SUPPORTING INFORMATION

TABLE OF CONTENTS

SUPPLEMENTAL TEXT

Estimation of library complexity and uniformity

In order to achieve a library containing 1,000 copies of 10^9 sequences, we also needed to compensate for losses during library synthesis and purification. Synthesis of 10^9 functional protein sequences requires an input DNA complexity of ~2.5×10^9 due to nonsense mutations arising from stop codons in the random loops and frame shifts based on the analysis by Olson and Roberts, 2007.[1] To achieve an excess of 1,000 copies of each clone, we used 400 pmoles of puromycin-labeled mRNA template for translation per target. Assuming ~5% fusion efficiency,[1] and accounting for losses arising through oligo dT cellulose purification, anti-Flag tag purification, or imperfect reverse transcription efficiency, we estimated recovery of at least 1,000 copies of all functional sequences just prior to the affinity purification step of selection.

To estimate complexity after assembling the DNA library template, we measured the DNA concentration, as each molecule is unique,[1, 2] and then aliquoted and PCR amplified 4.15 femtomoles (2.5×10^9 molecules). After sequencing, 31.3 million sequences from this pool were analyzed using two statistical methods which confirmed input complexity was ~2.5 billion. First, we applied a Monte Carlo analysis. Assuming the simplest possible model—approximately equally abundant numbers of each sequence—we repeatedly simulated random selection-with-replacement of 31.3 million items from a pool with a complexity of *n* possible unique items, and counted the number of singletons and doubletons. Our simulations indicated that *n* in the region of ~2.55 billion is highly consistent with the fraction of singletons/doubletons observed. The actual sequencing result gave 30,934,216 singletons and 187,326 doubletons. Eleven simulations with *n* ranging from 2.55 billion through 2.56 billion averaged 30,934,084 singletons and 189,406 doubletons.

To verify this, we also used non-linear regression models to estimate the complexity by fitting a least-squares model to the copy number distribution. We used a model of the form

$N(k) = C^* \lambda^k * e^{-\lambda} / k!$,

which resembles sampling from a Poisson distribution with mean λ, where N(k) is the count (number of different sequences) for a given number of occurrences k (copy number), and C is the complexity of the library (total number of unique sequences). Given the large number of samples, λ can be assumed to be s, the number of sequences observed, divided by the complexity of the library (*i.e.,* $\lambda = s/C$). Fitting the model for C, using the number of occurrences and respective count numbers, resulted in a comparable estimate of the complexity. The clone frequency profile

of the approximately 5,000-fold PCR amplified sample was very close to a Poisson distribution, with a complexity of approximately 2.5 billion unique sequences (Figure S3A).

After preparing fusions from the naïve library, we expect complexity to decrease towards our target of 10^9 due to the removal of incorrectly translated sequences (*e.g.*, sequences with frame shifts or stop codons) following Flag-tag affinity purification. We also wanted to determine if there is a significant skew in relative sequence frequencies resulting from the display method. A similar non-linear Poisson-based model fit of the input fusion pool copy number distribution model also reasonably fits the frequency data of single and double copy numbers. However, there is a significant difference in the frequency data for higher copy numbers (Figure S3B), where the probabilities associated with sampling clones with higher copy numbers were much less than what was observed. Our initial Poisson analysis assumed that each clone was equally likely to be sequenced, but this assumption is invalid if clone frequencies are not uniform. To estimate the extent of skew in this pool, we generated a simple mixture model of two discrete populations where a fraction of the input population exhibits higher fold over-representation compared to the remaining group. Hence, we added an additional component to the model that corresponds to sampling from a second population based on its fraction of the total complexity and relative frequency. $N(k)_{observed}$ is therefore a linear combination of N(k) for which sampling can occur from either of the two populations with complexities C_1 and C_2. The total complexity is $C = C_1 + C_2$, and the model is given by

$$N(k) = C_1 * \lambda_1^k * e^{-\lambda 1} / k! \; + \; C_2 * \lambda_2^k * e^{-\lambda 2} / k!$$

where

$\lambda_1 = s * \alpha / C_1$ and $\lambda_2 = s * (1- \alpha) / C_2$

α is the sampling fraction that is dependent on the complexities of the two populations and the fold over-representation of population 2 ($fold_o$) through

$\alpha = C_1 / (C_1 + C_2 * fold_o)$

By solving the regression for the parameters C_1, C_2, and α, a number of solutions could emerge as a result of choice in initial values for the regression parameters. To ensure our solution would not be confined to a local minimum around the initial values, we conducted the regression by sweeping through a large number of possible initial values. Of 12,500 trials approximately one-half converged to an identical model with the lowest mean square error (MSE) of 3850. The remainder of the trials produced unrealistic and poorly fitting solutions (Figure S4). The best fit MSE of 3850 is only a small fraction of the data values and is considerably smaller than the best single population Poisson-based model, where the MSE equals 7.94×10^8. The optimal solution predicts a total complexity of about 904 million unique clones, which is very close to our target of

1 billion functional protein sequences (Figure S3B,C). The model suggests that the fraction of the skewed over-represented population is approximately 7.1%, and that sequences within this population are about 5.46 times more frequent than those within the other population (Figure S3C). Therefore, a simple model of two discrete populations in which the majority of the library sequences (~92.9%) amplify uniformly while a small fraction (~7%) behaves with enhanced replication kinetics generates a prediction that closely matches the observed clone frequency distribution. These biases may be due to variable efficiencies in PCR, translation, or fusion formation.

Calculating Fold Enrichment

Using the two population Poisson mixture model, we determined the frequency of unique input library clones from population 1 to be 8.24×10^{-10} (Figure S3C). eFn-anti-IgG1 was sequenced 36 times (1.25×10^{-6} per sequence) yielding a fold enrichment of ~1517. However, the probability distribution around the observed value yields a broad range in frequency from 9.37×10^{-7} to 1.63×10^{-6} using a 97% confidence interval. This indicates enrichment of at least 1,137-fold and up to 1,979-fold for eFn-anti-IgG1. The most frequent MBP-binder, eFn-anti-MBP1, was predicted to range in frequency between 5.59×10^{-7} and 1.15×10^{-6} per sequence, indicating enrichment between 678- and 1,401-fold. However, this clone was identified in the input fusion pool twice, which results in a probability of ~0.66 that it originates from the over-represented population 2 based on our mixture model (Figure S3B). In this case enrichment would fall between 123- and 254-fold. The affinity of eFn-anti-MBP1 is poorer than eFn-anti-MBP2, which supports the hypothesis that the eFn-anti-MBP1 copy number is a product of both functionality and higher initial frequency. No other top clones were identified in the input fusion pool, which suggests the clones likely belong to population 1 based on our mixture model (probability of ~0.94).

