 **Abstract**

16 Antibody engineering technologies face increasing demands for speed, reliability and scale. We
17 developed CeVICA, a cell-free antibody engineering platform that integrates a novel generation
18 method and design for camelid heavy-chain antibody VHH domain-based synthetic libraries,
19 optimized *in vitro* selection based on ribosome display and a computational pipeline for binder
20 prediction based on CDR-directed clustering. We applied CeVICA to engineer antibodies against
21 the Receptor Binding Domain (RBD) of the SARS-CoV-2 spike proteins and identified >800
22 predicted binder families. Among 14 experimentally-tested binders, 6 showed inhibition of
23 pseudotyped virus infection. Antibody affinity maturation further increased binding affinity and
24 potency of inhibition. Additionally, the unique capability of CeVICA for efficient and
25 comprehensive binder prediction allowed retrospective validation of the fitness of our synthetic
26 VHH library design and revealed direction for future refinement. CeVICA offers an integrated
27 solution to rapid generation of divergent synthetic antibodies with tunable affinities *in vitro* and
28 may serve as the basis for automated and highly parallel antibody generation.

29 Antibodies and their functional domains play key roles in research, diagnostics and therapeutics.
30 Antibodies are traditionally made by immunizing animals with the desired target as antigen, but
31 such methods are time consuming, their outcome is often unpredictable, and their use is
32 increasingly restricted in the European Union ¹. Alternatively, antibodies can be generated and
33 selected *in vitro*, where libraries of antibody-encoding DNA, either fully synthetic or derived from
34 animals, are displayed *in vitro* followed by selection and recovery of those binding the intended
35 target ^{2,3}. However, broad application of such *in vitro* methods remains a challenge, possibly due
36 to throughput limitations and concerns over functional fitness and *in vivo* tolerance of antibodies
37 generated *in vitro* ⁴. Advances in antibody library design and construction, *in vitro* display and
38 selection methods, post-selection binder identification and maturation will all help increase the
39 utility of *in vitro* antibody generation ².

40 For typical antibodies, antigen binding is co-determined by the variable domains of both its heavy
41 chain (VH) and light chain (VL/VK), but camelids produce unconventional heavy-chain-only
42 antibodies that bind to antigens solely based on the variable domain of their heavy chain, the VHH
43 domain (also known as nanobodies). VHHs are increasingly used as functional antibody domains
44 because of their small size (~14 kD) ⁵ and high stability (T_m up to 90°C) ⁶. VHH libraries have
45 been successfully screened for binders by phage and yeast display ⁷⁻⁹. However, the screen
46 diversity of such cell-based systems is often limited by limited efficiency of DNA library delivery
47 into cells (typically <10¹⁰). Conversely, cell-free approaches, such as ribosome display ¹⁰, are not
48 limited by transfection efficiency and cell culture constraints. Despite the advantage, ribosome
49 display remains underutilized compared to cell-based display systems ² and recent efforts to build
50 *in vitro* system based on ribosome display alone produced inconsistent results ¹¹, suggesting that
51 further optimization is required.

52 To leverage the advantages of cell-free display, we developed CeVICA (**C**ell-free **V**VHH
53 **I**dentification using **C**lustering **A**nalysis) (**Fig. 1**), an integrated platform for *in vitro* VHH domain
54 antibody engineering, distinct from previous systems^{11–13}, that combines a novel design and
55 generation method for CDR-randomized VHH libraries, optimized ribosome display and selection
56 cycle with built-in background reduction, and a computational approach to perform global binder
57 prediction from post-selection libraries. CeVICA first takes as input a linear DNA library, in which
58 each sequence is unique and encodes for an artificial VHH with three fully-randomized CDRs, and
59 where the 5' and 3' ends of the DNA molecules contain elements required for downstream *in vitro*
60 ribosome display (**Fig. 1a, Materials and Methods**). Next, CeVICA uses ribosome display to link
61 genotype (RNAs transcribed from DNA input library that are stop codon free, and stall ribosome
62 at the end of the transcript) and phenotype (folded VHH protein tethered to ribosomes due to the
63 lack of stop codon in the RNA) (**Fig. 1b, Materials and Methods**). In each selection cycle (**Fig.**
64 **1c, Materials and Methods**), the displaying ribosomes bind to an immobilized target, followed
65 by RT-PCR of the RNA attached to the bound ribosomes, which leads to double stranded cDNA,
66 which is then *in vitro* transcribed/translated in a new round of ribosome display. The double
67 stranded DNA in any chosen round is sequenced to obtain full-length VHH sequences (**Fig. 1d,**
68 **Materials and Methods**). CeVICA then groups the sequences into clusters based on similarity of
69 their CDR sequences, such that each cluster represents a unique binding family (**Fig. 1e, Materials**
70 **and Methods**). Finally, one representative sequence from each cluster is synthesized and
71 characterized for specific downstream applications (**Fig. 1f, Materials and Methods**). The
72 combination of linear DNA libraries (**Fig. 1a**), ribosome display (**Fig. 1b**) and selection cycles
73 (**Fig. 1c**) allow display of libraries with much larger diversity ($>10^{10}$) than methods depending on
74 cells¹⁴ at similar experimental scale. As selection increases the representation of sequences

75 encoding binders, each binder sequence leads to a cluster of sequences in the output library.
76 Clustering following high throughput sequencing identifies them more efficiently than methods
77 that rely on the analysis of individual colonies or sequences^{7,8}, promising a more comprehensive
78 view of the landscape of binder potential, with minimal time and resources.

79 We made VHH libraries containing highly random CDRs, based on analysis of natural VHH
80 sequences and using a three-stage PCR and ligation process (**Fig. 1g**). First, to guide our VHH
81 library sequence design, we analyzed the sequence characteristics of 298 unique camelid VHHs
82 (representing natural VHHs) from the Protein Data Bank (PDB) (**table S1**), highlighting three
83 CDR regions, CDR1-3⁵, separated by four regions of low diversity, frame1-4 (**Fig. S1a**). The four
84 frames share high homology with human IGHV3-23 or IGHJ4 (**Fig. S2a,b**), and most of the
85 remaining non-identical residues are present in other human IGHV genes (**Fig. S2c**). We used
86 consensus sequences extracted from this profile to design VHH DNA templates encoding the four
87 frames (**Fig. 1g**), and included additional frames to the final mixture of frame templates (**Materials**
88 **and Methods**), based on well-characterized VHHs^{6,15}. The mixture of VHH frames serves as a
89 template in PCR reactions, where DNA oligonucleotides with a 5' NNB sequence were used to
90 introduce randomization in CDRs, while hairpin DNA oligonucleotides were used to block ligation
91 of one end of the PCR product (**Fig. 1g** and **Fig. S3, Materials and Methods**). We introduced 7
92 random amino acids for CDR1, 5 for CDR2, and 6, 9, 10 or 13 for CDR3 to match the most
93 commonly observed CDR lengths in natural VHHs. CDR3s longer than 13 amino acids only
94 account for a minority of natural VHHs (36%, **Fig. S1a, table S2**) and were not included in our
95 VHH library. CDRs randomized in earlier stages are subject to duplication in later stages that
96 reduces their diversity. We thus chose to randomize CDR2 first, followed by CDR1, and then
97 CDR3, imposing a diversity hierarchy of CDR3>CDR1>CDR2, because this is the overall ranking

98 of diversity we observed in CDRs in natural VHHs (**Fig. S1a,c**). The sequence profile of the
99 resulting randomized VHH library met our design objectives, and largely mirrored the sequence
100 features of natural VHHs (**Fig. S1** and **table S2**). Finally, the VHH DNA library contains an
101 upstream T7 promoter to allow transcription of VHH RNA, a 3xMyc tag, and a spacer downstream
102 of the VHH coding region that stalls peptide release, to enable ribosome display (**Fig. 1h**).

103 To test the performance of our library in ribosome display, and to reduce unproductive sequences,
104 such as VHHs that contain frame shifts or early stops, we ribosome displayed a library only with
105 randomized CDR1 and CDR2 and performed one round of anti-Myc selection. Functional VHH
106 sequences will express Myc tag at the C-terminal of VHH and are expected to be enriched after
107 anti-Myc selection. Indeed, there was a large decrease of unproductive sequences and an increase
108 of full-length VHHs (from 25.3% to 51.9%) after anti-Myc enrichment (**Fig. 1i**). At the DNA level,
109 there was an increase of all in-frame CDR1 DNA lengths and decrease of frame-shift lengths (**Fig.**
110 **1j**, arrows). We used the resulting full-length enriched CDR1 and 2 randomized library as PCR
111 template for randomization of CDR3. The final library with all three CDRs randomized (hereafter,
112 “the input library”) contained 27.5% full-length sequences, and 3.68×10^{11} full-length diversity per
113 μg of library DNA.

114 We performed *in vitro* selection from the input library for sequences that encode binders to two
115 target proteins: EGFP and the receptor binding domain (RBD) of the spike protein of SARS-CoV-
116 2¹⁶ (**Fig. 2**). We fused each of the two proteins with a 3xFlag tag and immobilized them on beads
117 coated with protein G and anti-Flag antibody (**Fig. 2a**). For each screen, we used input library
118 DNA corresponding to $\sim 1 \times 10^{11}$ full-length diversity, and performed 3 rounds of selection. After
119 round 3, RNA yield markedly increased in both screens (**Fig. S4a**) and the recovered sequences
120 were primarily composed of *E. coli* ribosomal RNAs and VHH library RNA (*e.g.*, **Fig. S4b**).

121 Comparing the input and output library sequences shows a marked increase in the proportion of
122 stop-free VHH sequences after 3 rounds of selection (**Fig. 2c**), fitting our expectation that
123 successful binding to targets depends on intact VHH structure.

124 We identified target specific binders by clustering CDR sequences enriched after selection into
125 families. First, we examined the distribution of the sequence match scores (**Materials and**
126 **Methods**) between randomly selected pairs of sequences within a CDR in a library, and compared
127 these distributions for each CDR between the input and output libraries (**Fig. 2b, Materials and**
128 **Methods**). In the pre-selection input libraries, the mean match score is low and the distribution is
129 unimodal, as expected given the randomization; whereas after selection, there is a multi-modal
130 distribution, with one low mode (similar to input) and at least one high mode (**Fig. 2b**), which is
131 further distinguished when combining the CDR1 and CDR2 match scores (**Fig. 2b**). This high
132 mode should reflect binders enriched by the selection rounds. Notably, sequences with a high
133 match score in one CDR are more likely to have a higher match score in other CDRs (**Fig S4c-f**).
134 We clustered the likely binder sequences exceeding a combined match score threshold (**Fig. 2b**,
135 dashed horizontal line), yielding 862 unique clusters for RBD and 71 for EGFP, with 52 clusters
136 shared by the two targets (**Fig. 2d, table S4 and 5**). The shared clusters likely target the shared
137 components (protein G, anti-Flag antibody) present on the solid support surfaces, and thus
138 represent background binders. Notably, RBD unique clusters span a wide range of cluster sizes
139 (**Fig. 2e**).

140 Focusing on RBD binders, we chose one representative VHH gene from each of the 14 top-ranking
141 RBD unique clusters and validated it for spike RBD binding and SARS-CoV-2 pseudovirus
142 neutralization (**Fig. 2f-h, Materials and Methods**). RBD binding ELISA assays of the 14 tested
143 VHHs (SR1-14) showed 3 strong binders (SR1,2,12), 7 weak binders (SR4,6,7,8,11,13,14) and 4

144 non-binders (**Fig. 2f,g**). SARS-CoV-2 S pseudotyped lentivirus neutralization assays revealed 6
145 VHHs inhibiting infection above 30% at 1 μ M (**Fig. 2h**), which included the 3 strong binders and
146 three of the weak binders (SR4,6,8).

