Supporting Information

Prediction of single-mutation effects for fluorescent immunosensor engineering with an end-to-end trained protein language model

Akihito Inoue,[a#] Bo Zhu,[b#] Keisuke Mizutani,[c] Ken Kobayashi,[c] Takanobu Yasuda,[b] Alon Wellner,[d] Chang C. Liu[d] and Tetsuya Kitaguchi[b]*

[a] Graduate School of Life Science and Technology, Institute of Science Tokyo, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan

[b] Laboratory for Chemistry and Life Science, Institute of Integrated Research, Institute of Science Tokyo, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan

[c] School of Engineering, Institute of Science Tokyo, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

[d] Department of Biomedical Engineering, University of California, Irvine, CA, 92697, USA

[#] Equal contributions

*** Corresponding Author**

Tetsuya Kitaguchi, Ph. D.
Laboratory for Chemistry and Life Science, Institute of Integrated Research, Institute of Science Tokyo, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan
E-mail: kitaguc.t.aa@m.titech.ac.jp

Table of Contents

Table S1. Characterization of Q-bodies recognizing RBD

	Max response (-fold)	EC_{50} (nM)	LOD (nM)
RBD1i13	1.1	6.6	5.7
L113W	1.2	8.2	4.8
H115W	1.7	26	0.45
RBD10i14	1.4	1.7	0.37
E112R	1.9	5.4	0.20

Table S2. Primers used for library construction / NGS / sub-cloning

Library construction

Primer name	Nucleotide sequence (5'-3')
Inf_AgeI_Krese_back	GAAGGGAGGCACCGGTCAGGTGCAGCTGCAGGAA
Inf_BamHI_Kruse_for	CCTTGTAGTCGGATCCGCTCGAGACGGTCACCTGGGTGC
Adapter_E4_for	ACCGGTGCCTCCCTTCTC
Adapter_flag_back	GGATCCGACTACAAGGACGATG
pYD1_back_long	AGTAACGTTTGTCAGTAATTGCGGTTC
pYD1_for_long	GTCGATTTTGTTACATCTACACTGTTG

NGS

Primer name	Nucleotide sequence (5'-3')
pYD1_initial_HA_back	**AAATGACTGAAA**AGTAACGTTTGTCAGTAATTGCG
pYD1_HA_quench_back	**AAAGCTTCTAAA**AGTAACGTTTGTCAGTAATTGCG
pYD1_HA_noquench_back	**AAAGCTCTGAAA**AGTAACGTTTGTCAGTAATTGCG
pYD1_round2_for	**AAACTGATCAAA**GTCGATTTTGTTACATCTACAC

* Bold: Barcode sequence

Sub-cloning

Primer name	Nucleotide sequence (5'-3')
Inf_E4_AgeI_back	CTTGAGAAGGGAGGCACCG
Inf_Kruse_XhoI_for	GGTGGTGGTGCTCGAGGCTGCTCACGGTCACCTG

Table S3. Sequences used in NGS data processing

Dataset name	Sequences for finding reverse complementary reads using Manipulate FASTAQ (5'-3')	Barcode sequences for grouping the reads using Cutadapt (5'-3')
Initial E4-tagged library	TCAGTCATT	AATGACTGA
After round 1 (HQ1)	TAGAAGCTT	AAGCTTCTA
After round 1 (LQ1)	TCAGAGCTT	AAGCTCTGA
After round 2 (HQ2)	AAACTGATCAAA	AAGCTTCTA / TTTGATCAGTTT
After round 2 (LQ2)	AAACTGATCAAA	AAGCTCTGA / TTTGATCAGTTT
CDR1	-	AGCTGCGCG / TATCGCCAG
CDR2	-	CCGGGCAAA / AGCGTGAAA
CDR3	-	GCGGTGTAT / CAGGGCACC

Table S4. The number of total valid sequences in NGS

	E4-tagged library	After round 1 (HQ1)	After round 1 (LQ1)	After round 2 (HQ2)	After round 2 (LQ2)
CDR1	39,336	49,631	49,809	87,967	83,189
CDR2	40,726	43,268	51,357	89,465	87,397
CDR3	16,429	33,748	34,080	56,649	50,809
CDR1+3	-	13,362	13,511	18,934	18,999

HQ: Selection for high TAMRA quenching properties,

LQ: Selection for low TAMRA quenching properties.