Determination of the eFn-anti-IgG1 binding site

It has been hypothesized that IgG(Fc) contains a convergent binding site, with numerous proteins—including protein G, protein A, and peptides evolved *in vitro*—binding at a single consensus site.[3] In order to determine if our selected e10Fn3 also binds in a similar manner, we tested the ability of protein G to compete with eFn-anti-IgG1 in binding to IgG. As shown in Figure S6B, protein G is able to inhibit eFn-anti-IgG1 binding in a concentration dependent manner, indicating that eFn-anti-IgG1 binds IgG at the consensus site. In this experiment, IgG was coated at a concentration that yields half maximal signal by eFn-anti-IgG1-SA-HRP. The K_i

determined by nonlinear regression was 11.6 nM with a Hill slope of 2.2, indicating positive cooperativity by the protein G fragment, which contains two IgG-binding domains. This competition by protein G suggests that the two eFn-anti-IgG1 FG loop cysteines do not mediate binding to IgG by disulfide bond formation—for example, to the exposed hinge region.

Using e10Fn3-based affinity reagents as western blot probes.

We wanted to test whether our e10Fn3-based binders can be used to replace antibodies for western blot- ting. Figure S12 demonstrates that both eFn-anti-IgG1-SA-HRP and eFn-anti-MBP1-SA-HRP specifically recognize target protein transferred onto nitrocellulose after SDS-PAGE. Figure S12A demonstrates that eFn-anti-IgG1 possesses excellent sensitivity for detection of IgG (in the low ng range) and also high specificity, with no detectable background from albumin, IgA, or crude mammalian cell lysate. All samples were loaded without reducing agent, except for the reduced IgG sample. The limited detection of reduced IgG may indicate that refolding is able to take place on the membrane if intrachain disulfides are intact. To demonstrate MBP detection by eFn-anti-MBP1-SA-HRP, we detected the expression of an unrelated e10Fn3-MBP fusion (Figure S12b). The specificity of eFn-anti-MBP1-SA-HRP is excellent as no other proteins were detected in the crude *E. coli* lysate. eFn-anti-MBP2, which has a better monomeric equilibrium binding constant than eFn-anti-MBP1, was also tested in this assay but did not display improved functionality (data not shown).

Determining the minimum sequencing needed. The level of enrichment achieved is significant enough to clearly identify the best target binders without screening beyond the initial validation (see Figure 2). In order to determine if even less sequencing could be performed, thereby increasing throughput, we fit each selected pool to a Poisson mixture model, and estimated the probability of observing the top target binders over the random carry-over with half as much sequencing (Figure S13). Using the observed binder frequency, we determined that eFn-anti-IgG1 would be identifiable without screening even at the lowest copy number values with significant probabilities, while eFn-anti-MBP1 and eFn-anti-MBP2 could require screening. There is an approximately 7% probability that eFn-anti-MBP1 would not be identifiable at all by screening (identified 7 times or less). To test this assumption, we sampled two random halves of the selected pools and identified the top binders. The observed samplings match the models well. eFn-anti-IgG1 copy numbers vary widely over the probability range expected for the observed frequency, yet is still sufficiently separated from bulk carryover to not require screening. eFn-anti-IgG2 however would be detected only after considerable screening in one sampling. eFn-

anti-MBP1 and eFn-anti-MBP2 were visible within the top six of both halves, indicating that ~14 million sequences may be sufficient for identifying binders.

Further advancements could be made to improve the efficiency and economy necessary for highly multiplexed ligand generation. While our method results in significant improvements over our previous batch affinity enrichment method, even greater partitioning efficiencies could be achieved through optimization of bead quantity, target loading, and enrichment stringency. However, without changing the selection conditions as described here and using the newest Illumina platform available at this time (Hi-seq 2000), approximately 12-14 target-specific pools could be analyzed per lane, or approximately 200 targets per run on one instrument. To accomplish this, barcodes must be used to distinguish target-specific pools. One strategy would be to insert an identifying code into the reverse transcription primer between the Fn annealing region and the flow cell annealing region. This would enable separation the naïve pool into 12 or more barcoded pools at the reverse transcription step, which is the last step in the fusion preparation process (see Figure 1 in the main text). However, this would require an increased sequencing length in order to read through the Fn annealing region in order to sequence the barcode. As an alternative, an identifying code could be placed closer to the random regions within the template DNA by unique assignments at third bases of codons. For example, the third base of Thr70 preceding the random FG loop and Pro81 following the FG loop (see Figure S1) could be assigned to produce up to 16 unique templates. This, however, would require the entire fusion preparation process to be performed separately for each of 12-16 barcoded libraries.

Main text full reference

[6] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, E. C. M. Chiara, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers , M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S.w W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F.L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, A. J. Smith. *Nature* **2008**, *456*, 53.