147 We next compared input, output and natural CDR sequence distributions to assess whether starting
148 with a fully random CDR amino acid profile may be generally detrimental to the fitness of binders,
149 and whether selection mimics a natural amino acid distribution. In natural VHHs, CDR1 and
150 CDR2 are less diverse than CDR3 with an amino acid profile that favors certain residues (**Fig.**
151 **S1a,c**). Previous synthetic VHH library designs sought to recapitulate the CDR1 and CDR2 amino
152 acid preferences of natural VHHs^{8,11,13}, whereas we used fully-randomized NNB codons to encode
153 all CDR positions. In principle, such a design might be less ideal if the natural CDR1 and CDR2
154 amino acid profile is required for functional VHHs. To determine whether our fully random CDR
155 amino acid profile is detrimental to the fitness of binders, we compared the CDR amino acid profile
156 of 932 representative sequences across all unique clusters from both the EGFP and RBD output
157 libraries (“output binders”) (**Fig. S5**) to the sequence profiles of either the input library or natural
158 VHHs (**Fig. S1a,b**). We reasoned that if the amino acid profile in the input library leads to a
159 distribution of proteins that are less fit in binding, the binder selection process should shift this
160 distribution to a more fit profile in the output library, such that there is a low correlation between
161 the amino acid profiles of the input library and output binders. Surprisingly, there was an overall
162 smaller shift in CDR1 and CDR2 compared to CDR3, as indicated by higher r^2 values (**Fig. S6a-**
163 **c, mean $r^2 = 0.45, 0.51, \text{ and } 0.36$ respectively), and lower similarity distances (as the RMSE
164 relative to $y = x$ line, **Materials and Methods, Fig. S6d,e, RMSE = 2.96, 2.40 and 3.51**
165 respectively), implying that a fully random profile at CDR1 and CDR2 may not have had a
166 substantial binding fitness cost at most positions, whereas CDR3 not only shifted away from the**

167 input profile, it was even further shifted from the natural profile (**Fig. S6d,e**). Moreover,
168 correlation of amino acid profiles between output binders and natural VHHs are significantly less
169 than between output binders and input library at most CDR positions (**Fig. S6**). A few positions
170 (CDR1 position 7 and CDR3 position 1-3) had much lower input-output binders r^2 than others.
171 This suggests that these positions may benefit from specifically-designed amino acid profiles (to
172 adjust off diagonal amino acids percentages (**Fig. S6b**) accordingly), even though their input
173 distributions were not particularly distinct from the native sequence distribution compared to other
174 positions (**Fig. S6a,d**). Thus, the output binder CDR profile is predominantly influenced by the
175 input library rather than by selection towards a natural VHH profile, a natural VHH CDR amino
176 acid profile is not required for VHH binding properties, and a fully random CDR design offers
177 high diversity without a major binding fitness cost (although may have other fitness drawbacks *in*
178 *vivo*).

179 To perform affinity maturation, a critical stage in antibody development in animals, we designed
180 and performed an affinity maturation strategy based on CeVICA to increase the affinity of RBD
181 binding VHHs (**Fig. 3a, Materials and Methods**). We used error-prone PCR to introduce random
182 mutations across the full-length sequence of six selected VHHs (SR1,2,4,6,8,12) and generated the
183 mutagenized library. A library size of 4.18×10^{10} diversity (sufficient to contain the full diversity
184 of VHHs with three mutations per sequence) was used as input and three rounds of stringent
185 selection were performed. We sequenced the libraries pre- and post-affinity maturation, and
186 observed about 3 mutations in the pre-library and about 2 mutations in the post-library per
187 sequence (**Fig. 3a**). We calculated their position-wise amino acid profiles, and determined, for
188 each VHH, the change in each amino acid proportion at each position, generating a percent point
189 change table. We defined putative beneficial mutations as those with a percent point increase above

190 a set threshold (**Fig. 3b, Materials and Methods** and **table S6**), highlighting between 8 to 25
191 putative beneficial mutations for each of the selected VHHs. Finally, we assembled a list of
192 identified putative beneficial mutations for each VHH and incorporated different combinations of
193 them into each VHH parental sequence to generate multiple mutated variants of each VHH for
194 final assessment (**table S7**).

195 Variants in the SR4 and SR6 families had both increased binding and neutralization, while the SR2
196 and SR12 family variants had only increased neutralization but not binding, based on an ELISA
197 binding assay and a pseudotyped virus neutralization assay (**Fig. 3c,d**). Multiple VHH variants
198 outperformed VHH72, a previously described VHH antibody that neutralizes SARS-CoV-2
199 pseudoviruses (Wrapp et al., 2020), in binding (*e.g.*, SR12_c3), neutralization (*e.g.*, SR4_t6), or
200 both (*e.g.*, SR6_c3) (**Fig. 3c,d** and **table S8**). Neutralization and binding performance were poorly
201 correlated across variants ($r^2 = 0.07$), as previously reported¹⁷. However, when considering each
202 VHH family separately, trends were stronger, and neutralization and affinity were more highly
203 correlated for SR4 and SR6 VHHs (**Fig. 3e**). This may be because variants within the same family
204 share the same binding site and orientation. One intriguing hypothesis is that the slope of each
205 VHH family's linear trend reflects the sensitivity of the virus to the blocking of the family's
206 binding site. A dose response curve of selected VHHs showed SR6_c3 as the most potent
207 neutralizer (**Fig. 3f**) with an IC₅₀ of 62.7 nM (**Fig. 3g**), comparable to the Fab domains of potent
208 SARS-CoV-2 neutralizing antibodies identified from human patients¹⁸. Importantly, the original
209 SR6 cluster contained only 679 sequences, representing 0.67% of the 101,674 sequenced from the
210 initial selection output, highlighting the power of CeVICA in rapidly identifying high performance
211 antibodies among a vast number of potential candidates.

212 Finally, we examined the potential impact that our VHH sequences may have on immunogenicity
213 in humans, as a major concern related to the therapeutic use of VHH antibodies is the possibility
214 that, as camelid proteins, they would elicit an immune response. In particular, VHH hallmark
215 residues in frame2 constitute a major difference between camelid VHHs and human VHs (**Fig.**
216 **S2**). We used our affinity maturation data to identify potential conversion options for these VHH
217 hallmark residues. In three of the four VHH hallmark residues there were VHHs where the residues
218 were converted to the corresponding human residue as a result of affinity maturation (**Fig. S7**,
219 arrows). These data imply that at least some of the VHH hallmark residues can be converted to
220 human residues without loss of binding fitness. Such conversions may serve as frame features of
221 future VHH library designs and improve tolerance of *in vitro* engineered VHHs by humans.
222 Overall, the extension of CeVICA for affinity maturation offers a strategy for improving antibody
223 function and additional iterations of the affinity maturation process may provide further
224 enhancement of antibody properties.

225 In conclusion, CeVICA is a new system for synthetic VHH based antibody library design, *in vitro*
226 selection optimization, post-selection screening, and affinity maturation. Using CeVICA, we
227 generated a large collection of antibodies that can bind the RBD domain of the SARS-CoV-2 spike
228 protein and can neutralize pseudotyped virus infection, thus providing an important resource.
229 Given its seamlessly integrated procedure, CeVICA is amenable to automation and could provide
230 an important tool for antibody generation in a rapid, reliable and scalable manner. CeVICA further
231 provides a technology framework for incorporation of future refinements that could overcome
232 limitations of *in vivo* fitness of *in vitro* generated antibodies and overall efficiency.

233

234

235 **Materials and Methods**

236

237 **Constructs**

238 DNA encoding VHHs were obtained by gene synthesis (IDT) and cloned into pET vector in frame
239 with a C-terminal 6XHis tag by Gibson assembly (NEBuilder® HiFi DNA Assembly Master Mix,
240 New England Biolabs). DNA encoding SARS-CoV-2 S RBD (S a.a. 319-541) were obtained by
241 gene synthesis and cloned into pcDNA3 with an N-terminal SARS-CoV-2 S signal peptide (S a.a.
242 1-16) and a C-terminal 3xFlag tag by Gibson assembly. EGFP was cloned into pcDNA3 with a C-
243 terminal 3xFlag tag by Gibson assembly. SARS-CoV-2 S was amplified by PCR (Q5 High-
244 Fidelity 2X Master Mix, New England Biolabs) from pUC57-nCoV-S (kind gift from Jonathan
245 Abraham lab). SARS-CoV-2 S was deleted of the 27 a.a. at the C-terminal and fused to the
246 NRVRQGYS sequence of HIV-1, a strategy previously described for retroviruses pseudotyped
247 with SARS-CoV S¹⁹. Truncated SARS-CoV-2 S fused to gp41 was cloned into pCMV by Gibson
248 assembly to obtain pCMV-SARS2ΔC-gp41. psPAX2 and pCMV-VSV-G were previously
249 described²⁰. pTRIP-SFFV-EGFP-NLS was previously described²¹ (a gift from Nicolas Manel;
250 Addgene plasmid # 86677; <http://n2t.net/addgene:86677> ; RRID:Addgene_86677). cDNA for
251 human TMPRSS2 and Hygromycin resistance gene was obtained by synthesis (IDT). pTRIP-
252 SFFV-Hygro-2A-TMPRSS2 was obtained by Gibson assembly.

253

254 **Cell culture**

255 HEK293T cells were cultured in DMEM, 10% FBS (ThermoFisher Scientific), PenStrep
256 (ThermoFisher Scientific). HEK293T ACE2 were a kind gift of Michael Farzan. HEK293T ACE2
257 cells were transduced with pTRIP-SFFV-Hygro-TMPRSS2 to obtain HEK293T ACE2/TMPRSS2

258 cells. The transduced cells were selected with 320 µg/ml of Hygromycin (Invivogen) and used as
259 a target in SARS-CoV-2 S pseudotyped lentivirus neutralization assays. Transient transfection of
260 HEK293T cells was performed using TransIT®-293 Transfection Reagent (Mirus Bio, MIR 2700).

261

262 **Amino acid profile construction and analysis of natural VHHs**

263 VHH protein sequences were downloaded from the Protein Data Bank (only entries deposited prior
264 to Sep 2nd, 2020 were included; **table S1**). VHHs were separated into CDRs and frames (segments)
265 by finding regions of continuous sequence in each VHH that best matched to the following
266 standard frame sequences:

267 frame1 standard: EVQLVESGGGLVQAGDSLRLSCTASG,

268 frame2 standard: MGWFRQAPGKEREVVAIS,

269 frame3 standard: AFYADSVRGRFSISADSAKNTVYLQMNSLKPEDTAVYYCAA,

270 frame4 standard: DYWGQGTQVTVSS,

271 Each matched region is the corresponding frame of the VHH, the region between frame1 and
272 frame2 is CDR1, the region between frame2 and frame3 is CDR2, the region between frame3 and
273 frame4 is CDR3 (**Fig. 1g**). Only VHH sequences with at least one unique CDR were selected to
274 represent natural VHHs and used for constructing amino acid profile (a.a. profile). 298 sequences
275 fit this selection criteria (**table S1**). The amino acid (a.a.) profile at each position within each
276 segment was calculated by finding the percentage of each of the 20 universal proteinogenic amino
277 acid at that position among all selected VHHs, all frame lengths were set to the same length as
278 frame standards. CDR lengths were manually set to accommodate different CDR lengths, CDR1
279 and CDR2 lengths was set to 10, CDR3 length was set to 30. VHHs with CDR lengths shorter than
280 the corresponding set length had their CDR filled from the C-terminal end with empty position

281 holders up to the set length. Numbers in amino acid profile table are the percentage of each amino
282 acid.