Table S5. The number of sequences for the training dataset for NanoQ-model 1.0

	HQ2	LQ2
CDR1	5186	4148
CDR2	2903	4084
CDR3	4449	3912
CDR1+3	2492	2163

Table S6. Nucleotide/amino acid sequences of nanobodies for Q-bodies

Name	Sequence
pYD1-E4-RBD1i13 (Yeast surface display)	ATGCAGTTACTTCGCTGTTTTTCAATATTTTCTGTTATTGCTTCAGTTTTAGCACAGG AACTGACAACTATATGCGAGCAAATCCCCTCACCAACTTTAGAATCGACGCCGTAC TCTTTGTCAACGACTACTATTTTGGCCAACGGGAAGGCAATGCAAGGAGTTTTTGA ATATTACAAATCAGTAACGTTTGTCAGTAATTGCGGTTCTCACCCCTCAACAACTAG CAAAGGCAGCCCCATAAACACACAGTATGTTTTTAAGGACAATAGCTCGACGATTG AAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTGGTGGTGGT GGTTCTGGTGGTGGTGGTTCTGGTGGTGGTGGTTCTGCTAGCATGGCTGAAATCGC TGCACTTGAAAAGGAAATTGCAGCCCTAGAAAAGGAAATAGCAGCGCTGGAAAA GGAAATCGCAGCACTTGAGAAGGGAGGCACCGGTCAGGTGCAGCTGCAGGAAAG CGGCGGCGGCCTGGTGCAGGCGGGCGGCAGCCTGCGCCTGAGCTGCGCGGCGAG CGGCACTATTTCTTACGAAAACTTCATGGGCTGGTATCGCCAGGCGCCGGGCAAAG GACGCAAACTTGTTGCCGGTATTAATGACGGTACTAATACCTATTATGCGGATAGCG TGAAAGGCCGCTTTACCATTAGCCGCGATAACGCGAAAAACACCGTGTATCTGCAG ATGAACAGCCTGGAACCGGAAGATACCGCGGTGTATTATTGCGCGGTTATCGGTGC TTCTGTTCTGGGTCATGCTTATTGGGGCCAGGGCACCCAGGTGACCGTGAGCAGC GATCCGACTACAAGGACGATGACGACAAGTAA MQLLRCFSIFSVIASVLAQELTTICEQIPSPTLESTPYSLSTTTILANGKAMQGVFEYYK SVTFVSNCGSHPSTTSKGSPINTQYVFKDNSSTIEGRYPYDVPDYALQASGGGGSGGG GSGGGGSASMAEIAALEKEIAALEKEIAALEKEIAALEKGGTGQVQLQESGGGLVQA GGSLRLSCAASGTISYENFMGWYRQAPGKGRKLVAGINDGTNTYYADSVKGRFTISR DNAKNTVYLQMNSLEPEDTAVYYCAVIGASVLGHAYWGQGTQVTVSSGSDYKDDD DK

pYD1-E4-RBD10i14 (Yeast surface display)	ATGCAGTTACTTCGCTGTTTTTCAATATTTTCTGTTATTGCTTCAGTTTTAGCACAGG AACTGACAACTATATGCGAGCAAATCCCCTCACCAACTTTAGAATCGACGCCGTAC TCTTTGTCAACGACTACTATTTTGGCCAACGGGAAGGCAATGCAAGGAGTTTTTGA ATATTACAAATCAGTAACGTTTGTCAGTAATTGCGGTTCTCACCCCTCAACAACTAG CAAAGGCAGCCCCATAAACACACAGTATGTTTTTAAGGACAATAGCTCGACGATTG AAGGTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTGGTGGTGGT GGTTCTGGTGGTGGTGGTTCTGGTGGTGGTGGTTCTGCTAGCATGGCTGAAATCGC TGCACTTGAAAAGGAAATTGCAGCCCTAGAAAAGGAAATAGCAGCGCTGGAAAA GGAAATCGCAGCACTTGAGAAGGGAGGCACCGGTCAGGTGCAGCTGCAGGAAAG CGGCGGCGGCCTGGTGCAGGCGGGCGGCAGCCTGCGCCTGAGCTGCGCGGCGAG CGGCACTATTTTTCAGGTTGGTTCTGTGGGCTGGTATCGCCAGGCGCCGGGCAAAG GACGCAAATTTGTTGCCACTATTGCTGACGGTAGTAGTACCAATTATGCGGGTAGC GTGAAAGGCCGCTTTACCATTAGCCGCGATAACGCGAAAAACACCGTGTATCTGCA GATGAACAGCCTGAAACCGGAAGATACCGCGGTGTATTATTGCGCGGCTCTGGGTC AGGTTTCTGAATACAACTCTGCTTCTTACGAATGGACTTATCCGTATTGGGGCCAGG GCACCCAGGTGACCGTGAGCAGCGATCCGACTACAAGGACGATGACGACAAGTA A MQLLRCFSIFSVIASVLAQELTTICEQIPSPTLESTPYSLSTTTILANGKAMQGVFEYYK SVTFVSNCGSHPSTTSKGSPINTQYVFKDNSSTIEGRYPYDVPDYALQASGGGGSGGG GSGGGGSASMAEIAALEKEIAALEKEIAALEKEIAALEKGGTGQVQLQESGGGLVQA GGSLRLSCAASGTIFQVGSVGWYRQAPGKGRKFVATIADGSSTNYAGSVKGRFTISRD NAKNTVYLQMNSLKPEDTAVYYCAALGQVSEYNSASYEWTYPYWGQGTQVTVSSG SDYKDDDDK