SUPPLEMENTAL FIGURES

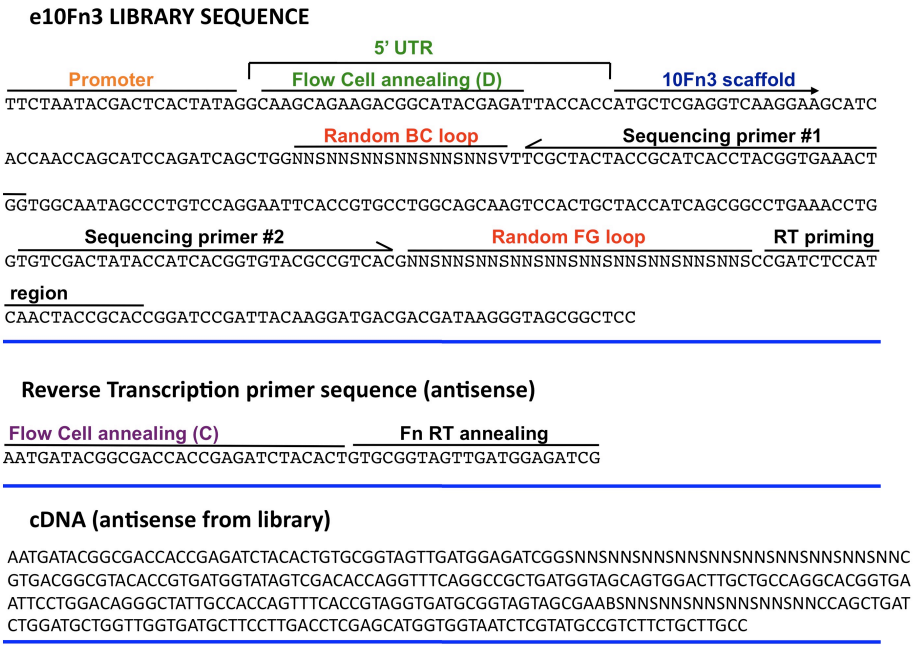

Figure S1. e10Fn3 library sequence adapted for Illumina deep sequencing. Our library encodes a variation on the "monobody" scaffold, which is based on the tenth fibronectin type III domain of the human fibronectin protein (10Fn3), first adapted by Koide et al. for phage display.[4] Our modifications include the truncation of the first seven N-terminal residues and the implementation of an expanded loop randomization strategy targeting seven BC loop residues and ten FG loop residues.[1] Here, we used a variant of this scaffold, termed "e10Fn3", which contains additional mutations that improve solubility and expression of both in *E. coli* and *in vitro*.[5, 6] We have also developed a strategy to incorporate Illumina flow cell annealing and bridge amplification regions within this library DNA template, while utilizing e10Fn3-specific primers for sequencing, eliminating the need for adapter ligation and restricting sequencing to the coding region of interest. Placing the sequencing primers adjacent to the diverse random region also facilitates cluster identification thereby maintaining read quality for amplicons with identical 5' and 3' ends. The DNA-encoded library is assembled from synthetic oligonucleotides[1, 5] (Table S1) and converted into a naïve mRNA-protein fusion pool as previously described,[5,7] except for the inclusion of DNA sequences that enable eventual Illumina flow-cell annealing and bridge amplification. One flow-cell annealing region (D/D') is incorporated into an altered 5' UTR. The second flow-cell annealing region (C'/C) is incorporated during reverse transcription as a tail on the reverse primer (Figure 1b step 6, Figure S1). This strategy enables directional sequencing of the library. After a single round of CFMS-based affinity enrichment performed as described in Olson et al. 2011,[5] cDNA is collected and can either be amplified with oligo C and oligo D (Table S1) or directly subjected to sequencing with an Illumina instrument (either GAIIx or HiSeq) (optimal chip density was more easily achieved after PCR amplification, data not shown). During Illumina sequencing, the BC and FG loop sequences are determined via paired reads using two e10Fn3-specific primers. The first e10Fn3-specific primer anneals to the fibronectin scaffold DNA one base downstream of the BC loop while the FG loop requires a second e10Fn3-specific primer that anneals one base upstream from the FG loop. Once constructed, this single library can be scaled up and screened against any desired number of targets.

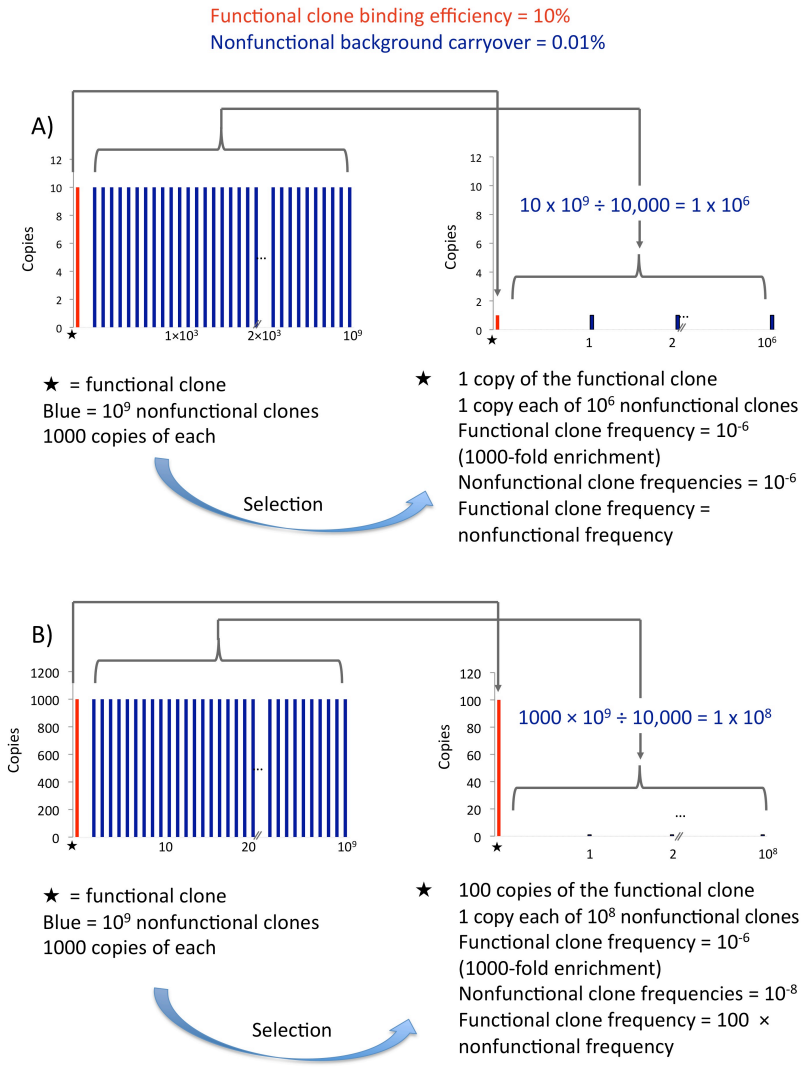

Figure S2. Functional vs. nonfunctional clone frequencies after one round of selection depend on copy number.

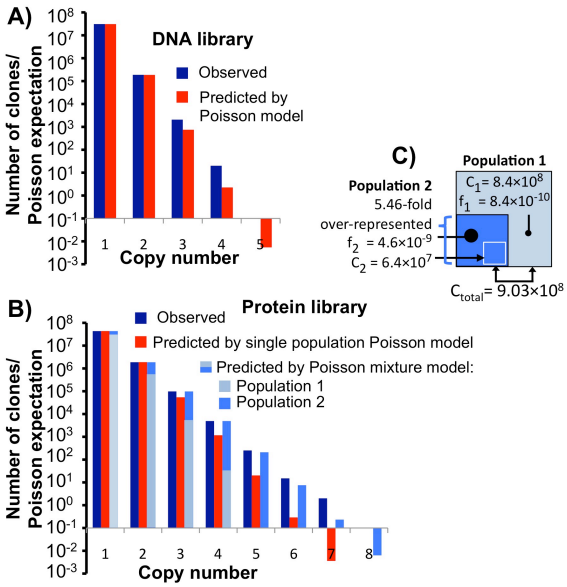

Figure S3. Input library complexity and uniformity. A) The input DNA pool was sequenced using one lane on an Illumina GAIIx, generating 31 million sequences that passed quality filters. The number of unique clones sequenced at each copy number is shown. This distribution was used to calculate a complexity of 2.5 billion, which agrees well with the theoretical target complexity. The calculated Poisson expectation based on this complexity given the amount of sampling is also shown. B) Input library fusions were sequenced using two lanes (47.5 million sequences). Input fusions were purified to eliminate improperly translated clones. A uniform population would generate a Poisson expectation that diverges significantly from what is observed at higher copy numbers. Including a second discrete population of higher relative frequency generates a distribution that tightly fits the observed data, providing a measure of skew. The linear combination of the number of clones predicted to be sequenced from each population is illustrated by the dark (population 1) and light (population 2) blue bars, demonstrating the likelihood that naïve pool clones are part of either population given the observed copy number. C) The total complexity (C_1+C_2) estimated by this fit is ~904 million, with 7.1% of all clones over-represented by 5.46-fold (population 2) as compared to the bulk (93%) of the library clones. This model was used to estimate initial clone frequencies for each group (f_1 and f_2) in order to determine fold enrichment of selected molecules.