283

284 **VHH library construction**

285 VHH libraries were constructed by ligation of PCR products in three stages, with each stage
286 randomizing one of the three CDRs. Primers used and PCR cycling conditions for each primer pair
287 are listed in **table S3**. At each stage, PCR was performed using a high-fidelity DNA polymerase
288 without strand displacement activity, using Phusion DNA polymerase (New England Biolabs,
289 M0530L). Importantly, 65°C was used as the elongation temperature to avoid hairpin opening
290 during DNA elongation. PCR products with correct size were purified by DNA agarose gel
291 extraction. Ligation and phosphorylation of PCR products were performed simultaneously using
292 T4 DNA ligase (New England Biolabs, M0202L) and T4 Polynucleotide Kinase (New England
293 Biolabs, M0201L). Ligation products with the correct size were purified by DNA agarose gel
294 extraction using NucleoSpin Gel and PCR Clean-Up Kit (Takara, 740609.250, this kit was used
295 for all DNA agarose gel extraction steps in this study). Purified ligation products were quantified
296 with Qubit 1X dsDNA HS Assay Kit (ThermoFisher Scientific, Q33230, this kit was used for all
297 Qubit measurements in this study) using Qubit 3 Fluorometer.

298

299 CDR2 was randomized in stage one, PCR templates at this stage were equal molar mixtures of
300 plasmids carrying DNA encoding frames, including three frame1 versions, one frame2, three
301 frame3 versions and one frame4. The three versions of frame1 and frame3 were derived from
302 consensus sequence extracted from natural VHH a.a. profile, the A3 VHH⁶ and a GFP binding
303 VHH¹⁵. Amino acid sequences of the frames are shown in **fig. S1**.

304
305 CDR1 was randomized in stage two, 200 ng of ligation product from the first stage were digested
306 by Not I-HF (New England Biolabs, R3189S) and heat denatured, the entire digestion product was
307 used as template for PCR in stage two. Ligation product of stage two was subject to one round of
308 ribosome display and anti-Myc selection (below), the entire recovered RNA was reverse
309 transcribed and PCR amplified and purified.

310
311 270 ng of this RT-PCR product was used as template for PCR in stage three to randomize CDR3.
312 Ligation product of stage three was purified by DNA agarose gel extraction. The purified ligation
313 product was then digested by DraI (New England Biolabs, R0129S) and a fragment of ~680 bp in
314 size was purified by DNA agarose gel extraction to get the final VHH library, referred to as the
315 input library.

316
317 **High throughput full-length sequencing of VHH library**
318 Sequencing libraries from VHH DNA libraries were prepared by two PCR steps using primers and
319 PCR cycling conditions listed in **table S3**. Equal mixtures of Phusion DNA polymerase (New
320 England Biolabs, M0530L) and Deep Vent DNA polymerase (New England Biolabs, M0258L)
321 were used for both PCRs to ensure efficient amplification. PCR cycle number was chosen to avoid
322 over-amplification and typically falls between 5 to 15.

323
324 In the first PCR, Illumina universal library amplification primer binding sequence and a stretch of
325 variable lengths of random nucleotides were introduced to the 5' end of library DNA. And
326 similarly, Illumina universal library amplification primer binding sequence and a stretch of

327 variable lengths of index sequence are introduced to the 3' end of library DNA. Eight different
328 lengths were used for both random nucleotides and index to create staggered VHH sequences in
329 the sequencing library, this arrangement is required for high quality sequencing of single amplicon
330 libraries on an Illumina Miseq instrument. The product of the first PCR was purified by column
331 clean-up using NucleoSpin Gel and PCR Clean-Up Kit and the entire sample was used as template
332 for the second PCR.

333

334 In the second PCR, Illumina universal library amplification primers were used to generate
335 sequencing library. Sequencing libraries were purified by DNA agarose gel extraction, quantified
336 using Qubit 3 Fluorometer, and sequenced on an Illumina Miseq instrument using MiSeq Reagent
337 Nano Kit v2 (500-cycles) (Illumina, MS-103-1003), no PhiX control library spike-in was used.
338 Sequencing run setup was: paired end 2X258 with no index read. Index in the library was designed
339 as inline index, so a separate index read was not required.

340

341 **Ribosome display**

342 VHH DNA library containing a specified amount of diversity was first amplified using a DNA
343 recovery primer pair listed in **table S3**. Equal mixtures of Phusion DNA polymerase (New England
344 Biolabs, M0530L) and Deep Vent DNA polymerase (New England Biolabs, M0258L) were used
345 for the PCR. PCR cycle number was chosen to avoid over-amplification and typically falls between
346 5 and 15. In a standard preparation, 200-500 ng of the purified PCR product was used as DNA
347 template in 25 µl of coupled *in vitro* transcription and translation reaction using PURExpress In
348 Vitro Protein Synthesis Kit (New England Biolabs, E6800L). The reaction was incubated at 37°C
349 for 30 minutes, then placed on ice, and 200 µl ice cold stop buffer (10 mM HEPES pH 7.4, 150

350 mM KCl, 2.5 mM MgCl₂, 0.4 µg/µl BSA (New England Biolabs, B9000S), 0.4 U/µl SUPERase•In
351 (ThermoFisher Scientific, AM2696), 0.05% TritonX-100) was then added to stop the reaction.
352 This stopped ribosome display solution was used for binding to immobilized protein targets during
353 *in vitro* selection. The amount of DNA template, volume of coupled *in vitro* transcription and
354 translation reaction, and volume of stop buffer were scaled proportionally when different volumes
355 of stopped ribosome display solution was needed. 1 to 8X standard preparations were used for
356 each selection cycle.

357

358 ***In vitro* selection**

359 Target proteins were immobilized to magnetic beads by first coating protein G magnetic beads
360 (ThermoFisher Scientific, 10004D) with anti-Flag antibody (Sigma-Aldrich, F1804), then
361 incubating antibody-coated beads with cell lysate or cell media containing 3xFlag tagged target
362 proteins at 4°C for 2 hours. For anti-Myc selection, magnetic beads were coated by anti-Myc
363 antibody (ThermoFisher Scientific, 13-2500) only. The beads were washed three times with PBST
364 (PBS, ThermoFisher Scientific, with 0.02% TritonX-100). Beads were then incubated with
365 stopped ribosome display solution at 4°C for 1 hour, and then washed 4 times with wash buffer
366 (10 mM HEPES pH 7.4, 150 mM KCl, 5 mM MgCl₂, 0.4 µg/µl BSA (New England Biolabs,
367 B9000S), 0.1U/µl SUPERase•In (ThermoFisher Scientific, AM2696), 0.05% TritonX-100). After
368 washing, beads were resuspended in TRIzol Reagent (ThermoFisher Scientific, 15596026), and
369 RNA was extracted from the beads, 25 µg of linear acrylamide (ThermoFisher Scientific,
370 AM9520) were used as co-precipitant during RNA extraction. Reverse transcription of extracted
371 RNA was performed using Maxima H Minus Reverse Transcriptase (ThermoFisher Scientific,
372 EP0752). The reverse transcription reaction was purified using SPRIselect Reagent (Beckman

373 Coulter, B23317) to obtain purified cDNA. Purified cDNA was amplified by PCR using equal
374 mixtures of Phusion DNA polymerase and Deep Vent DNA polymerase. PCR cycle number (**table**
375 **S3**) was chosen to avoid over-amplification and typically falls between 10 to 25. The PCR product
376 was purified by DNA agarose gel extraction. The purified PCR product was used for library
377 generation for high throughput full-length sequencing or as DNA template for ribosome display
378 reaction (coupled in vitro transcription and translation) to perform additional rounds of in vitro
379 selection.

380

381 **CDR-directed clustering analysis**

382 Computational analysis for CDR-directed clustering was performed using custom python scripts.
383 Paired end sequences were merged to form full-length VHH sequences. Merged VHH sequences
384 were quality trimmed and translated into VHH protein sequence, which were separated into CDRs
385 and frames (segments) as described in the *Amino Acid Profile Construction* section. Two VHHs
386 were determined to have similar CDRs via the following steps. First, the ungapped sequence
387 alignment score (match score) was calculated for each CDR of the two VHHs as the sum of
388 BLOSUM62²² amino acid pair scores at each aligned position. (If two CDRs have different
389 lengths, their sequence alignment score was set to -5 by default.) The alignment scores of any two
390 pairs of CDRs were summed to yield three scores, and if at least one of the three was larger than
391 35 (**Fig. 2b**), the two VHHs were defined as having similar CDRs. Next, VHHs with similar CDRs
392 were grouped by a two-step process. In the first step, we chose as VHH cluster-forming “seeds”
393 those VHHs that were called as similar to at least 5 other VHHs (all remaining VHHs were not
394 considered for clustering). In the second step, we iteratively selected a seed VHH with at least 5
395 other similar (>35 match score) seed VHHs, and grouped all of them into one cluster, removing

396 them from the seed VHH pool, and iterated this procedure until no seed VHHs remained. For RBD,
397 there were 83,433 seeds in the first step, and 83,392 were grouped in clusters in the second step.
398 For EGFP, 71,210 of 71,220 seeds were grouped in clusters (**table S9**). This heuristic was fast in
399 a standard computing environment with multiprocessing capabilities.

400

401 A representative sequence to illustrate each CDR in each cluster was chosen as the most frequent
402 CDR sequence in the cluster (the chose representatives for CDR1,2, and 3 may not necessarily be
403 from the same sequence, and are used only for illustrative purposes for each cluster as in **table S4**
404 and **S5**; whole VHH sequences were used for gene synthesis and all downstream experiments). A
405 consensus sequence was generated for each CDR, where each position in the CDR was represented
406 by a 6 character string, such that the first and fourth character were the single letter code for the
407 top and the second most abundant amino acid at the position, respectively, and the following two
408 characters (second and third for the most abundant; fifth and sixth for the second most abundant),
409 were their frequency, respectively (ranging from 00 for <34% to 99 for 100%). The consensus
410 sequence for a CDR was recorded as a single “B00” when the standard deviation of the lengths of
411 all CDRs was greater than 1. CDR scores were calculated by summing a score for each position in
412 the CDR consensus sequence, with scores of 3, 2, 1 for positions where the most abundant amino
413 acid had frequencies greater than 80%, 50%, or less, respectively, and a score of 0 for CDRs with
414 a consensus sequence of a single “B00” (**table S4** and **table S5**). Representative whole VHH
415 sequence for each cluster was selected as the one with the maximal sum of all CDR similarity
416 score between each VHH and all other VHHs in the cluster.