pSQ-E4-RBD1i13 (Bacterial expression)	ATGGCTGAAATCGCTGCACTTGAAAAGGAAATTGCAGCCCTAGAAAAGGAAATA GCAGCGCTGGAAAAGGAAATCGCAGCACTTGAGAAGGGAGGCACCGGTCAGGTG CAGCTGCAGGAAAGCGGCGGCGGCCTGGTGCAGGCGGGCGGCAGCCTGCGCCTG AGCTGCGCGGCGAGCGGCACTATTTCTTACGAAAACTTCATGGGCTGGTATCGCC AGGCGCCGGGCAAAGGACGCAAACTTGTTGCCGGTATTAATGACGGTACTAATA CCTATTATGCGGATAGCGTGAAAGGCCGCTTTACCATTAGCCGCGATAACGCGAA AAACACCGTGTATCTGCAGATGAACAGCCTGGAACCGGAAGATACCGCGGTGTA TTATTGCGCGGTTATCGGTGCTTCTGTTCTGGGTCATGCTTATTGGGGCCAGGGCA CCCAGGTGACCGTGAGCAGCTCGAGCACCACCACCACCACCACGGATCCGACTAC AAGGACGACGATGACAAATAA MAEIAALEKEIAALEKEIAALEKEIAALEKGGTGQVQLQESGGGLVQAGGSLRLSCA ASGTISYENFMGWYRQAPGKGRKLVAGINDGTNTYYADSVKGRFTISRDNAKNTVY LQMNSLEPEDTAVYYCAVIGASVLGHAYWGQGTQVTVSSLEHHHHHHGSDYKDDD DK

pSQ-E4-RBD10i14 (Bacterial expression)	ATGGCTGAAATCGCTGCACTTGAAAAGGAAATTGCAGCCCTAGAAAAGGAAATA GCAGCGCTGGAAAAGGAAATCGCAGCACTTGAGAAGGGAGGCACCGGTCAGGTG CAGCTGCAGGAAAGCGGCGGCGGCCTGGTGCAGGCGGGCGGCAGCCTGCGCCTG AGCTGCGCGGCGAGCGGCACTATTTTTCAGGTTGGTTCTGTGGGCTGGTATCGCC AGGCGCCGGGCAAAGGACGCAAATTTGTTGCCACTATTGCTGACGGTAGTAGTAC CAATTATGCGGGTAGCGTGAAAGGCCGCTTTACCATTAGCCGCGATAACGCGAA AAACACCGTGTATCTGCAGATGAACAGCCTGAAACCGGAAGATACCGCGGTGTA TTATTGCGCGGCTCTGGGTCAGGTTTCTGAATACAACTCTGCTTCTTACGAATGGA CTTATCCGTATTGGGGCCAGGGCACCCAGGTGACCGTGAGCAGCTCGAGCACCAC CACCACCACCACGGATCCGACTACAAGGACGACGATGACAAATAA MAEIAALEKEIAALEKEIAALEKEIAALEKGGTGQVQLQESGGGLVQAGGSLRLSCA ASGTIFQVGSVGWYRQAPGKGRKFVATIADGSSTNYAGSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCAALGQVSEYNSASYEWTYPYWGQGTQVTVSSLEHHHHHHG SDYKDDDDK

*Aga2 signal sequence, Aga2 protein, affinity tag (HA, His, FLAG), E4 tag, nanobody

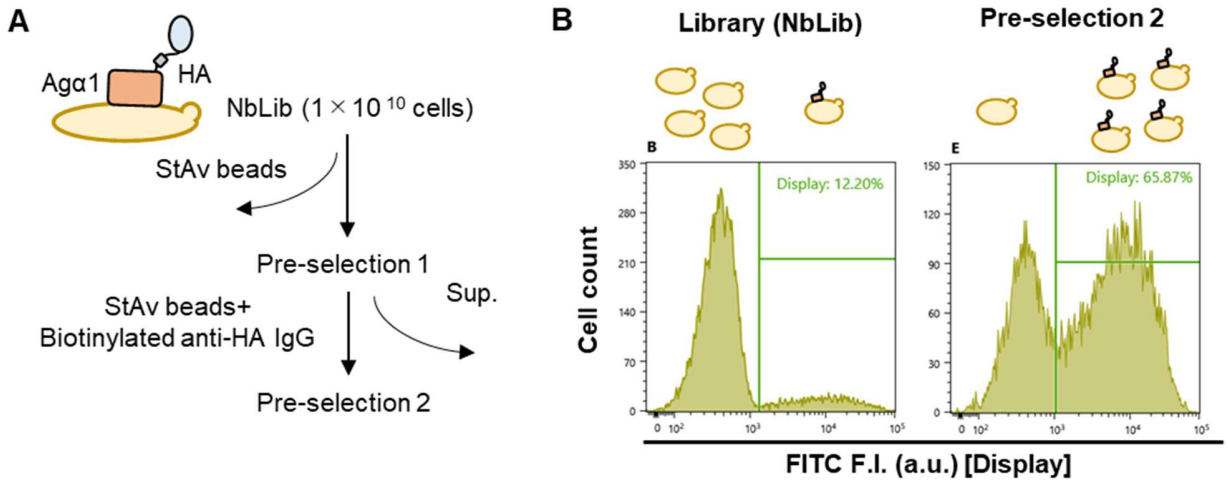

Figure S1. Pre-selection for the enrichment of the yeasts displaying nanobodies from NbLib. (A) Schematic image of magnetic sorting for pre-selection. The yeasts non-specifically binding with StAv beads were removed (Pre-selection 1), followed by the collection of the yeasts displaying the nanobodies (Pre-selection 2). (B) Flow cytometric analysis of yeasts before and after pre-selection.