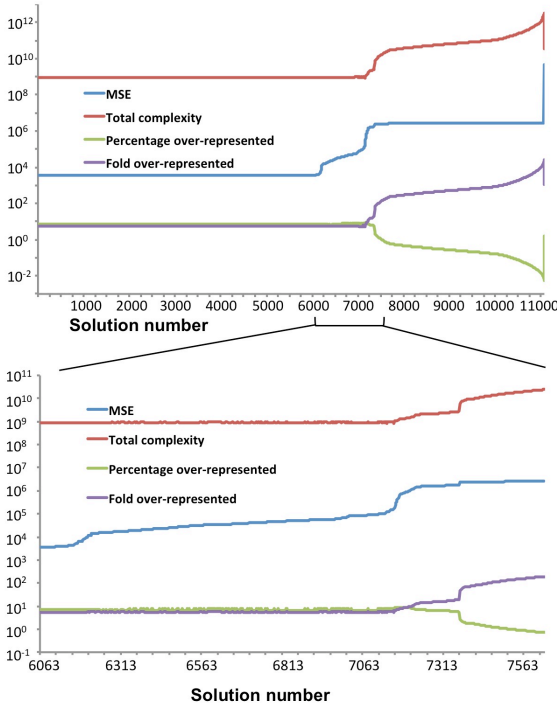

Figure S4. Distribution of solutions from non-linear regression of Poisson mixture models ranked by mean square error. Approximately 1250 trials were run with variations in the initial values of parameters C_1, C_2 and α. The top chart shows all solutions converted to total complexity (red), percentage of the complexity that belongs to the higher frequency pool (green), and the extent of over-representation of pool 2 sequences (purple). Solutions with negative values and equal frequencies (single populations) were eliminated. One identical solution with an excellent fit of the data is found in approximately half of the trials (MSE equals 3,580). Slight variations in the three parameters rapidly increase MSE, demonstrating the sharpness of the solution minimum. The lower chart expands this region to show the transition from these solutions to a set of alternative solutions with approximately 1,000-fold higher MSE. These solutions are characterized by increasing complexity with an increasing skew in over-representation of a decreasing fraction of the pool. The best fit of a single population Poisson-based model has a much higher MSE, of approximately 8×10^8 (data not shown).

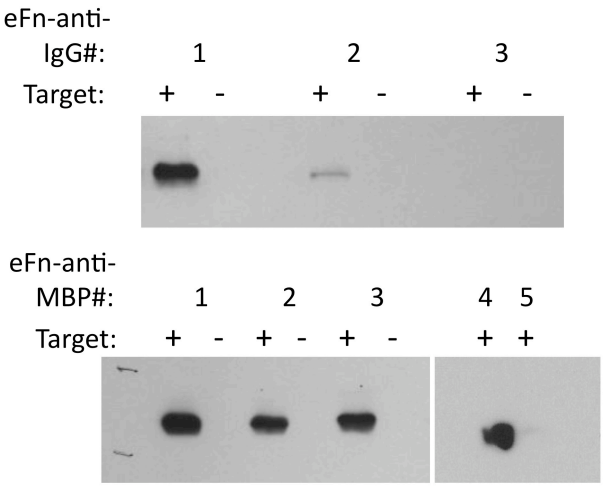

Figure S5. Western blots for binder validation. After reconstruction by PCR, each clone was transcribed and translated *in vitro* and bound to beads plus and minus target. After washing and elution with SDS loading buffer, samples were analyzed by anti-Flag western blot. Samples were characterized as ++ (efficient in binding, e.g. eFn-anti-IgG1 or eFn-anti-MBP1), + (e.g. eFn-anti-IgG2 or eFn-anti-MBP5), or – (eFn-anti-IgG3) (Figure 2D-E).

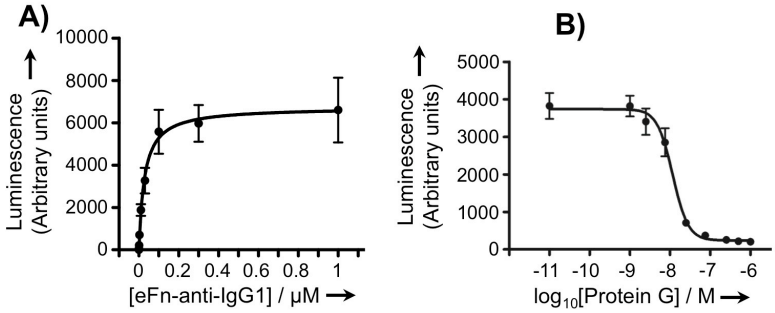

Figure S6. Affinity of eFn-anti-IgG1 to huIgG. A) We determined the affinity of monomeric biotinylated eFn-anti-IgG1 by binding various concentrations to IgG coated multiwell plates, followed by SA-HRP detection after washing. Non-linear regression generates a K_D of 27.7 nM. The hill coefficient is 1.005, indicating that surface effects do not generate a cooperative response. B) Binding is blocked by protein G, indicating that eFn-anti-IgG1 binds a convergent epitope on IgG(Fc). IgG was coated at a concentration that generates half-maximal signal by eFn-anti-IgG1-SA-HRP, and the K_i for protein G is 11 nM. A hill slope of 2.2 indicates positive cooperativity of the two Fc-binding domains per protein G molecule. Both assays were performed in triplicate, with error bars representing standard deviations.

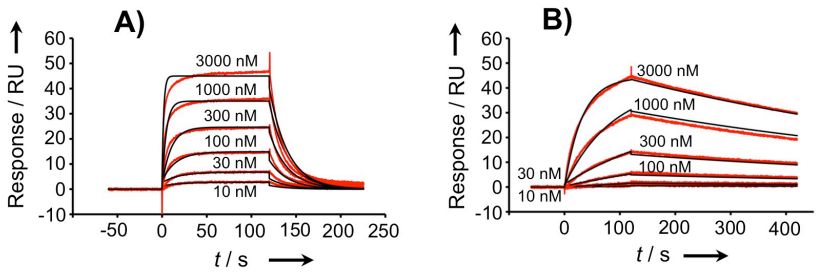

Figure S7. Kinetics and affinity of eFn-anti-MBP1 and eFn-anti-MBP2. A) SPR data from eFn-anti-MBP1 binding to MBP, fit to an on-rate of 1.69×10^5 M^{-1} s^{-1} and an off rate of 4.77×10^{-2}, resulting in K_D = 282 nM. B) SPR data from eFn-anti-MBP2, fit to an on-rate of 1.01×10^4 M^{-1} s^{-1} and k_{off} = 1.30×10^{-3} s^{-1} resulting in K_D = 129 nM. Black curves denote normalized observed data while red curves denote fitted data.

pJD3: C-terminal Alkaline Phosphatase (D153G/D330N), His-tag

From pET-28

BglII — T7 promoter → — lac operator — XbaI

AGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAAT

rbs — NdeI — PelB — XhoI

AATTTTGTTTAACTTTAAGAAGGAGATATACATATGAAATACCTGCTGCCG...48bp...GCGATGGCCATGCTCGAG

MetLysTyrLeuLeuPro...16aa...AlaMetAlaMetLeuGlu

KpnI — BamHI — Alkaline Phosphatase

GGTACCGGATCCGGTGGTAGCGGGACACCAGAAATGCCTGTT...1311bp...AAAGCCGCTCTGGGGCTGAAAGTCGAG

GlyThrGlySerGlyGlySerGlyThrProGluMetProVal... 437aa...LysAlaAlaLeuGlyLeuLysValGlu

His-Tag

CACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGA

HisHisHisHisHisHisEnd

T7 Terminator

GCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG...