417

418 **Protein expression and purification**

419 Target proteins used for *in vitro* selection and ELISA were prepared by transiently transfecting
420 HEK293T cells with plasmids carrying either spike RBD with C-terminal 3xFlag tag and N-
421 terminal signal peptide of spike (RBD-3xFlag), or EGFP with C-terminal 3xFlag tag (EGFP-
422 3xFlag). Cell culture media (for RBD-3xFlag) or lysate of cell pellet (for EGFP-3xFlag) were used
423 for coating magnetic beads or plates. VHHs with C-terminal 6XHis tag (VHH-6XHis) were
424 purified by expressing in *E. coli.*, followed by purification using HisPur Cobalt Resin
425 (ThermoFisher Scientific, 89964). Briefly, VHH-6xHis plasmids were transformed into T7
426 Express *E. Coli.* (New England Biolabs, C2566I), single colonies were transferred into 10 ml LB
427 media and grown at 37°C for 2-4 hours (until OD reached 0.5-1), the culture was chilled on ice,
428 then IPTG was added to a final concentration of 10 µM. The culture was then incubated on an
429 orbital shaker at room temperature (RT) for 16 hours. Bacterial cells were pelleted by
430 centrifugation and lysed in B-PER Bacterial Protein Extraction Reagent (ThermoFisher Scientific,
431 78248) supplemented with rLysozyme (Sigma-Aldrich, 71110), DNase I (New England Biolabs,
432 M0303S), 2.5 mM MgCl₂ and 0.5 mM CaCl₂. Bacterial lysates were cleared by centrifugation and
433 mixed with wash buffer (50 mM Sodium Phosphate pH 7.4, 300 mM Sodium Chloride, 10 mM
434 imidazole) at 1:1 ratio, and then incubated with 40 µl HisPur Cobalt Resin for 2 hours at 4°C. The
435 resins were then washed 4 times with wash buffer. Proteins were eluted by incubating resin in
436 elution buffer (50 mM Sodium Phosphate pH 7.4, 300 mM Sodium Chloride, 150 mM imidazole)
437 at RT for 5 minutes. Purified protein samples were quantified by measuring absorbance at 280 nm
438 on a NanoDrop Spectrophotometer.

439

440 **ELISA assay for VHH binding to RBD**

441 Maxisorp plates (BioLegend, 423501) were coated with 1µg/ml anti-Flag antibody (Sigma
442 Aldrich, F1804) in coating buffer (BioLegend, 421701) at 4°C overnight. Plates were washed once
443 with PBST (PBS, ThermoFisher Scientific, with 0.02% TritonX-100), a 1:1 mixture of HEK293T
444 cell culture media containing secreted RBD-3xFlag and blocking buffer (PBST with 1% nonfat
445 dry milk) was added to the plates and incubated at RT for 1 hour. RBD coated plates were then
446 blocked with blocking buffer at RT for 1 hour. Plates were washed twice with wash buffer and
447 purified VHHs-6xHis diluted in blocking buffer were added to the plates and incubated at RT for
448 1 hour. Plates were washed three times with wash buffer, HRP conjugated anti-His tag secondary
449 antibody (BioLegend, 652503) diluted 1:2000 in blocking buffer was then added to the plates and
450 incubated at RT for 1 hour. Plates were washed three times with wash buffer and TMB substrate
451 (BD, 555214) was added to the plate and incubate at RT for 10 to 20 minutes. Stop buffer (1N
452 Sulfuric Acid) was added to the plates once enough color developed. Quantification of plates was
453 performed by measuring absorbance at 450 nm on a BioTek synergy H1 microplate reader. Data
454 reported were background subtracted. Two levels of background subtraction were performed: (1)
455 subtracting absorbance measured from wells incubated with blocking buffer only (without purified
456 VHHs-6xHis) from sample measurements (reflecting background absorbance by plates); and (2)
457 subtracting absorbance from each VHH incubated wells coated only with anti-Flag antibody and
458 without RBD (reflecting non-specific binding of each VHH).

459

460 **Pseudotyped SARS-CoV-2 lentivirus production and lentivirus production for transductions**

461 Lentivirus production was performed as previously described ²⁰. Briefly, HEK293T cells were
462 seeded at 0.8x10⁶ cells per well in a 6 well plate and were transfected the same day with TransIT®-
463 293 Transfection Reagent and a mix of DNA containing 1 µg psPAX, 1.6 µg pTRIP-SFFV-EGFP-

464 NLS and 0.4 μg pCMV-SARS2 Δ C-gp41. Medium was changed after overnight transfection.
465 SARS-CoV-2 S pseudotyped lentiviral particles were collected 30-34 hours post medium change
466 and filtered on a 0.45 μm syringe filter. To transduce HEK293T ACE2 the same protocol was
467 followed, with a mix containing 1 μg psPAX, 1.6 μg pTRIP-SFFV-Hygro-2A-TMPRSS2 and 0.4
468 μg pCMV-VSV-G.

469

470 **SARS-CoV-2 S pseudotyped lentivirus neutralization assay**

471 The day before the experiment, 5×10^3 HEK293T ACE2/TMPRSS2 cells per well were seeded in
472 96 well plates in 100 μl . On the day of lentivirus harvest, SARS-CoV-2 S pseudotyped lentivirus
473 was incubated with VHHs or VHH elution buffer in 96 well plates for 1 hour at RT (100 μl virus
474 + 50 μl of VHH at appropriate dilutions). Medium was then removed from HEK293T
475 ACE2/TMPRSS2 cells and replaced with 150 μl of the VHH + pseudotyped lentivirus solution.
476 Wells in the outermost rows of the 96 well plate were excluded from the assay. After overnight
477 incubation, medium was changed to 100 μl of fresh medium. Cells were harvested 40-44 hours
478 post infection with TrypLE (Thermo Fisher), washed in medium, and fixed in FACS buffer
479 containing 1% PFA (Electron Microscopy Sciences). Percentage GFP was quantified on a Cytotflex
480 LX (Beckman Coulter) and data were analyzed with FlowJo.

481

482 **Affinity maturation**

483 Error-prone PCR was used to introduce random mutations across the full length of selected VHH
484 DNA sequences. 0.1 ng of plasmid carrying DNA sequence encoding each selected VHH were
485 used as template in PCR reactions using Taq DNA polymerase with reaction buffer (10 mM Tris-
486 HCl pH 8.3, 50 mM KCl, 7mM MgCl₂, 0.5 mM MnCl₂, 1 mM dCTP, 1 mM dTTP, 0.2 mM dATP,

487 0.2 mM dGTP) suitable for causing mutations in PCR products. Mutagenized library for input to
488 CeVICA was made by ligating PCR products of error-prone PCR that carries VHH to DNA
489 fragment containing the remaining elements required for ribosome display. Three rounds of
490 ribosome display and *in vitro* selection were performed on the mutagenized library (pre-affinity
491 maturation, after error-prone PCR) as described in the *In vitro selection* section, during which the
492 incubation time of the binding step was kept between 5 seconds to 1 minute to impose a stringent
493 selection condition, additional error-prone PCR was not performed during the selection cycles.
494 The output library (post-affinity maturation) was sequenced along with the pre-affinity maturation
495 library as described in the *High throughput full-length sequencing of VHH library* section.

496

497 **Identification and ranking of beneficial mutations**

498 To identify potential beneficial mutations for each selected VHH we built an amino acid profile
499 (a.a. profile) table for each VHH family in the pre- and post-affinity maturation library, and
500 identified amino acids with increased frequency in the post-affinity maturation population
501 compared to their pre-maturation frequency. For each VHH parental sequence, an a.a. profile was
502 built of the percent of each a.a. across all VHH sequences originated from one parental VHH in
503 the pre-affinity maturation library (“pre-a.a. profile”) and in the post-affinity maturation library
504 (“post-a.a. profile”). A percent point change table was generated by subtracting the pre-a.a. profile
505 from the post-a.a. profile, describing the change of frequency of each observed amino acid at each
506 position of the VHH protein following affinity maturation.

507

508 We defined a putative beneficial mutation as either (1) the non-parental amino acid with the biggest
509 increase in frequency if its increase is at least 0.5 percentage points; the score is the difference

510 from the parental amino acid frequency; or (2) the non-parental amino acid with the biggest
511 increase after the parental amino acid if the increase is at least 1.5 percentage points; the score is
512 the percent point change of the beneficial mutation. To avoid too many proximal putative
513 beneficial mutations (which may cause structural incompatibility), a putative beneficial mutation
514 was discarded if it (1) is outside the CDRs; (2) is less than 3 positions away from another beneficial
515 mutation (“nearby mutation) and has a smaller beneficial mutation score than the nearby mutation;
516 and (3) co-occurs less than twice with the nearby mutation. From this final list of putative
517 beneficial mutations, different combinations were picked and incorporated into each VHH parental
518 sequence that include one combination of all beneficial mutations in CDRs, one combination of
519 the top-3 ranked (by beneficial mutation score) mutations in frames, and at least one combination
520 of both CDR mutations and frame mutations (**table S7**).

521

522

523 **References and Notes:**

- 524 1. Gray, A. C. *et al.* Animal-derived-antibody generation faces strict reform in accordance
525 with European Union policy on animal use. *Nat. Methods* **17**, 755–756 (2020).
- 526 2. Dübel, S., Stoevesandt, O., Taussig, M. J. & Hust, M. Generating recombinant antibodies
527 to the complete human proteome. *Trends Biotechnol.* **28**, 333–339 (2010).
- 528 3. Miersch, S. & Sidhu, S. S. Synthetic antibodies: Concepts, potential and practical
529 considerations. *Methods* **57**, 486–498 (2012).
- 530 4. Bradbury, A. R. M., Sidhu, S., Dübel, S. & McCafferty, J. Beyond natural antibodies: The
531 power of in vitro display technologies. *Nat. Biotechnol.* **29**, 245–254 (2011).
- 532 5. Muyldermans, S. Nanobodies: Natural single-domain antibodies. *Annu. Rev. Biochem.* **82**,
533 775–797 (2013).
- 534 6. Turner, K. B., Zabetakis, D., Goldman, E. R. & Anderson, G. P. Enhanced stabilization of
535 a stable single domain antibody for SEB toxin by random mutagenesis and stringent
536 selection. *Protein Eng. Des. Sel.* **27**, 89–95 (2014).
- 537 7. Huo, J. *et al.* Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block
538 interaction with ACE2. *Nat. Struct. Mol. Biol.* (2020). doi:10.1038/s41594-020-0469-6
- 539 8. McMahon, C. *et al.* Yeast surface display platform for rapid discovery of
540 conformationally selective nanobodies. *Nat. Struct. Mol. Biol.* **25**, 289–296 (2018).
- 541 9. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial
542 polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).
- 543 10. Hanes, J. & Plückthun, A. In vitro selection and evolution of functional proteins by using

- 544 ribosome display. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4937–4942 (1997).
- 545 11. Zimmermann, I. *et al.* Synthetic single domain antibodies for the conformational trapping
546 of membrane proteins. *Elife* **7**, 1–32 (2018).
- 547 12. Hanes, J., Schaffitzel, C., Knappik, A. & Plückthun, A. Picomolar affinity antibodies from
548 a fully synthetic naive library selected and evolved by ribosome display. *Nat. Biotechnol.*
549 **18**, 1287–1292 (2000).
- 550 13. Moutel, S. *et al.* NaLi-H1: A universal synthetic library of humanized nanobodies
551 providing highly functional antibodies and intrabodies. *Elife* **5**, 1–31 (2016).
- 552 14. He, M. & Taussig, M. J. Ribosome display: Cell-free protein display technology. *Briefings*
553 *Funct. Genomics Proteomics* **1**, 204–212 (2002).
- 554 15. Kirchhofer, A. *et al.* Modulation of protein properties in living cells using nanobodies.
555 *Nat. Struct. Mol. Biol.* **17**, 133–139 (2010).
- 556 16. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat
557 origin. *Nature* **579**, 270–273 (2020).
- 558 17. Rogers, T. F. *et al.* Isolation of potent SARS-CoV-2 neutralizing antibodies and protection
559 from disease in a small animal model. *Science* **7520**, eabc7520 (2020).
- 560 18. Hansen, J. *et al.* Studies in humanized mice and convalescent humans yield a SARS-CoV-
561 2 antibody cocktail. *Science* **0827**, eabd0827 (2020).
- 562 19. Moore, M. J. *et al.* Retroviruses Pseudotyped with the Severe Acute Respiratory
563 Syndrome Coronavirus Spike Protein Efficiently Infect Cells Expressing Angiotensin-
564 Converting Enzyme 2. *J. Virol.* **78**, 10628–10635 (2004).