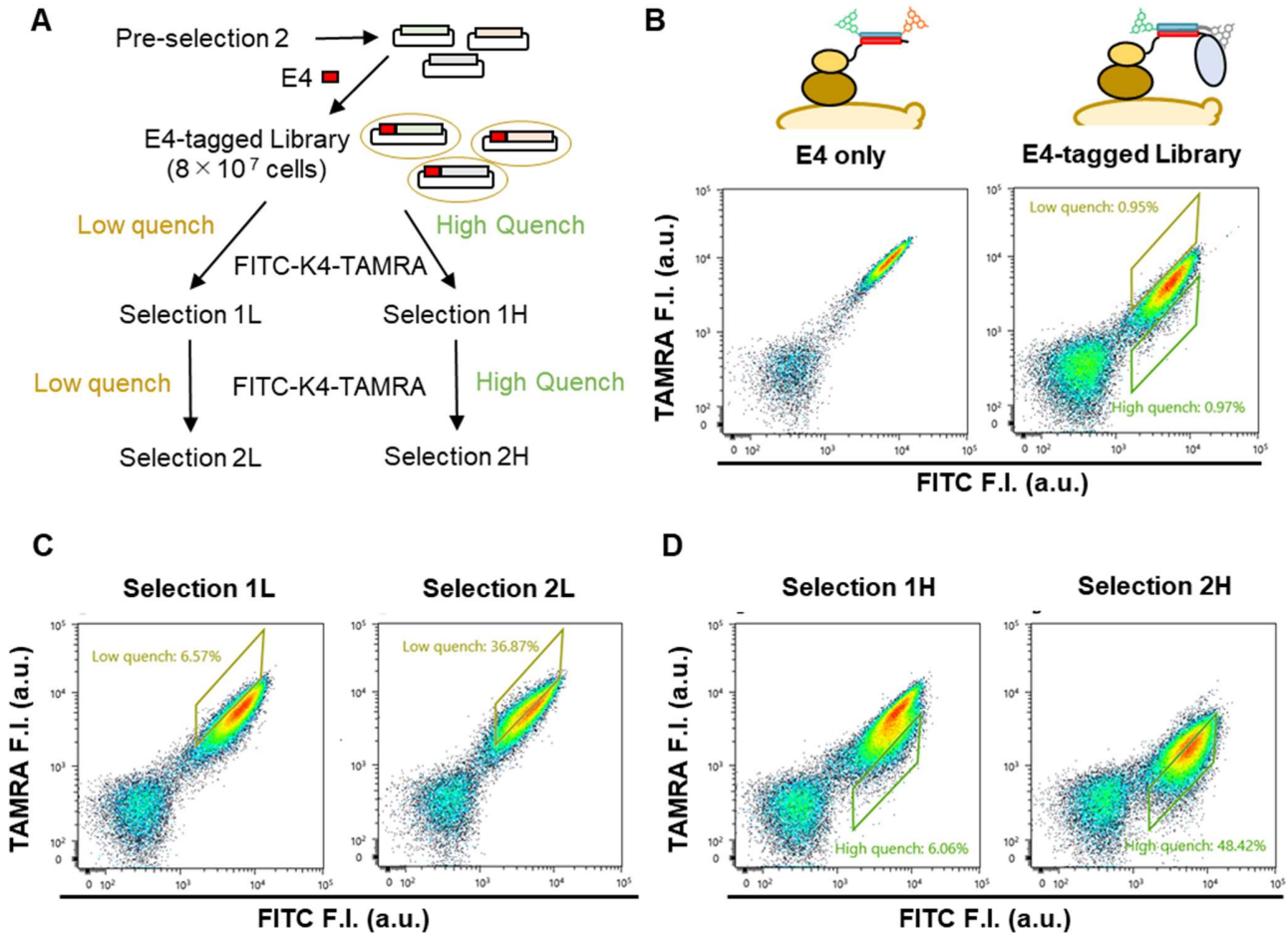

Figure S2. Collection of the yeasts displaying high-quenching and low-quenching nanobodies. (A) Schematic image of the collection step. After performing pre-selection as shown in Fig. S1, the plasmids were extracted from the collected yeast, and fused with E4 to construct the E4-tagged library. (B-D) Flow cytometric analysis of the yeast displaying E4-peptide or E4-tagged library during collection. Low quench or High quench was selected using the yellow gate or the green gate respectively.