Figure S8. pJD3 cloning and expression region.

pAO5-btn: C-terminal BirA biotinylation tag, His-tag

From pET 11

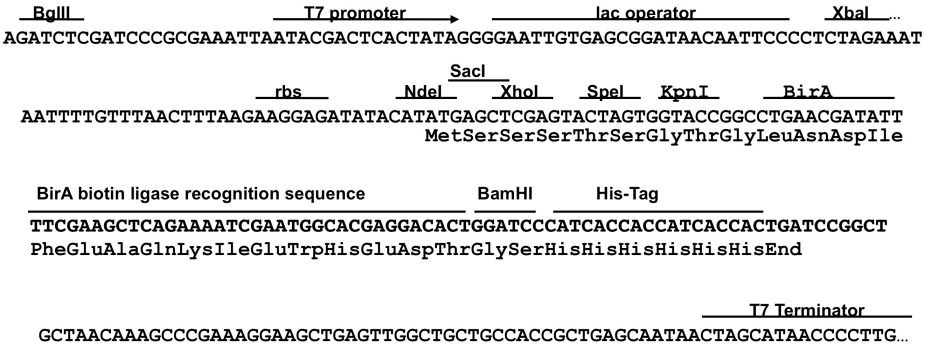

Figure S9. pAO5-btn cloning and expression region.

pAO9-btn: C-terminal biotin-, MBP-, and His-tag

From pET-28

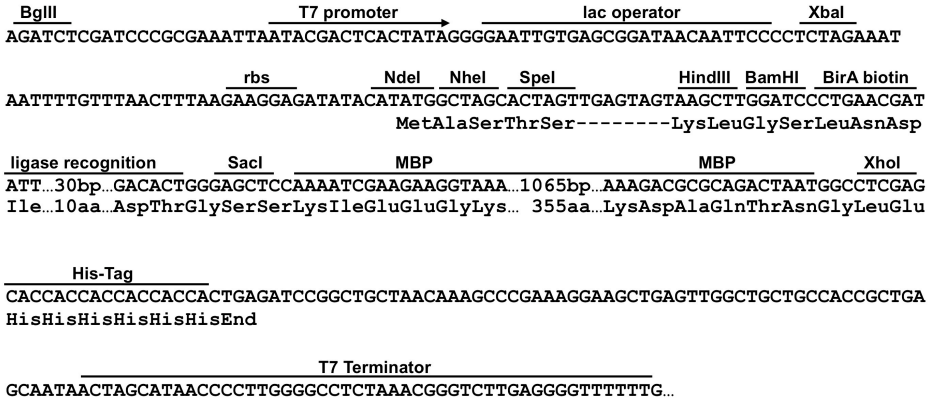

Figure S10. pAO9-btn cloning and expression region.

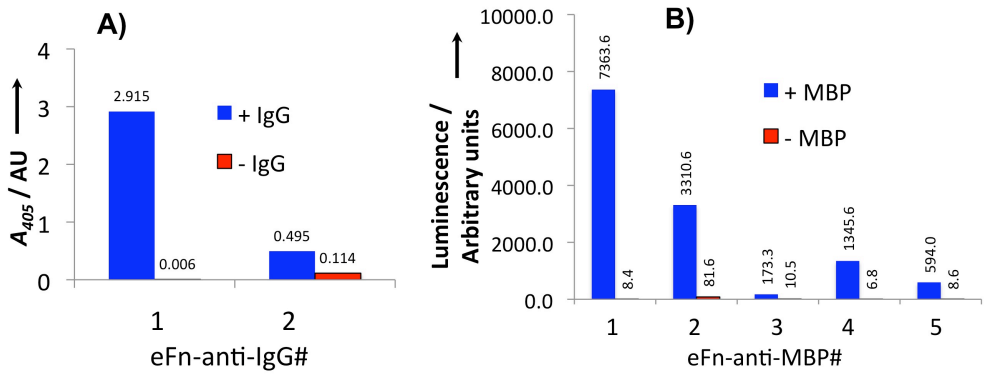

Figure S11. Enzyme-linked detection assays to compare binding efficiency and background. A) eFn-anti-IgG1-AP (5 µg/mL) and eFn-anti-IgG2-AP (5 µg/mL) were bound to multiwell polystyrene pates coated with IgG(Fc) or wells blocked with BSA as a negative control. Colorimetric detection was mediated by PNPP hydrolysis. B) MBP-binders were biotinylated with BirA for conjugation to SA-HRP (100 ng/mL e10Fn3, 25 ng/mL SA-HRP). Complexes were bound to wells with or without MBP and detected for luminescence using SuperSignal ELISA pico (Pierce) on an Lmax (Molecular Devices) luminometer.

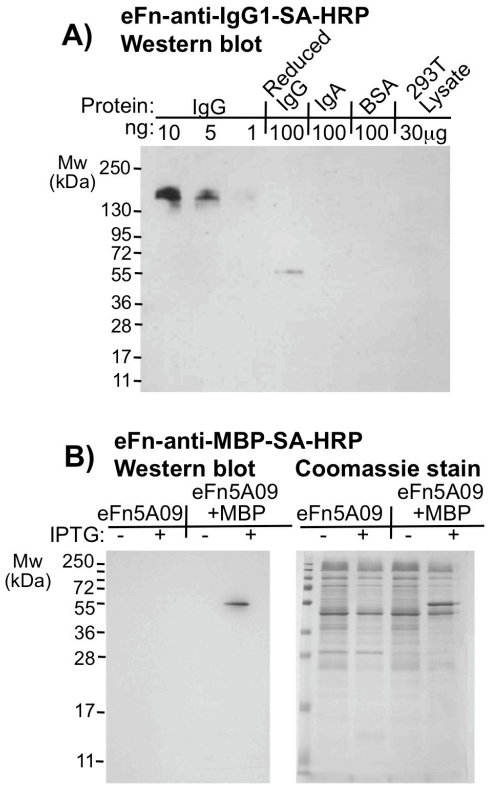

Figure S12. Antibody-free western blots with selected e10Fn3s. A) eFn-anti-IgG1-SA-HRP detects huIgG in the low nanogram range. Reducing IgG prior to gel loading results in poor detection of the IgG heavy chain, indicating that intrachain disulfides may be necessary to allow refolding on the membrane. There is little to no background from excess BSA, IgA, or cell lysate. B) Monitoring eFn-MBP fusion expression induction using eFn-anti-MBP-SA-HRP via Western blotting. eFn-anti-MBP1 only detected a control e10Fn3 (eFn6A09) when that protein was fused to MBP, and generated no background response to other cellular proteins. Blots were incubated with 100 ng/mL e10Fn3s plus 25-50 ng/mL SA-HRP in TBS plus 0.5% tween and 5% w/v milk at room temperature.