- 565 20. Gentili, M. *et al.* Transmission of innate immune signaling by packaging of cGAMP in
566 viral particles. *Science* **349**, 1232–1236 (2015).
- 567 21. Raab, M. *et al.* ESCRT III repairs nuclear envelope ruptures during cell migration to limit
568 DNA damage and cell death. *Science* **352**, 359–362 (2016).
- 569 22. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks.
570 *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).

571

572

573 **Acknowledgements.** We thank Christopher M Vockley for critical reading and editing of the
574 manuscript, Matthew H Bakalar for helping with cloning VHH72, Leslie Gaffney and Anna
575 Hupalowska for assistance in figure making, Michael Farzan for providing HEK293T expressing
576 ACE2 and for discussing the SARS-CoV-2 S pseudotyped lentivirus neutralization approach,
577 Jonathan Abraham for providing the pUC57-nCov19-S plasmid. Work was supported by the
578 Klarman Cell Observatory and Klarman Incubator at the Broad Institute, NHGRI
579 5RM1HG006193 (A.R.) and HHMI (A.R.). M.G. is the recipient of an EMBO Long-Term
580 Fellowship (ALTF 486-2018) and a Cancer Research Institute/Bristol-Myers Squibb Fellow
581 (CRI2993). Until July 31, 2020, A.R. was an Investigator of the Howard Hughes Medical Institute.

582

583 **Author contributions.** X.C. and A.R. conceived the study. X.C. designed and developed the
584 CeVICA platform, performed selection and identification of EGFP and RBD binders, performed
585 affinity maturation of RBD binders. M.G. developed and performed SARS-CoV-2 S pseudotyped
586 lentiviruses neutralization assay. N.H. provided support for pseudotyped lentiviruses
587 neutralization assay. X.C. and A.R. wrote the manuscript, with contributions from all co-authors.

588

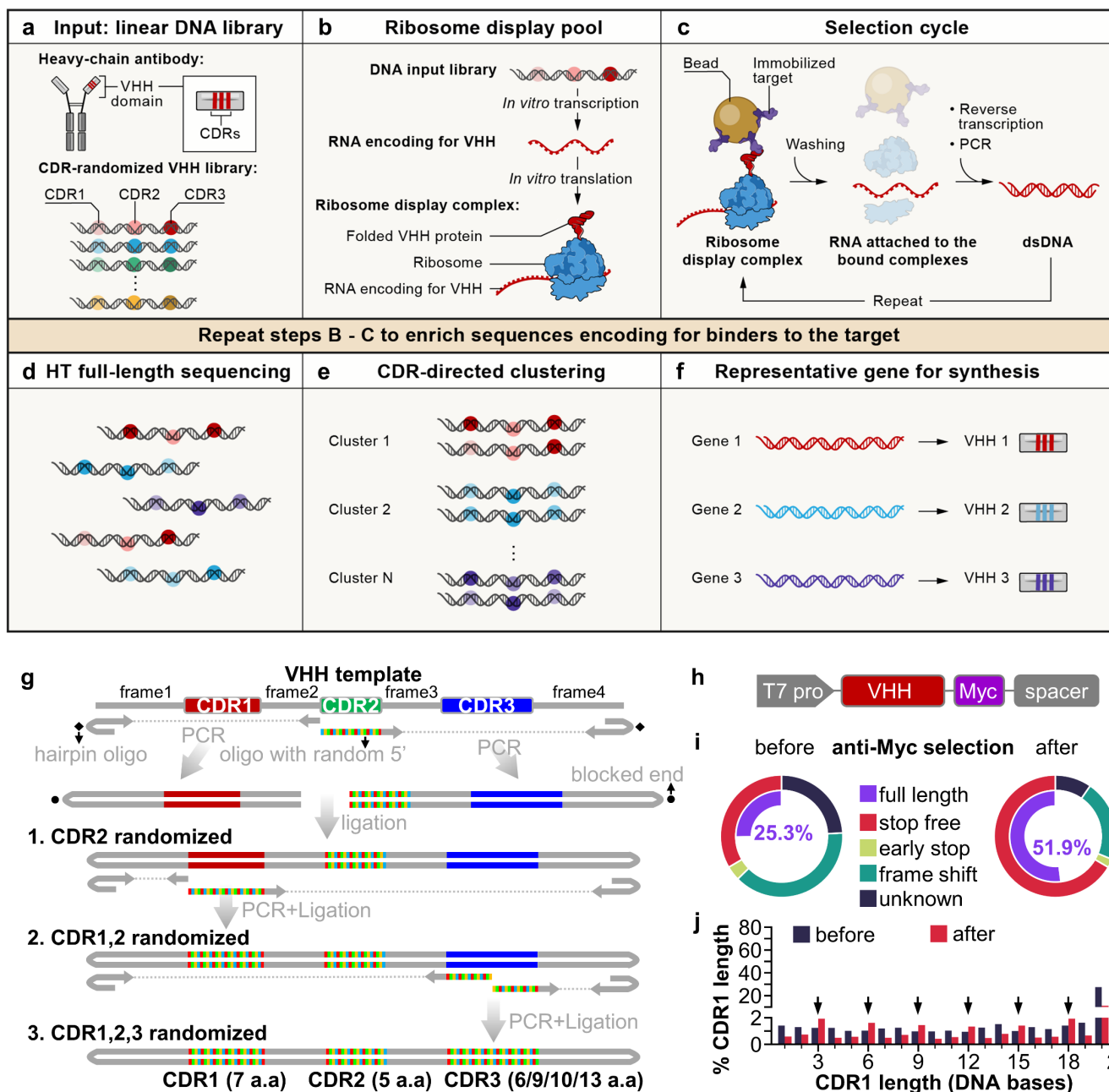
589 **Competing interests.** A.R. is a founder and equity holder of Celsius Therapeutics, an equity
590 holder in Immunitas Therapeutics and until August 31, 2020 was an SAB member of Syros
591 Pharmaceuticals, Neogene Therapeutics, Asimov and ThermoFisher Scientific. From August 1,
592 2020, A.R. is an employee of Genentech. N.H is an equity holder of BioNtech and is an advisor
593 for Related Sciences. X.C. and A.R. are named co-inventors on a patent application related to
594 CeVICA filed by the Broad Institute that is being made available in accordance with COVID-19
595 technology licensing framework to maximize access to university innovations.

596

597 **Data and materials availability.** Antibody sequences are in **table S7** and will be made publicly
598 available upon publication. Code for computational analysis will be available on Github. Key
599 plasmids generated in this study will be deposited in Addgene.

600

Fig 1



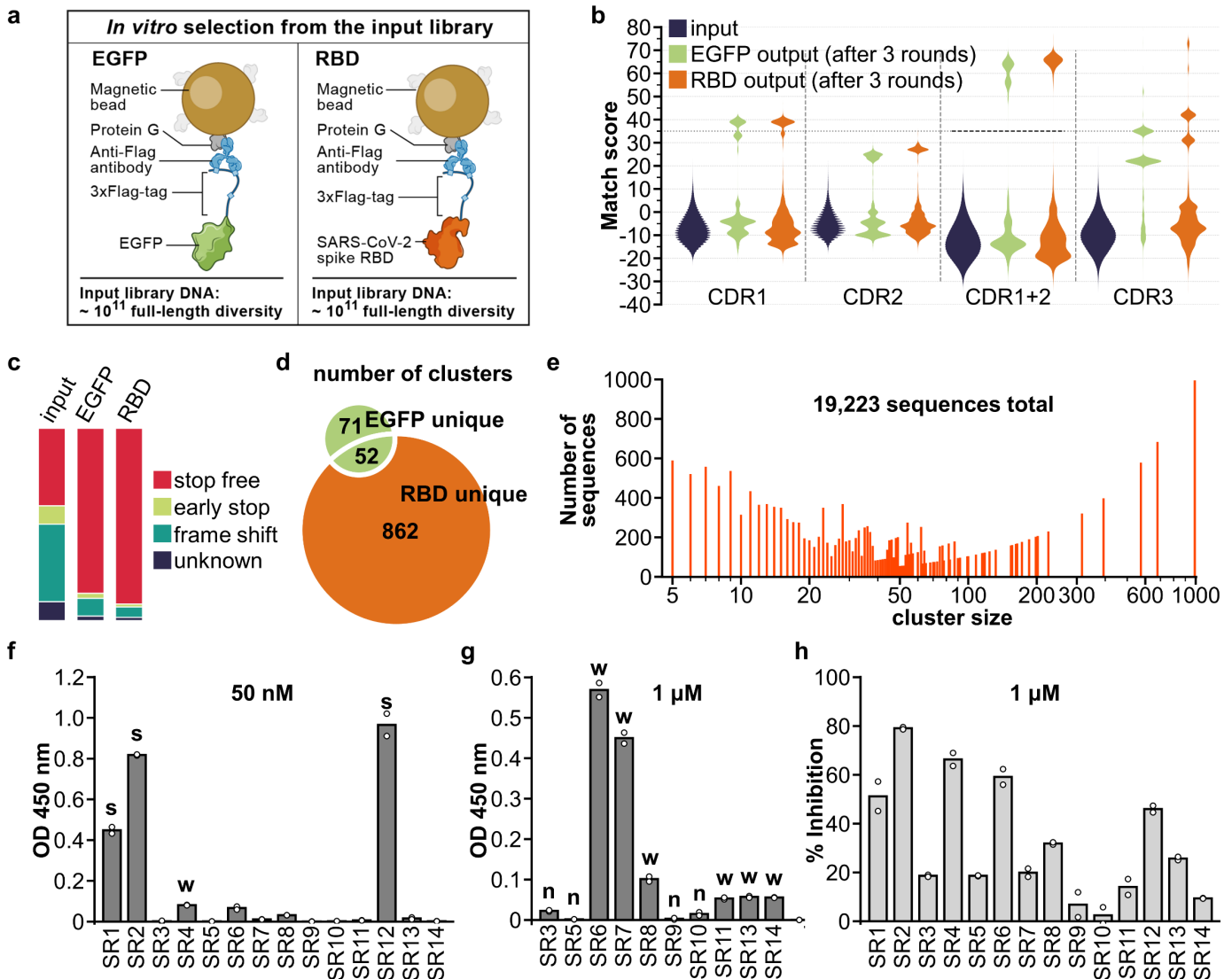
601 **Fig. 1. A cell-free antibody engineering platform for rapid isolation of antibodies from large**
 602 **synthetic libraries.** (a) The workflow takes linear DNA library as input. (b) Ribosome display
 603 links genotype (RNAs transcribed from DNA input library that are stop codon free, and stall
 604 ribosome at the end of the transcript) and phenotype (folded VHH protein tethered to ribosomes
 605 due to the lack of stop codon in the RNA). (c) Selection cycle that enriches DNA encoding for
 606 VHHs that binds immobilized targets. (d) High throughput sequencing of full-length VHHs. (e)
 607 Sequences are grouped into clusters based on similarity of their CDRs, each cluster is distinct and
 608 represent a unique binding family. (f) The system outputs one representative sequence from each
 609 cluster to be synthesized and characterized for specific downstream applications. (g) Workflow
 610 for generating VHH library. VHH CDR randomization was introduced by PCR using a hairpin

611 oligo (blocks DNA end from ligation) and an oligo with random 5' sequence, followed by
612 orientation-controlled ligation. Three successive PCR plus ligation cycles randomizes all three
613 CDRs. **(h)** The final DNA library sequence structure. **(i)** One round of ribosome display and anti-
614 Myc selection was performed after randomization of CDR1 and CDR2. The pie chart shows
615 percentage of indicated sequence categories before and after anti-Myc selection. **(j)** Length
616 distribution of DNA region encoding CDR1 of the VHH library before and after anti-Myc
617 selection. Arrows indicate all correct-frame lengths showing increased percentage after anti-Myc
618 selection.