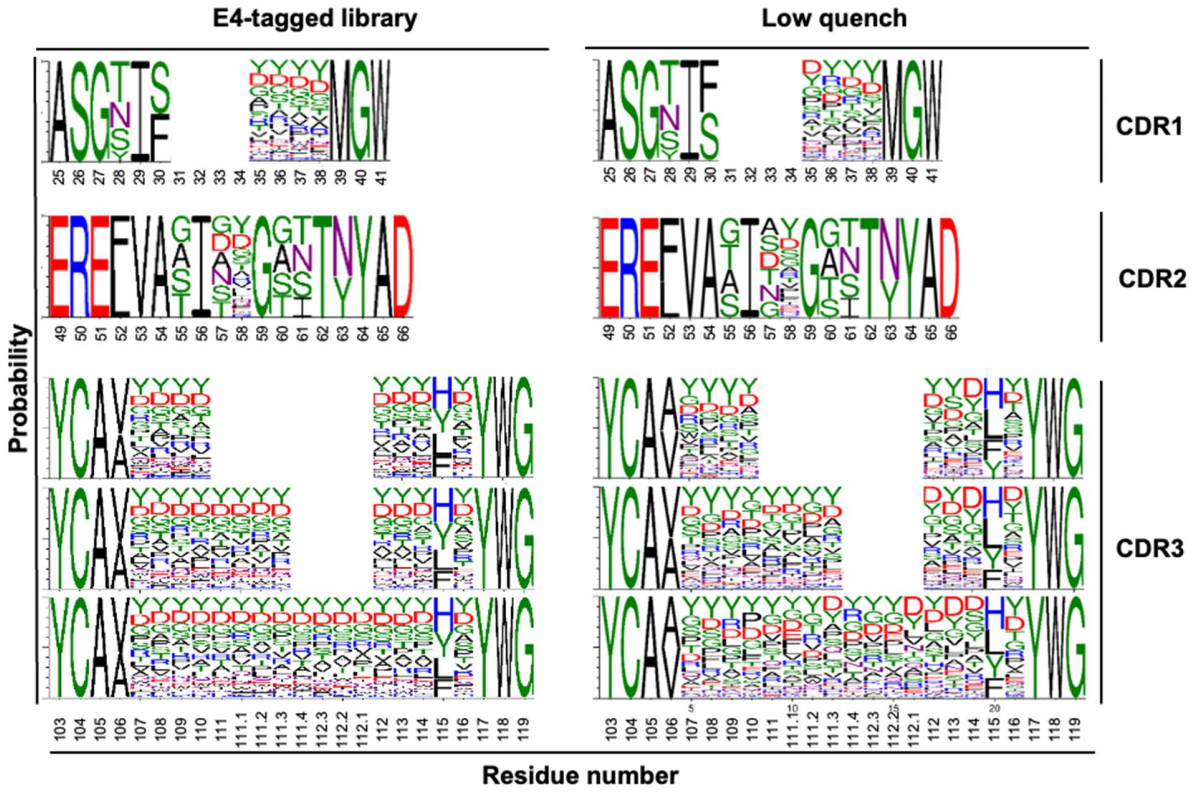

Figure S3. Additional information for NGS analysis. Weblogo of E4-tagged library and low quench sequences (LQ2).