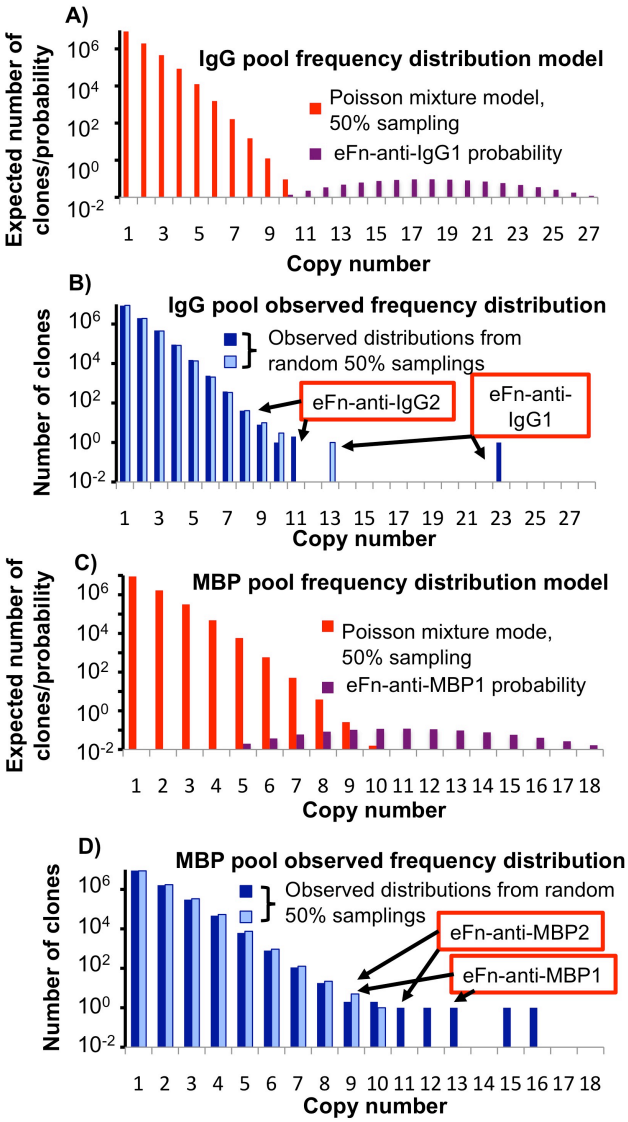

Figure S13. 50% less sequencing is sufficient to discover top binders. A-B) eFn-anti-IgG1 frequency varies significantly (13 and 26) in two randomly sampled halves of the sequencing data as expected based on the binomial probability distribution, yet is clearly separated from the random background (non-binder carryover). eFn-anti-IgG2 would not be detected in one of the two halves without extensive screening. C-D) eFn-anti-MBP1 and eFn-anti-MBP2 are identifiable with minimal sequencing (*i.e.*, both are within the top 3–6), however the probability density of eFn-anti-MBP1 suggests this level of sequencing will not guarantee discovery of true binders without an increase in separation efficiency. Note that this is only instructive in that if the frequencies are as observed or higher, it will be above noise within this reduced level of sequencing.

Table S1. Oligonucleotide sequences
All oligos longer than 49 bases were purified by denaturing urea PAGE

Library construction and sequencing oligos

FnOligo1DS	TTCTAATACGACTCACTATAGGCAAGCAGAAGACGGCATACGAGATTACCACCATGCTCGAGG
FnOligo2DS	GAGATTACCACCATGCTCGAGGTCAAGGAAGCATCACCAACCAGCATCCAGATCAGCTGG
FnOligo3C25	ACCAGCATCCAGATCAGCTGGNNSNNSNNSNNSNNSNNSVTTCGCTACTACCGCATCACCTACG
FnOligo4	GCACGGTGAATTCCTGGACAGGGCTATTGCCACCAGTTTCACCGTAGGTGATGCGGTAGTAGCG
FnOligo5	CCTACCGGTCTCAGCTGATGGTAGCAGTGGACTTGCTGCCAG
FnOligo6	CCTACCGGTCTCACAGCGGCCTGAAACCTGGTGTCGACTATACCATCACGGTGTACGCCGTCACG
FnOligo7	CGGTAGTTGATGGAGATCGGSNNSNNSNNSNNSNNSNNSNNSNNSNNCGTGACGGCGTACACCGTGA
FnOligo9	GGAGCCGCTACCCTTATCGTCGTCATCCTTGTAATCGGATCCGGTGCGGTAGTTGATGGAGATCGG
FnRTDS	AATGATACGGCGACCACCGAGATCTACACTGTGCGGTAGTTGATGGAGATCG
Oligo C	AATGATACGGCGACCACCGAGATC
Oligo D	GCAAGCAGAAGACGGCATACGAGA
FNBCSEQ	CCAGTTTCACCGTAGGTGATGCGGTAGTAGCGA
FNFGSEQ	TGTCGACTATACCATCACGGTGTACGCCGTCAC

Clone reconstruction oligos

I1.36BC	CAGCATCCAGATCAGCTGGCCGGCGACCACGTACAC
I1.36FG1	GATCGGGTTCGGGCGGAGCTGGCACTGGTTGCAGGG
I1.36FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGGTTCGGGCGGAG
I3.19BC	CAGCATCCAGATCAGCTGGGGGTCCGCCAAGACGAC
I3.19FG1	GAGATCGGCATGGACTTCTTGAGGAAGAAGGTGCCCAAC
I1.36FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGCATGGACTTCTTG
I5.15BC	CAGCATCCAGATCAGCTGGCTGACGCTGGAGATCGC
I5.15FG1	GATCGGCCTGATCAACCGCTCGATGTACTCCCACCA
I1.36FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGCCTGATCAACCG
M3.22BC	CAGCATCCAGATCAGCTGGGCCGGGTTGTGGTTGG
M3.22FG1	GATCGGGCTGTAGTTCGGCGTCCAGGACGCCCGGTA
M3.22FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGGCTGTAGTTCGG
M4.20BC	CAGCATCCAGATCAGCTGGCAGGCGCCCCACCTGTTC
M4.20RG1	GATCGGGCGGCCGGGCAGGAGCAAGTACATGAACTG
M4.20FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGGCGACCGGGCAGG
M6.17BC	CAGCATCCAGATCAGCTGGAACATGAGGAGCTGGTTCC
M6.17FG1	GATCGGCCACAACGTGTAGGGAGGCGTGCGGGGCCA
M6.17FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGCCACAACGTGTAG
M8.13BC	CAGCATCCAGATCAGCTGGCAGCTGTACACGCGGATG
M8.13FG1	GATCGGGGAGATGTGGTACAGGAAGAAGCTGGGGATC
M8.13FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGGGAGATGTGGTACAG
M10.13BC	CAGCATCCAGATCAGCTGGCTGCGCTGGACGAGCAG
M10.13FG1	GATCGGCCAGTTGGCCATCAGCCAGGGCCAGGC
M10.13FG2	TCGGATCCGGTGCGGTAGTTGATGGAGATCGGCCAGTTGGCCATC
3FLAG30dA	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTACTTATCGTCGTCATCCTTGTAATCGGATCCGGTGCGGTAG