619

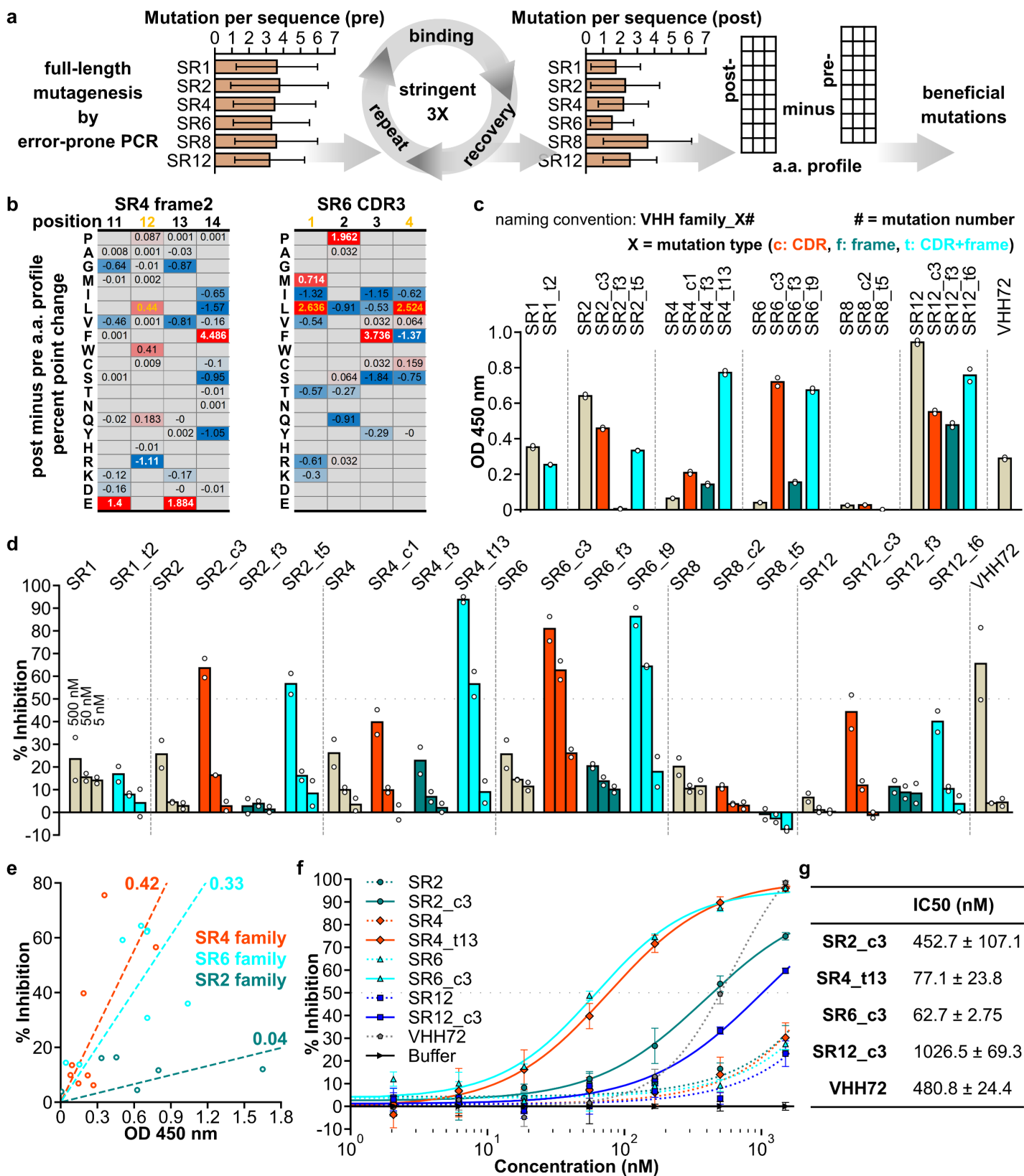
620

Fig 2



621 **Fig. 2. Isolation and characterization of synthetic VHhs that binds SARS-CoV-2 spike RBD.**
 622 (a) Immobilization strategy for the target proteins: 3xFlag-tagged EGFP or RBD. (b) Pair-wise
 623 CDR match score (based on BLOSUM62 matrix) were calculated for 2000 randomly selected
 624 sequences from input library and output libraries after 3 rounds of selection. High match score
 625 populations appeared in the output libraries. Combining CDR1 and 2 match scores further
 626 separated high and low score population and a match score of 35 (black dashed line) was chosen
 627 as cut-off for downstream clustering analysis. (c) Percentage of indicated sequence categories in
 628 the input library and output libraries (EGFP, RBD). (d) Number of unique and shared clusters
 629 identified in EGFP and RBD output libraries. (e) Number of sequences for each size of RBD
 630 unique clusters. (f) ELISA assay revealed 3 strong binders (“s”) to RBD, 7 weak binders (“w”)
 631 and (g) 4 non-binders (“n”) among the 14 VHhs chosen for characterization. (h) SARS-CoV-2 S
 632 pseudotyped lentivirus neutralization assay showed 6 VHhs inhibiting infection >30% at 1 μ M on
 633 HEK293T expressing ACE2 and TMPRSS2. Data shown are two technical replicates, bars indicate
 634 the average of data, circles indicate values of each replicate.

Fig 3



635 Fig. 3. An affinity maturation strategy enhances binding and neutralization properties of

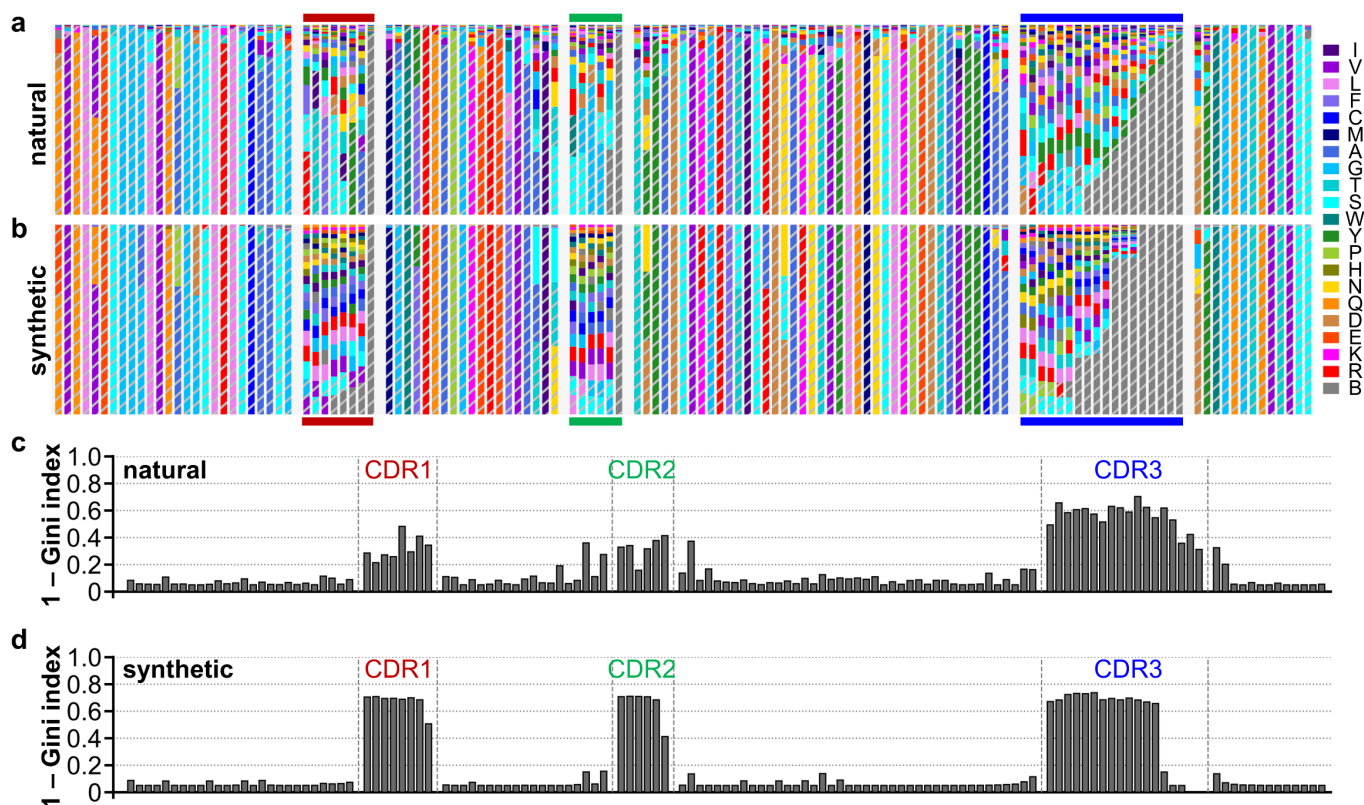
636 **synthetic VHHs.** (a) Affinity maturation workflow. (b) Two representative sections of position-
637 wise post- minus pre-affinity maturation amino acid percent point change profile. White values
638 indicate the original amino acid, yellow values indicate the beneficial mutation. Empty positions
639 indicate amino acids not detected in either the pre- or post- selection libraries. (c) ELISA assay of
640 VHH variants. (d) SARS-CoV-2 S pseudotyped lentivirus neutralization assay of VHHs on
641 HEK293T expressing ACE2 and TMPRSS2. For (c) and (d), data shown are two technical
642 replicates, bars indicate the average of data, circles indicate values of each replicate. (e) Scatter
643 plot of ELISA assay absorbance versus pseudotyped lentivirus neutralization as percent infection
644 inhibited. VHH concentration for both assays were 50 nM. Values are average of two technical
645 replicates. Numbers on linear fitting lines were r^2 value for data within each family. (f) Dose-
646 response curve for neutralization of pseudotyped lentiviral infection by VHHs. Markers are
647 average of three technical replicates, error bars are standard deviation. (g) IC50 calculated from
648 data in (f), presented as mean \pm standard deviation.