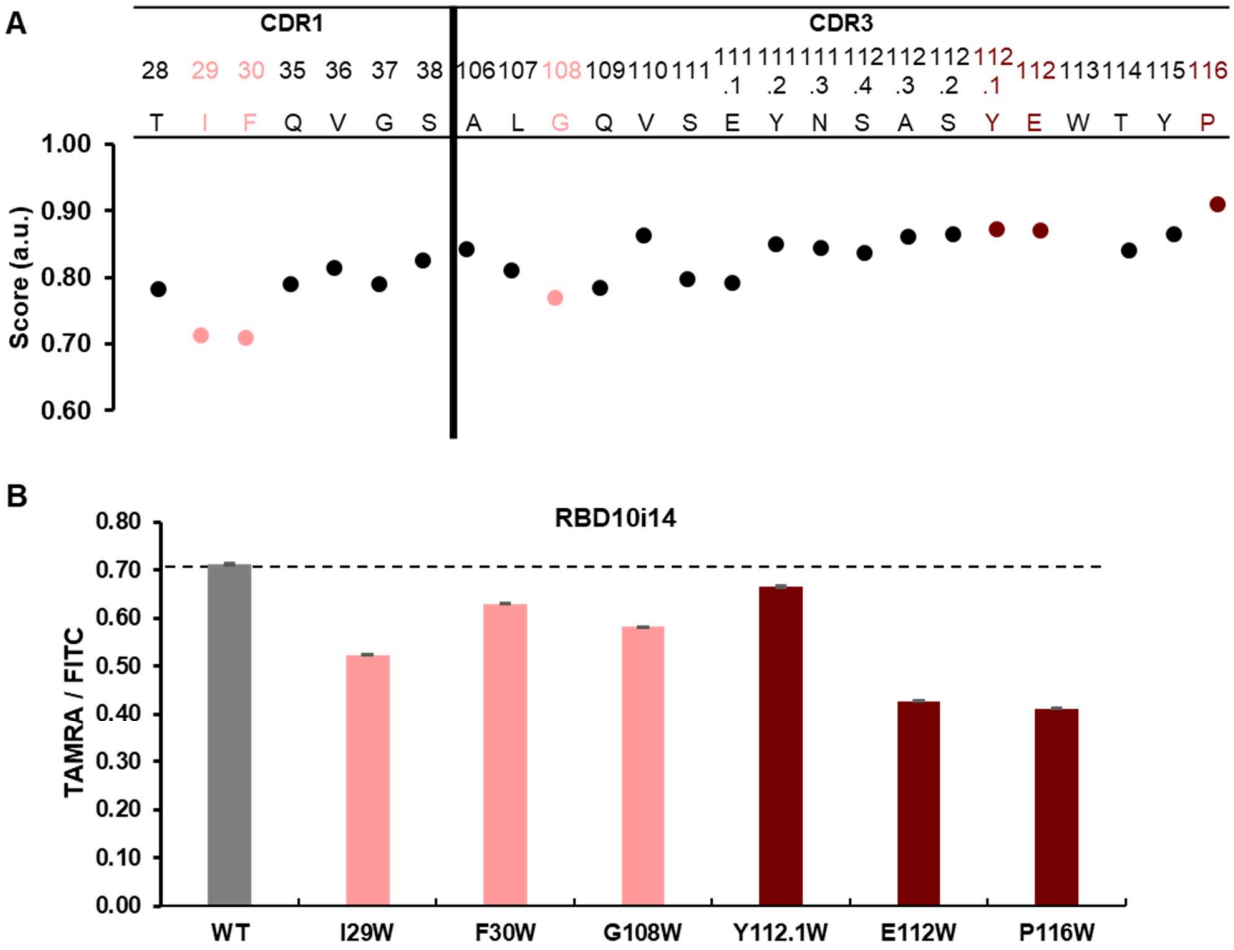

Figure S4. In silico Trp scanning on RBD10i14 and its validation on yeast cell surface. (A) The probability score during in silico Trp scanning on RBD10i14. (B) TAMRA/FITC ratio of mutants on yeast cell surface selected during in silico Trp scanning. The bar graphs represents the mean of TAMRA/FITC ratio ± standard error of mean. The 3 highest scores were dark red, and 3 lowest scores were pink.

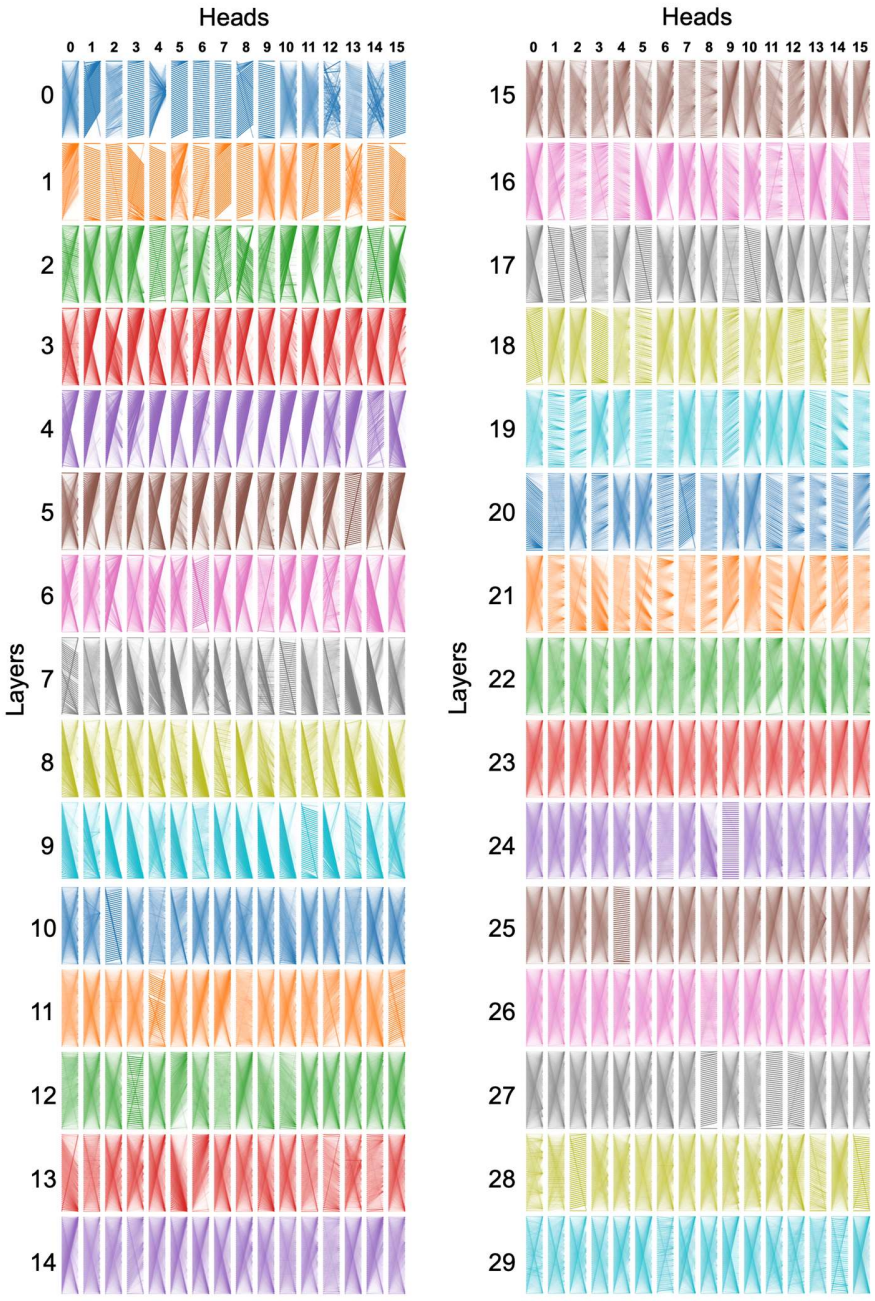

Figure S5. Attention visualization on the CDR sequences of RBD10i14-E112R nanobody of the pre-trained protein language model ProtBert-BFD. Model view of the self-attention in each head of each layer.

