Supporting Information

Larman et al. 10.1073/pnas.1215549109

SI Methods

Construction of the Ribosome Display Vector. Plückthun and colleagues have optimized vectors capable of displaying single-chain variable fragments (scFvs) on ribosomes. (1, 2) We adapted components of these and other such vectors to our present purpose. Beginning from the 5′ end of the DNA vector, the following parts were assembled as a synthetic gene product (DNA2.0):

i) T7 promoter for in vitro transcription from the DNA library (TAATACGACTCACTATAGGGAGACCACAACGGTTTCCC);
ii) 5′ mRNA stemloop (5′-GGGAGACCACAACGGTTTCCC-3′) to improve transcript stability;
iii) Ribosome binding site for translation of the library;
iv) Kozak sequence for potential use in eukaryotic translation systems;
v) N-terminal 6xHis tag for detection and potential purification of scFv protein;
vi) The variable domain of the light chain was encoded N-terminal to the heavy chain so that PCR recovery of the three diversified complementarity-determining regions (CDRs) (L3, H2, H3) would require the shortest amplicon;
vii) Between the N-terminal variable light chain (V_L) and C-terminal variable heavy chain (V_H) is a "$(G_4S)_3$" linker with optimized codon use (5′-ggtggtggtggttctggtggtggttctggcggcggctccagtggtggtggatcc-3′);
viii) The C terminus of V_H is fused to a linker segment derived from the TolA *Escherichia coli* protein (accession: NP_415267, position 131–214), which provides a spacer between the displayed scFv and the ribosomal tunnel;
ix) 3′ mRNA stemloop (5′-CCGCACACCTTACTGGTGTGCGG-3′) to improve transcript stability.

NotI sites flank the 3′ and 5′ ends of the construct for isolation of the in vitro transcription template. Directional SfiI sites flank the minimal scFv for facile movement of clones into and out of alternative expression vectors.

Quality Control During scFv Selection. We used a positive control scFv and bait pair to optimize our ribosome display selection protocol. Pluckthun and colleagues have used ribosome display to affinity mature an scFv (4c11L34Ser, "Pluck-scFv") to high affinity (K_d = 40 pM) for a peptide derived from the yeast GCN4 protein (3). Our eventual goal was to perform selections on GST-fusion proteins, and so we expressed GST-GCN4 in BL21 *E. coli* cells as a positive-control bait. As a negative-control scFv, a random clone ("rand-scFv") was picked from a naive human repertoire (4) and expressed in the same ribosome display vector backbone. A negative-control peptide, "GST-pep" was used as nonspecific bait. Protocol optimization experiments were undertaken to maximize the amount of both enrichment and recovery of the Pluck-scFv that be could be attained. For most experiments, Pluck-scFv was diluted 1,000-fold into a background of rand-scFv, and GST-GCN4 was diluted 1,000-fold into a background of GST-pep. Our selection protocol typically achieved enrichments of several hundred-fold, and recovery of ∼0.2%. This relatively low rate of recovery is consistent with known inefficiencies inherent to the ribosome display technology (5).

We incorporated a system of quality-control measures to ensure the success of each round of hidden Markov model (HMM) scFv library selection. First, we spiked Pluck-scFv into our HMM scFv library and GST-GCN4 into our selection bait (GST-PVRL4), both at a dilution of 1:1,000. In this way, the efficiency of enrichment and recovery for each selection could be quantitatively monitored using a probe specific for the Pluck-scFv control. If enrichment or recovery of Pluck-scFv was below a threshold, then the selection was considered a failure and repeated. For our selections, we required enrichment of Pluck-scFv to be at least 50-fold and the recovery of Pluck-scFv be at least 0.04%. Second, degradation of mRNA transcripts is a concern with ribosome display, and so we used TaqMan probes targeting the constant 3′ and 5′ ends of the scFv transcript. In the absence of mRNA degradation, these two signals arise with equal strength. The distal, 5′ signal is differentially diminished by degradation, and so the ratio of the two signals can be used to measure the amount of degradation that occurred during the selection. If the ratio of the 5′ signal to the 3′ signal was below our threshold of 1:5, the selection was considered a failure and repeated.

Illumina Sequencing. Libraries for Illumina sequencing were prepared by two rounds of PCR amplification to add the Illumina adapters and barcode sequences. Libraries were PCR-amplified from the in vitro transcription template DNA using the TaKaRa EX HS kit (Clontech). The conditions for the first round of PCR were: 1× TaKaRa EX HS Buffer, 0.2 mM dNTP, 0.4 μM IS7_L3F_PE primer, 0.4 μM IS8_H3R_PE_Multi primer, 0.5 μL TaKaRa Ex HS enzyme, and 1 μL of template per 50-μL reaction. The thermal profile was: [(*i*) 98 °C for 10 s, (*ii*) 50 °C for 30 s, (*iii*) 72 °C for 1 min 30 s] × 10 cycles, (*iv*) 72 °C for 7 min. The conditions for the second round of PCR were: 1× TaKaRa EX HS Buffer, 0.2 mM dNTP, 0.5 μM of IS4_L3F_PE primer, 0.5 μM of the barcoding primer, 0.5 μL TaKaRa Ex HS enzyme, and 1 μL of the first round PCR product per 50-μL reaction. The thermal profile was: [(*i*) 98 °C for 10 s, (*ii*) 60 °C for 30 s, (*iii*) 72 °C for 1 min 30 s] × 10 cycles, (*iv*) 72 °C for 7 min.