649

650

651

Fig S1



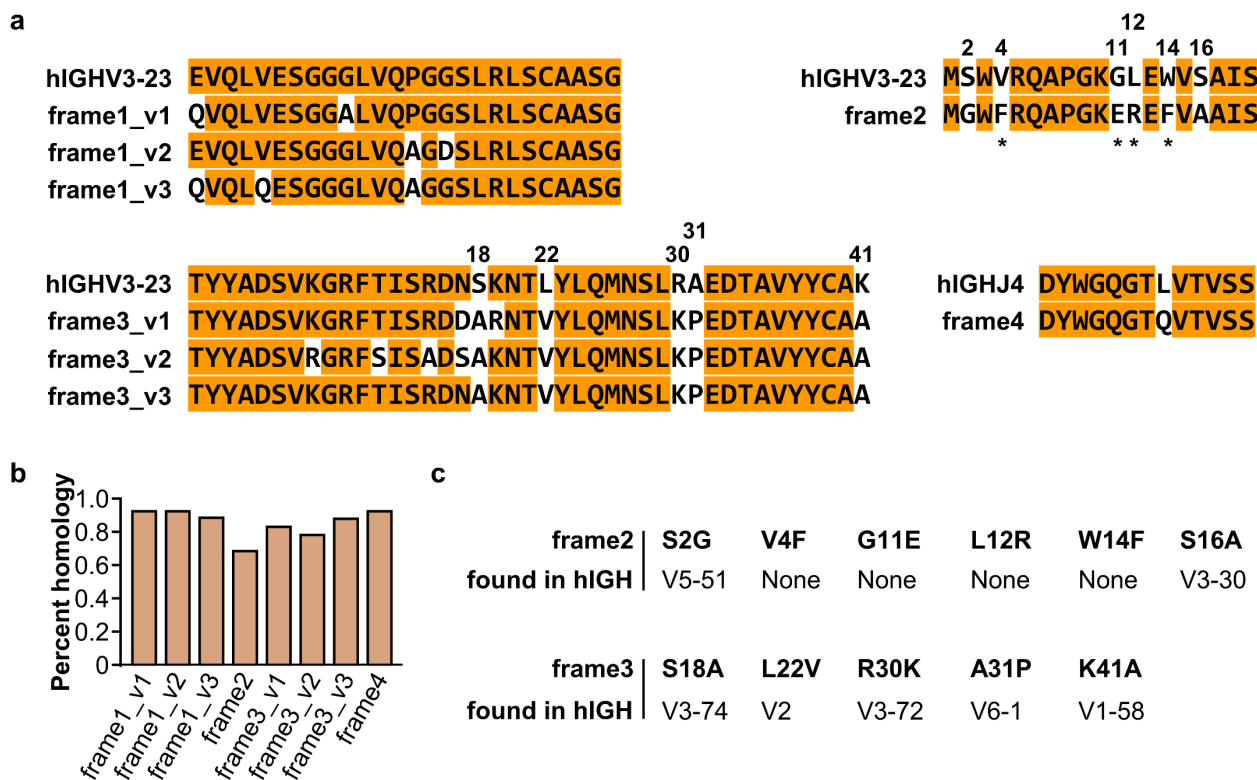
652 **Fig. S1. Amino acid profiles of natural and synthetic VHHs.** (a) Position-wise amino acid
653 profile of natural VHHs (298 VHHs, PDB) and (b) synthetic VHHs. Amino acids were color
654 coded according to labels to the right, B indicates an empty position. Bar height is the relative
655 percentage of each amino acids. The two most common amino acids were shown as patterned
656 bars while others were shown as solid bars. (c) Plot of diversity index (as 1 - Gini index) for
657 each amino acid position of natural VHHs and (d) synthetic VHHs.

658

659

660

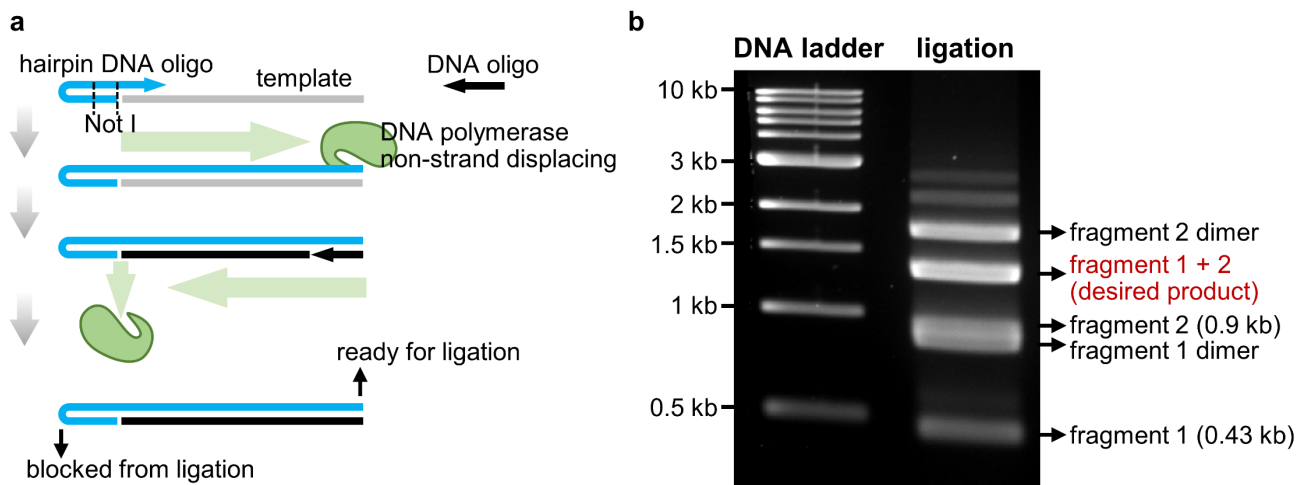
Fig S2



661 **Fig. S2. Design of VHH frames and their homology to human IGH genes.** (a) Amino acid
 662 sequences encoded by frames that serve as templates for VHH library generation were aligned to
 663 the corresponding segments of the human IGHV3-23 (hIGHV3-23) or IGHJ4 (hIGHJ4). Positions
 664 in hIGHV3-23/hIGHJ4 that are identical to the corresponding position in at least one VHH frames
 665 are highlighted in orange. Positions in VHH frames that are identical to the corresponding position
 666 in hIGHV3-23/hIGHJ4 are highlighted in orange. hIGHV3-23 positions not identical to any VHH
 667 frames are numbered according to its position within the segment. Asterisks indicate VHH
 668 hallmark residues thought to be required for VHH's independence of light chain. (b) Percent
 669 homology of VHH frames to the closest human gene. (c) List of VHH residues at positions
 670 numbered in (a) and representative human IGHV genes that encode the same VHH residue at the
 671 corresponding position. None: no human IGHV genes has the VHH residue at the corresponding
 672 position.

673
 674
 675
 676

Fig S3



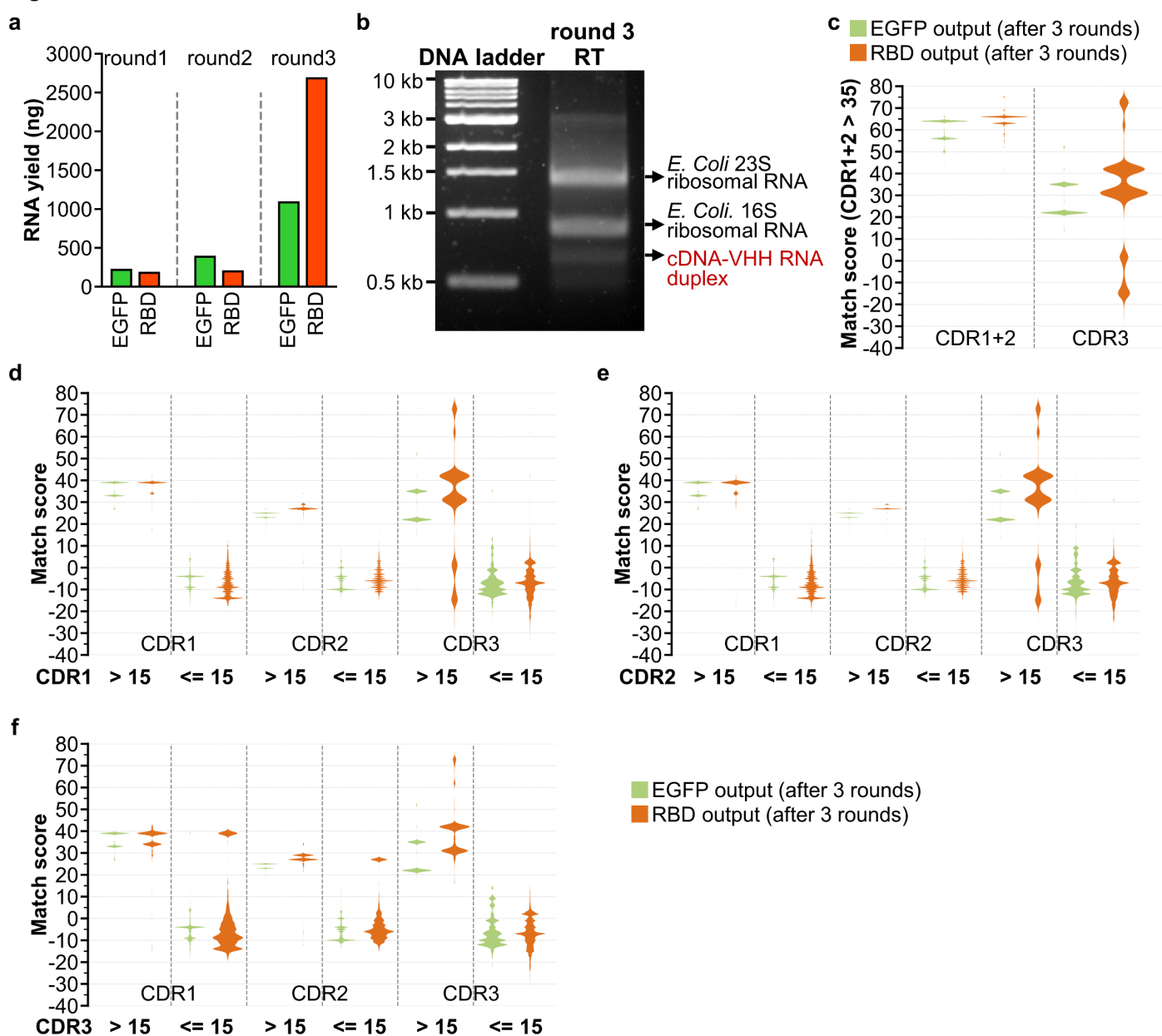
677 **Fig. S3. Working principles for orientation-controlled ligation by end blocking using**
678 **hairpin oligos.** (a) working principle for generating one end blocked DNA for orientation-
679 controlled ligation by PCR using a hairpin DNA oligo. (b) Representative orientation-controlled
680 ligation products visualized by agarose gel electrophoresis.

681

682

683

Fig S4

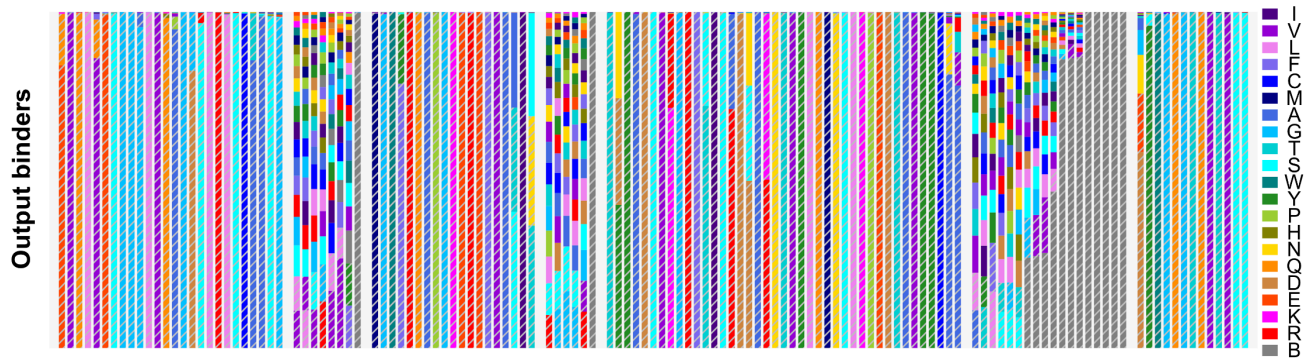


684 **Fig. S4. Evaluation of ribosome display and selection rounds.** (a) Yield of recovered RNA at
 685 each round of ribosome display and selection for EGFP or RBD targets. (b) Representative RT
 686 reaction (without heat denaturation) product for RBD selection after 3 rounds, visualized by
 687 agarose gel electrophoresis. (c) Plot of match scores of sequence pairs with a combined CDR1
 688 and CDR2 score > 35. (d) Plot of match scores of sequence pairs (from 2000 randomly sampled
 689 sequences) with indicated CDR1 scores, and (e) indicated CDR2 scores, and (f) indicated CDR3
 690 scores.