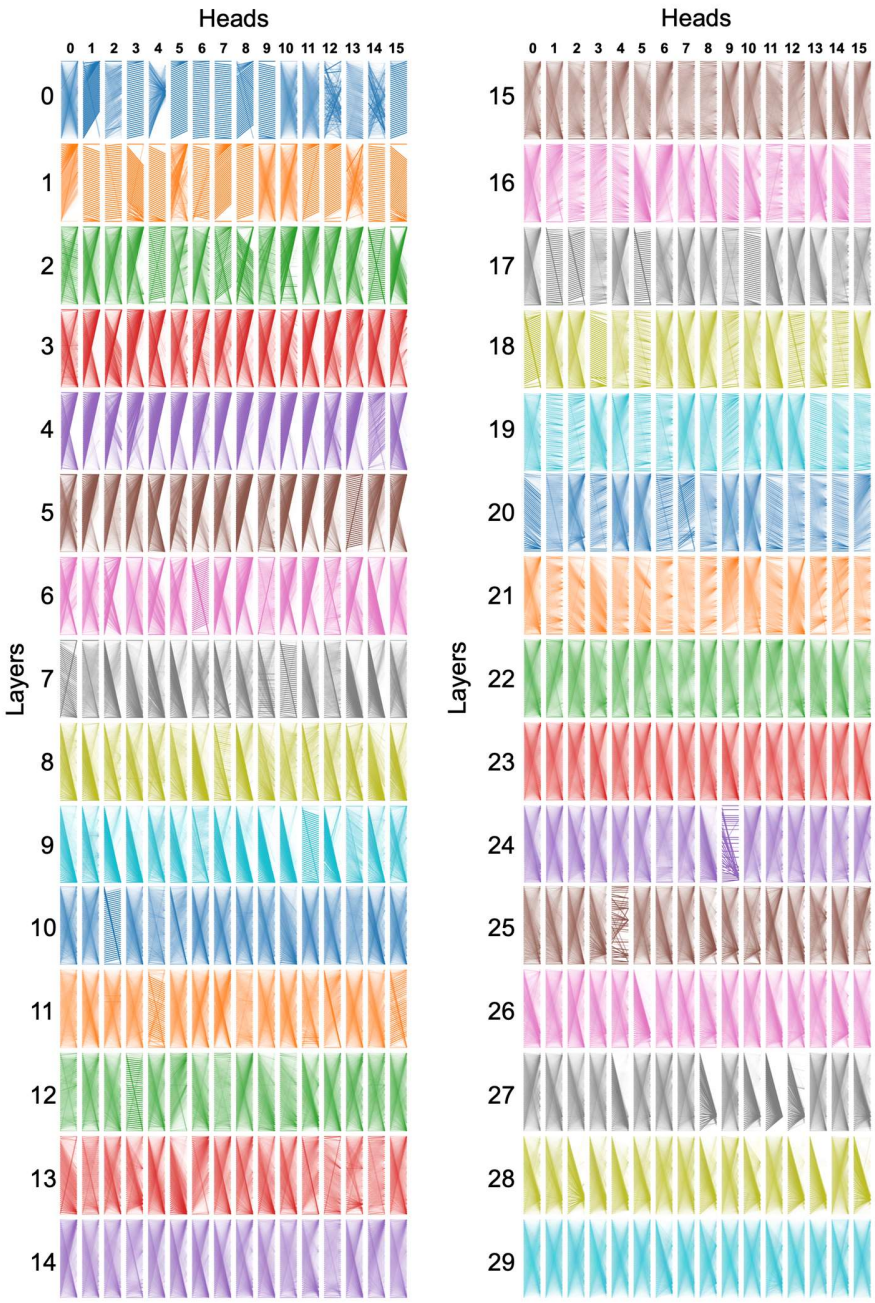

Figure S6. Attention visualization on the CDR sequences of RBD10i14-E112R nanobody of the NanoQ-model 1.0. Model view of the self-attention in each head of each layer.