Vector construction/cloning oligos

pelB-MCSFOR	GAGAGGCATATGAAATACCTGCTGCCGACCGCTGCTGCTGGTCTGCTGCTCCTCGCTGC
pelBREV	CCTCTCGGATCCGGTACCCTCGAGCATGGCCATCGCCGGCTGGGCAGCGAGGAGCAGCAG
phoAFOR	GAGGAGGGATCCGGTGGTAGCGGGACACCAGAAATGCCTGTTCTGG
phoAD153GREV	CCTCCTGGTCTCACCCTGCAACTCTGCGGTAGAAACG
phoAD153GFOR	GGAGGAGGTCTCCAGGGTGCCACGCCCGCTGC
phoAD330NREV	CTCCTCGGTCTCTTCTGTTTATCGATTGACGCACCTTC
phoAD330NFOR	GGAGGAGGTCTCACAGAATCATGCTGCGAATCCTTGTGG
phoAREV	CTCCTCGTCGACTTTCAGCCCCAGAGCGGCTTTC
AO5btnFOR	CCTTGGTCCCATATGAGCTCGAGTACTAGTGGTACCGGCCTGAACGATATTTTCGAAGCTCAG
AO5btnREV	CCACCAGGATCCAGTGTCCTCGTGCCATTCGATTTTCTGAGCTTCGAAAATATCGTTCAGG
AO9btnFOR	GAGAGGGGATCCCTGAACGATATTTTCGAAGCTCAGAAAATCGAATGGCACGAG
AO9btnREV	CCTTCCGAGCTCCCACGGCCCTCGATAGTGTCCTCGTGCCATTCGATTTTCTGAG
eFn5NdeI	TTTACAATTCATATGCTCGAGGTCAAGGAAGC
Fn3KpnI	GGAGGAGGTACCGGTGCGGTAGTTGATGGAG

DETAILED METHODS

Library construction. The e10Fn3 library was assembled from eight oligonucleotides similar to the scheme described in Olson et al. 2007.[1] This library incorporates the solubilizing modifications as described in Olson et al. 2011[5] (all oligonucleotide sequences are listed in Table S1). Here, FnOligo1(DS) and FnOligo2(DS) were modified at the 5' UTR to enable incorporation of a sequence for Illumina flow-cell annealing. All PCR reactions utilized the KOD polymerase system (EMD). Briefly, using primers FnOligo2(DS) and FnOligo4, 0.1 pmol of FnOligo3 were extended and amplified approximately 25-fold in a 25 µL PCR reaction. This cassette which contains the random BC loop sequence was further extended and amplified approximately 8-fold in a 200 µL PCR reaction using primers FnOligo1(DS) and FnOligo5. For the FG loop cassette, 1 pmol of FnOligo7 was extended and amplified approximately 20-fold in a 200 µL reaction using FnOligo6 and FnOligo9. After spin column purification (Qiagen), both cassettes containing the randomized loop regions were digested using BsaI in 100 µL reaction volumes. After spin column purification, the digested products were then ligated using T4 DNA ligase (NEB) in a 60 µL reaction volume. The ligation product was separated on a 2% agarose gel, extracted, and spin-column purified (Qiagen). The concentration was measured by spectrophotometry, with a yield of ~1.2 pmoles, or 7.2×10^{11} molecules, all of which are predicted to be unique due to the improbability of identical products being formed during ligation.[2] The library was serially diluted to 2.5×10^{9} molecules per aliquot for this selection experiment.

Fusion preparation. Our goal was to generate sufficient quantities of puromycin-labeled mRNA library for numerous selection experiments with a high excess of library clones. To do this, we amplified 2.5×10^{9} unique library DNA molecules in a 5 mL total reaction volume with 19 cycles of PCR using FnOligo1(DS) and FnOligo9. After phenol/chloroform extraction and ethanol precipitation, the entire sample of DNA was used as a template for T7 RNA polymerase runoff transcription in a 5 mL reaction (Ambion). The RNA sample was phenol/chloroform extracted and ethanol precipitated and the yield was calculated to be 33.6 nmoles. This sample was ligated to the pF30P linker (40 nmoles) via a splint oligo (37 nmoles) using T4 DNA ligase (NEB) in a 3 mL reaction. The ligation product was separated by denaturing urea PAGE (4.5%) and extracted by electroelution. After ethanol precipitation, the total mRNA-puromycin template yield was calculated to be 2.9 nmoles. In order to achieve our goal of greater than 1000 copies of all mRNA-protein fusions per target, we estimated that 400 pmoles of template was sufficient to carry through to selection. Therefore, this preparation yielded template sufficient for seven

selection experiments. The naïve library preparation was performed as described previously.[5, 7] For the selection of ligands against two targets, 800 pmoles of template were translated in a 2 mL reaction using rabbit reticulocyte lysate (Ambion) for one hour at 30 °C. Fusion formation was enhanced by addition of KCl (500 mM final) and $MgCl_2$ (60 mM final) and incubation at room temperature for 30 minutes. Fusions were purified by binding to 40 mg oligo dT cellulose (Invitrogen) in 20 mL of dT binding buffer (100 mM Tris-HCl, pH 8, 1 M NaCl, 10 mM EDTA, 0.2% Triton X-100) for one hour at room temperature. The sample was washed twice with binding buffer and twice with ice cold TBS. The purified fusions were eluted at room temperature using three 500 µL aliquots of 5 mM Tris-HCl, pH 8 and then were reverse transcribed with Superscript II (Invitrogen) using primer FnRTDS (2 µM). Following RT, EDTA was added and anti-Flag affinity purification was then used to remove nonfunctional clones (e.g., those sequences with frameshifts) as well as remove mRNA lacking fused protein. The fusion sample was bound to 40 µL M2 resin (Sigma) for one hour at 4 °C followed by washing four times with TBS, 0.05% tween 20. The purified fusions were eluted with 800 µL of 0.15 mg/mL 3X FLAG peptide (Sigma) in TBS, 0.05% tween 20. BSA was added to 1 mg/mL prior to the affinity enrichment step. One percent of the purified fusion pool was reserved for sequencing analysis.