As the complexity of the libraries is expected to decrease significantly with each round of selection, we divided the contribution of each library by two for each round of enrichment undergone. For example, if we added 100 ng of input library product to the multiplex pool, then we would add 50 ng of round 1 selected library, 25 ng of round 2 selected library, 12.5 ng of round 3 selected library, and so on. All second-round PCR products were gel-purified before sequencing on an Illumina HiSeq 2000 instrument.

Analysis of High-Throughput Sequencing Results. All reads were separated into samples according to the barcode sequence by the standard Illumina software. Framework sequences were trimmed according to the following rules: L3 and H2 reads were truncated to their respective lengths (36 nt and 39 nt, respectively). H3 reads were trimmed of 5′ and 3′ framework sequences with an error rate of 0.2 using cutadapt (6). Reads were then aligned to consensus sequences with up to two mismatches using bowtie software (7): First, all of the reads in a sample were tallied and an index was built for each sample. Second, each read in a sample was aligned globally against that sample's index with up to two mismatches allowed. The alignment with the highest tally (i.e., the read that occurred most frequently in that sample) was chosen as the consensus sequence for that read. Finally, reads that contain wildcards ("N") or stop codons were discarded. The paired L3-H3 or H2-H3 reads were then joined and the frequency of unique pairs was tallied. For paired L3-H3, we obtained 2.58×10^7, 2.12×10^7, 9.82×10^6, 1.08×10^6, 5.19×10^5, and 5.33×10^5 total reads for the input library, round 1, round 2, round 3, round 4, and GCN4 selections, respectively. After ap-

plying our filtering algorithm, we obtained 1.60×10^7, 1.97×10^7, 8.66×10^6, 8.77×10^5, 4.71×10^5, and 4.97×10^5 reads, respectively. For paired H2-H3, we obtained 2.02×10^7, 1.68×10^7, 7.55×10^6, 8.08×10^5, 3.90×10^5, and 4.10×10^5 total reads for the input library, round 1, round 2, round 3, round 4, and GCN4 selections, respectively. After applying our filtering algorithm, we obtained 1.89×10^7, 7.16×10^6, 6.71×10^6, 6.72×10^5, 3.55×10^5, and 2.94×10^5 reads, respectively. After four rounds of selection the median read depth of the top 10 L3-H3 paired clones was 55.5 and the median read depth of the top 10 H2-H3 paired clones was 289.5. In each of the libraries, there is a long tail of clones that are sequenced only once.

Live-Cell FACS Analysis. Telomerase-large T-antigen–immortalized human mammary epithelial cells (TL-HMECs) were transduced with retroviral constructs expressing human PVRL4 or control (empty vector). For labeling with in vitro-translated scFvs, cells were dissociated from the tissue-culture plate with enzyme-free cell dissociation buffer (Invitrogen), resuspended in Stain buffer (BD Biosciences), and filtered through a 35-μm nylon mesh cell strainer (BD Biosciences). Cells were incubated with in vitro-translated FLAG-tagged scFvs at a 1:100 dilution or anti-PVRL4 mouse monoclonal antibody (R&D Systems) for 30 min on ice, washed twice with Stain buffer, and incubated with M2 anti-FLAG antibody (Sigma) at a 1:100 dilution for 30 min on ice. Labeled cells were washed twice and incubated with Alexa Fluor 488-conjugated goat–anti-mouse secondary antibody (Invitrogen) at 1:500 dilution for 30 min on ice. After a final series of washes, cells were resuspended in Stain buffer. Fluorescent signal was measured on LSR II FACS Analyzer (BD Biosciences) and analyzed with FlowJo software.

1. Hanes J, Plückthun A (1997) In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci USA* 94:4937–4942.
2. Zahnd C, Amstutz P, Plückthun A (2007) Ribosome display: Selecting and evolving proteins in vitro that specifically bind to a target. *Nat Methods* 4:269–279.
3. Hanes J, Jermutus L, Weber-Bornhauser S, Bosshard HR, Plückthun A (1998) Ribosome display efficiently selects and evolves high-affinity antibodies in vitro from immune libraries. *Proc Natl Acad Sci USA* 95:14130–14135.
4. Feldhaus MJ, et al. (2003) Flow-cytometric isolation of human antibodies from a nonimmune *Saccharomyces cerevisiae* surface display library. *Nat Biotechnol* 21: 163–170.
5. Schaffitzel C, Hanes J, Jermutus L, Plückthun A (1999) Ribosome display: An in vitro method for selection and evolution of antibodies from libraries. *J Immunol Methods* 231:119–135.
6. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10–12.
7. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