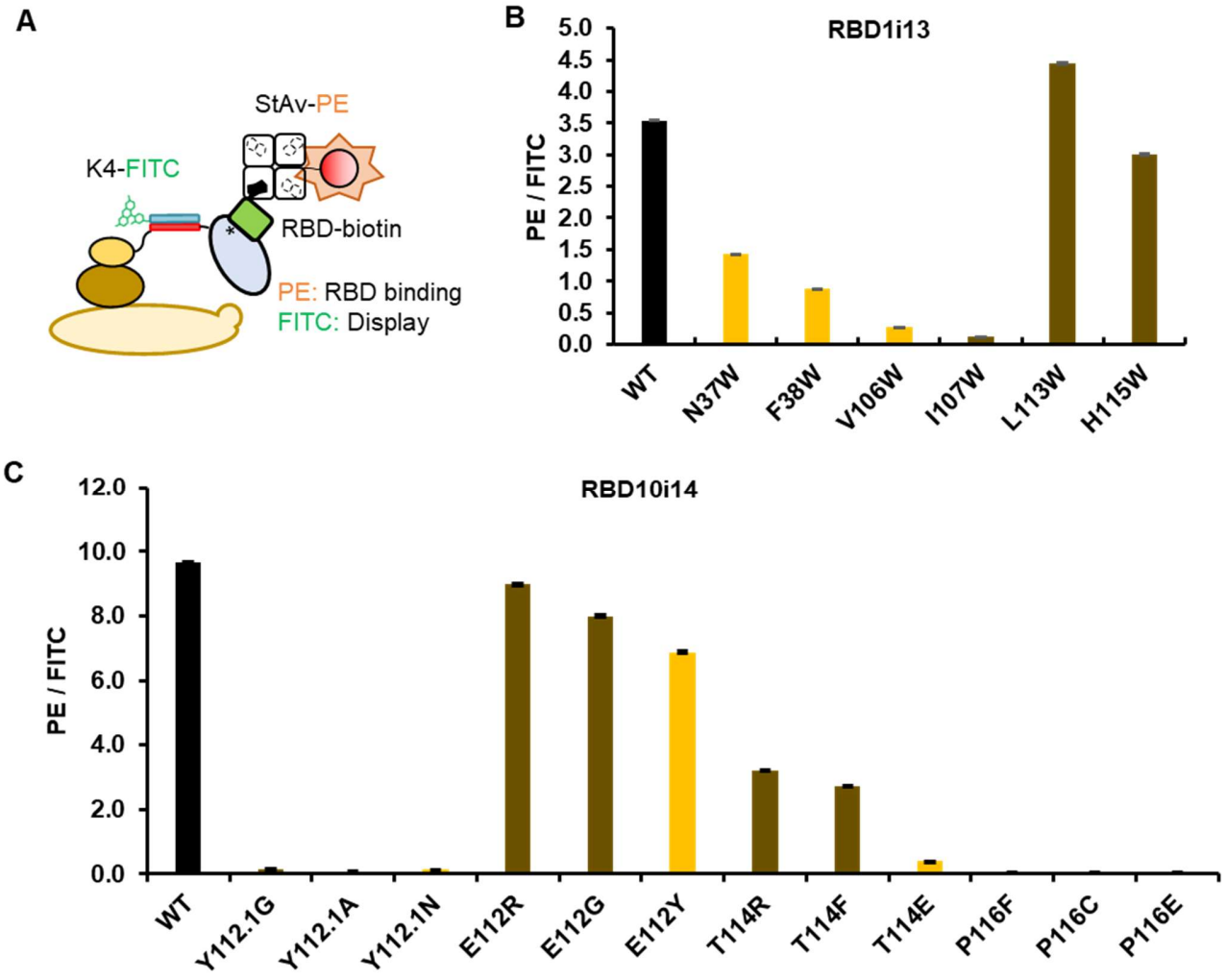

Figure S7. The antigen binding activity on yeast cell surface. (A) Schematic image of evaluation antigen binding to antibody on yeast cell surface. The fluorescence of PE represents the antigen binding activity against RBD, and FITC is used to correct for nanobody display. (B, C) PE/FITC ratio of mutants selected during in silico Trp scanning (B) and in silico single saturation mutagenesis (C). The bar graphs represent the mean of TAMRA/FITC ratio ± standard error of mean. The 3 highest scores were dark yellow, and 3 lowest scores were light yellow.

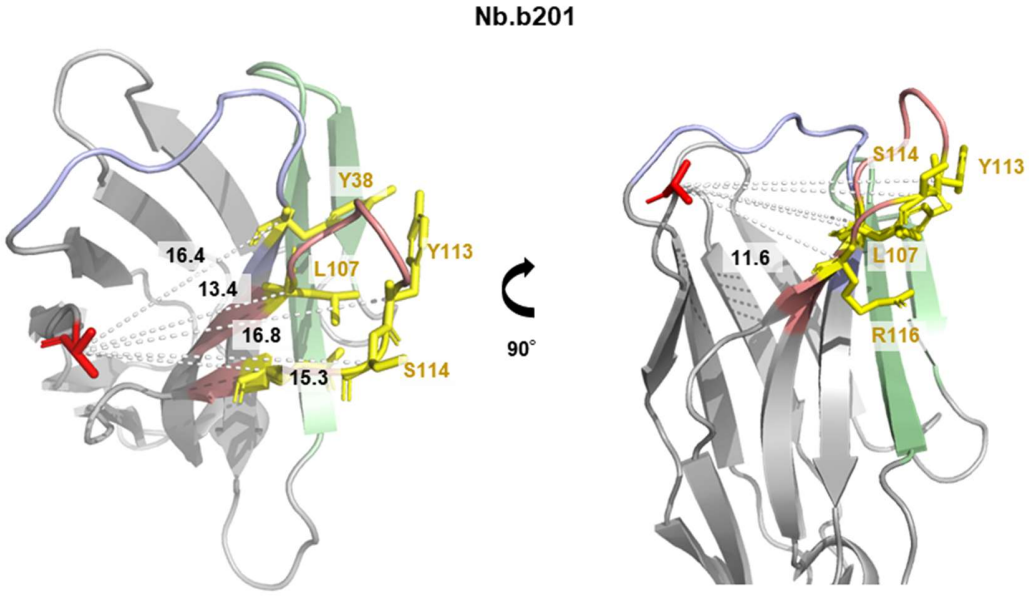

Figure S8. X-ray structure of Nb.b201 derived from NbLib (PBD: 5XVN). The CDR1, CDR2, and CDR3 are blue, green, and pink respectively. The N-terminus and the five positions where Trp residues are enriched were highlighted as red and yellow, respectively.