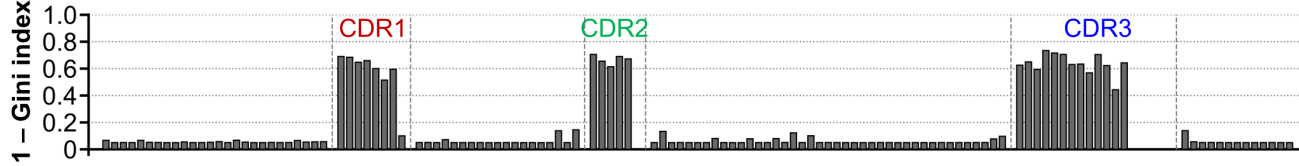
691

Fig S5

a



b



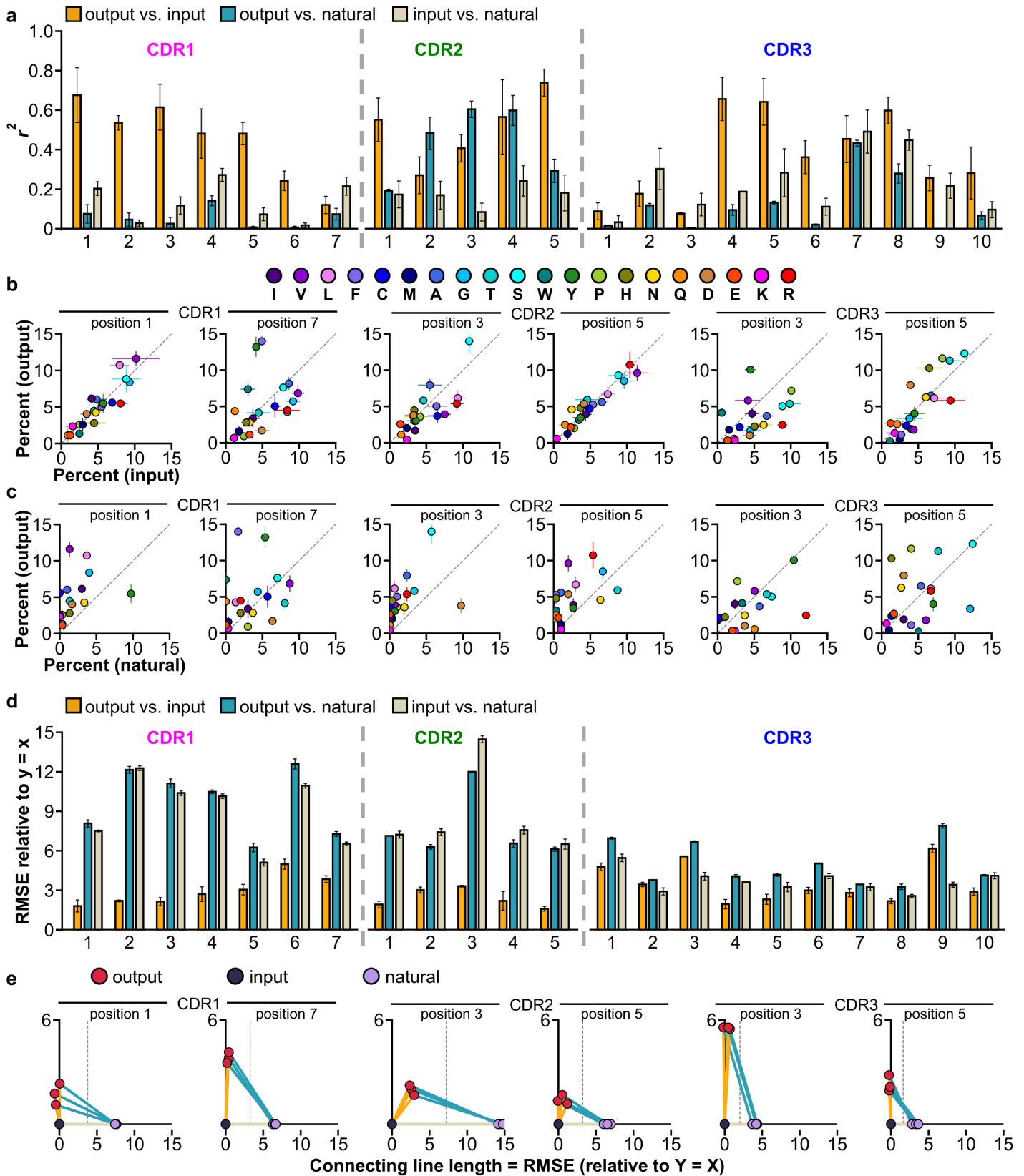
692 **Fig. S5. Amino acid profile for EGFP and RBD unique output binders.** (a) Amino acid
693 profile of representative VHH sequence for each unique cluster identified from RBD and EGFP
694 output libraries (“output binders”, 932 sequences). Plotted as described in **Fig S1a**. (b) Plot of
695 diversity index (as 1 – Gini index) for each amino acid position of output binder VHHs.

696

697

698

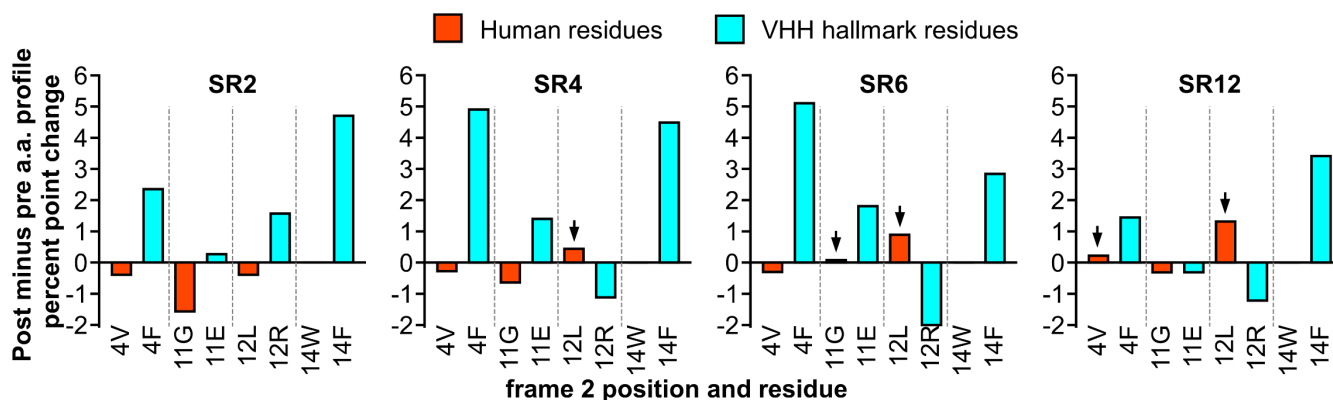
Fig S6



699 Fig. S6. Unique output binders amino acid profiles correlate more highly with that of input

700 **library than natural VHHs.** (a) r^2 values for the amino acid percentages in the indicated
701 sequence group pairs at each CDR position. 298 natural VHHs (natural) and 298 randomly
702 sampled sequences from input library (input) and output binders (output) were analyzed. Three
703 random sampling trials were performed to generate three r^2 for each position. (b) Scatter plots of
704 the percentage of each amino acid in the input library and the output binders and (c) that in the
705 natural VHHs and the output binders at representative CDR positions. Circles are the mean and
706 error bars are the standard deviation of data. (d) Root mean square error (RMSE, relative to $Y =$
707 X line) values for the indicated sequence group pairs at each CDR position. Using the same
708 randomly sampled sequences as (a). (e) Plot showing the similarity distances between the three
709 sequence groups, with each connecting line length between two sequence groups indicating their
710 RMSE. Vertical dashed lines indicate the middle point of the distance between output and
711 natural sequence groups
712

Fig S7



713 **Fig. S7. Affinity maturation leads to some VHH hallmark residues converting to the**
 714 **corresponding human VH residues.** The post- minus pre-affinity maturation percent point
 715 change of VHH hallmark residues and the corresponding human residues for each VHH. Arrows
 716 indicate human residues with increased frequency as a result of affinity maturation.

717

718

719 **Table S1. Natural VHH sequences selected for calculating natural VHH amino acid profile.**
 720 Amino acid sequences of all VHHs from Protein Data Bank (sheet: all_VHH_RSCB) and from
 721 which 298 unique VHHs were selected to represent natural VHHs (sheet: unique_VHH_RSCB),
 722 the sequences were separated into 4 frames and 3 CDRs.

723

724 **Table S2. Amino acid profile of natural VHHs and synthetic VHHs in the VHH input library.**
 725 Position-wise amino acid profile of natural VHHs and VHHs in the input library. Positions are
 726 relative positions within each segment and numbers are percentage of the corresponding amino
 727 acid labelled to the left of each segment.

728

729 **Table S3. Primers and templates used for generation, selection and sequencing of VHH**
 730 **library.** Primer sequences used in this study and VHH frame template sequences. PCR cycling
 731 conditions were also shown.

732

733 **Table S4. List of RBD binder clusters.** A list containing key information for all predicted RBD
 734 binding clusters (sheet: all clusters) and unique RBD binding clusters (sheet: all CDR unique).
 735 Cluster ID, size, CDR representative sequences, CDR consensus sequences, CDR scores
 736 (**Materials and Methods**), and whether each CDR is unique to RBD and not found in EGFP
 737 clusters were shown.

738

739 **Table S5. List of EGFP binder clusters.** A list containing key information for all predicted
 740 EGFP binding clusters (sheet: all clusters) and unique EGFP binding clusters (sheet: all CDR
 741 unique). Cluster ID, size, CDR representative sequences, CDR consensus sequences, CDR scores
 742 (**Materials and Methods**), and whether each CDR is unique to EGFP and not found in RBD
 743 clusters were shown. The cluster with ID 0, is a spike-in VHH¹⁵, and did not originate from the
 744 input library.

745
746 **Table S6. Affinity maturation subtracted amino acid profile for SR4 and SR6.** Position-wise
747 post- minus pre- affinity maturation amino acid profile for SR4 and SR6. Numbers are percent
748 point change of each amino acid after affinity maturation.
749
750 **Table S7. Amino acid sequences of VHH variant and the mutations they contain.** Amino acid
751 sequences of all VHH variants characterized in this study.
752
753 **Table S8. VHH variants ELISA and neutralization data.** ELISA binding assay and
754 pseudotyped virus neutralization assay results for all VHH variants characterized in this study.
755
756 **Table S9. High-throughput sequencing and analysis metadata.** Number of sequences obtained
757 by high-throughput sequencing for indicated analyses.
758
759 **Data S1. Cluster files for SR1, SR2, SR4, SR6, SR8, SR12.** Text file containing all sequences
760 belonging to each cluster. Each line in the file represent one sequence, both segments and full
761 length of the sequence were shown, shown items were divided by “#” and in the order from start
762 to end of each line was: CDR1 amino acid sequence, CDR2 amino acid sequence, CDR3 amino
763 acid sequence, full-length amino acid sequence, full-length DNA sequence.
764