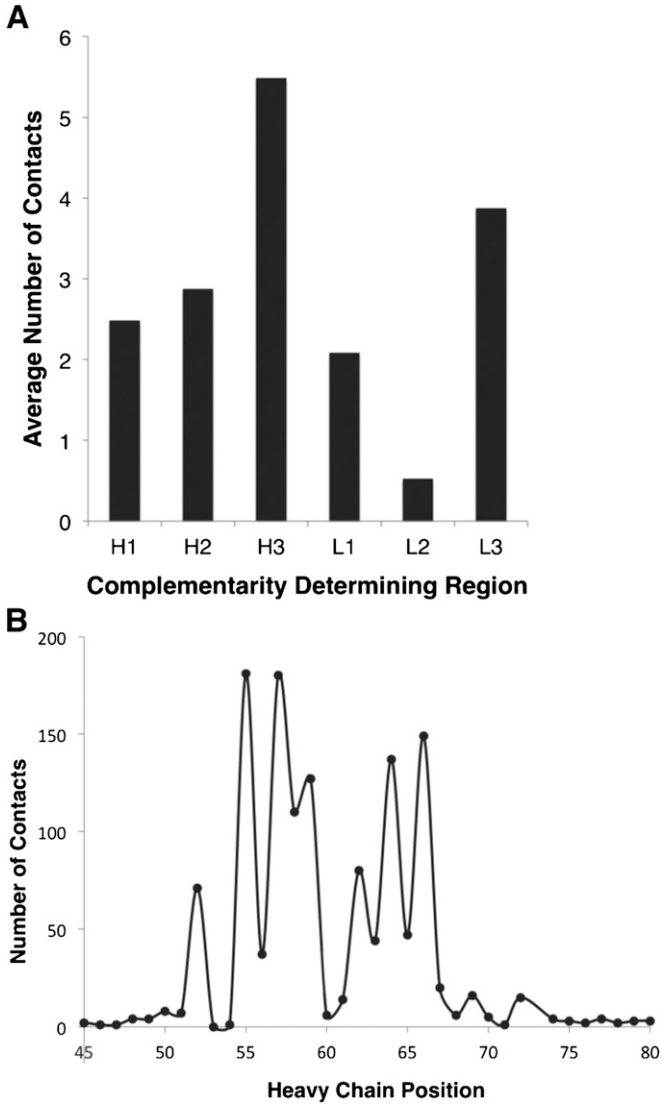

Fig. S1. CDR contact distribution and H2 contact profile. (*A*) Contacts reported in the international ImMunoGeneTics/3Dstructure-DB database. Contact assignment is based on international ImMunoGeneTics' definition of CDR positions. Data were obtained from 241 antibody-antigen cocrystal structures. (*B*) Position-dependent contact distribution in H2. Valleys represent amino acids more likely to play a role in framework stability.

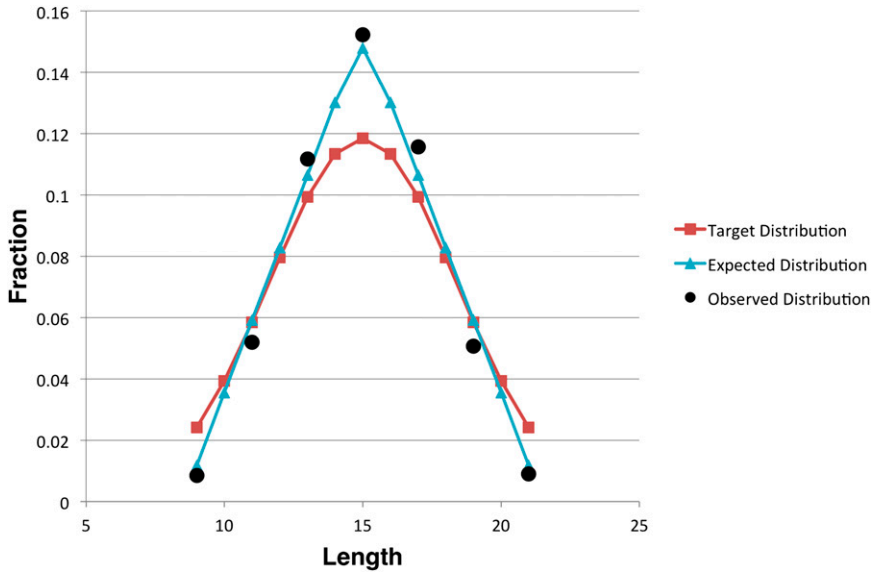

Fig. S2. Length distribution of the H3 CDR library. Target H3 length distribution is based on the high-throughput sequencing of an individual's heavy-chain repertoire. Expected distribution is the calculated fraction of each length based on random ligation of all H3L sequences with all H3R sequences. The observed distribution is based on the analysis of the Illumina sequencing data from the unselected HMM scFv library.

Dataset S1. Vector components and sequences

Dataset S1

Sequences of the single framework scFv construct for ribosome display. Shown are original sequences used for the screen, and the codon optimized sequences for protein expression. Real-time PCR primer and probe sets are also shown.