Affinity enrichment. For target-coupled bead preparation, both IgG(Fc) (Rockland Immunochemicals) and MBP purified from *E. coli* via amylose affinity resin (NEB) were biotinylated using sulfo-NHS-LC-biotin (Pierce) per the manufacturer's recommendations (50 µM protein, 250 µM sulfo-NHS-LC-biotin in PBS, incubated for 30 minutes at room temperature). Neutravidin-coupled epoxy M270 Dynabeads (Invitrogen, Carlsbad, CA) were prepared per the manufacturer's specifications. Briefly, we incubated 1 mg/mL neutravidin (Pierce) in 100 mM sodium phosphate buffer with 1 M ammonium sulfate with 10^9 beads for 48 hours at room temperature, followed by blocking with TBS. For both selections, 2 µg of each target protein was bound to 4×10^6 beads in PBS, 0.05% Tween 20. Prior to the affinity enrichment step, the purified fusions were incubated with 10^7 neutravidin-coupled beads for 30 minutes four times to help remove neutravidin and matrix-binding sequences. The CFMS affinity enrichment platform was assembled as previously described. Briefly, three 0.25"×0.25"×0.5" NdFeB magnets (B448, K & J Magnetics, Inc.) were affixed to ~10 cm of PFA tubing (0.0625" outer diameter, 0.04" inner diameter, IDEX) attached to a syringe pump (SP1000, Next Advance). Equal fractions of the purified, pre-cleared fusion pool were added to MBP and IgG(Fc)-coupled beads. After binding for one hour at room temperature, beads were separated by withdrawing the sample at 30 mL/hr. The trapped beads were washed with buffer (TBS, 0.05%

tween 20, 1 mg/mL BSA) for 3.5 minutes at 30 mL/hr. The tubing was then removed from the buffer reservoir and buffer was evacuated. The magnets were then removed and beads were resuspended with 100 μL of the PCR reaction mixture. The resuspended beads were incubated at 95 °C for 2 minutes after which the disassociated cDNA was PCR amplified (KOD, EMD) using flow-cell annealing region primers (Oligos C and D, see Table S1). Based on estimated recovery in mock trials, we determined that 17 PCR cycles were sufficient to enable quantitation by UV absorbance while limiting amplification to below saturation. Our library may be used for deep sequencing of the cDNA without amplification (*i.e.*, the cDNA present immediately after selection); however, optimal chip density was more easily achieved after PCR amplification (data not shown).

Sequencing. Amplified input DNA, input fusion cDNA, and selected samples were sequenced on the Illumina Genome Analyzer IIx. Samples were loaded at a concentration of 6 pM and hybridized to an Illumina flowcell via the Illumina cluster station. The cluster station performed bridge amplification to amplify single DNA molecules 35 times into clusters. Each cluster was then linearized, blocked, and hybridized with the BC loop sequencing primer (0.5 μM in hybridization buffer). The flowcell was loaded and run with the Paired-End recipe, 2×42 base pairs, adding the FG loop primer (0.5 μM) for read 2 sequencing. Individual nucleotides of each cluster were sequenced base by base. The Illumina Sequencing Control Software produced image intensities and quality-scored base calls in real time. We performed downstream analysis using software developed internally. The paired read 1 (BC loop) should have the pattern of A + 19-base random region + 24-base scaffold sequence, while read 2 (FG loop) should have the pattern of G + 30-base random region + 11-base scaffold sequence. Any sequences not matching this pattern were filtered out, allowing one miss match with the constant scaffold sequence. Constant scaffold sequences were trimmed out after filtering, leaving the two paired random regions for downstream analysis. Identical paired reads were combined and counted, followed by enrichment analysis.

Binder reconstruction and validation. Using semi-enriched selected pool cDNA as the template, putative binders were amplified in three PCR reactions. The first PCR reaction utilized primers that cover the entire random loop regions in order to accurately generate clone sequences. Primers were both annealed and extended at 70 °C. The next reaction used 1 μL of the first reaction to extend and amplify the products with FnOligo2C25K and a second FG-specific primer. We used 1 μL of the second reaction in the final reaction, amplifying and extending the

DNA with FnOligo1C25K and primer 3FLAGdA30 to full-length DNA templates with a stop codon and a dA(30) tail at the 3' end of the DNA. The crude PCR product was directly used for coupled transcription/translation using the TNT T7 Quick system (Promega). The reactions were terminated after 90 minutes by the addition of 1/10 volume of 0.5 M EDTA pH 8.0. The protein samples were bound to beads with or without target (streptavidin-coupled agarose beads, Pierce) in TBS plus 0.05% Tween 20 and 1 mg/mL BSA, for one hour at room temperature, followed by washing with the same buffer and elution in SDS-PAGE loading buffer. Binding was detected by western blotting using M2 anti-FLAG antibody directly coupled to HRP (Sigma).

pJD3 vector construction. A fragment containing the pelB leader peptide sequence and the multiple cloning site was generated from extension of two synthetic oligos (pelB-MCSFOR and pelBREV) which was then digested with NdeI and BamHI. The pelB sequence was designed as in Suzuki et al.[8] and is predicted to cleave between alanine and methionine. The *E. coli* alkaline phosphatase gene was amplified from *E. coli* genomic DNA (DH5α) using primers phoAFOR and phoAREV. Two mutations demonstrated to improve activity[9] were employed here. The D153G and D330N mutations were introduced by a PCR mutagenesis strategy. Three fragments were generated using primers phoAFOR plus D153GREV, D153GFOR plus D330NREV, and D330NFOR plus phoAREV. These fragments were digested with BsaI which produced desired in-frame, non-palindromic overhangs for reassembly of the full double mutant by T4DNA ligase. The full gene was amplified further using phoAFOR and phoAREV and digested with BamHI and SalI. The digested leader peptide-MCS and mutant AP fragment were ligated into the NdeI and XhoI site of pAO9[6] thereby replacing the original MCS and affinity fusion tags. Utilizing SalI for ligation into the pAO9 XhoI site eliminates this restriction site and allows use of a XhoI site placed in the new MCS. This enables direct insertion of library clones without additional PCR as XhoI and BamHI are intrinsic to the e10Fn3 library template DNA.

pAO5-btn and pAO9-btn vector constructions. Two synthetic oligos (AO5btnFOR and AO5btnREV) were extended to produce the multiple cloning site and BirA biotinylation sequence and were digested with NdeI and BamHI for ligation into pAO5.[6] For pAO9-btn two synthetic oligos (AO9btnFOR and AO9btnREV) encoding the BirA biotinylation sequence and a factor Xa recognition sequence were extended and digested with BamHI and SacI for ligation into pAO9, thereby replacing the Flag-tag in the original vector.[6] The fXa site can by used to remove the MBP and multihistidine tags. However we have not detected loss of function due to presence of the MBP fusion which may enhance solubility.[10]

Enzyme-linked detection assays. e10Fn3 clones were digested and ligated into pJD3 or pAO9-btn using XhoI and BamHI, which are intrinsic to the library template and therefore required no additional primers. Cloning into pAO5-btn utilized NdeI and KpnI and therefore required incorporation of restriction sites by PCR using primer eFn5NdeI and Fn3KpnI. All proteins were expressed in *E. coli* BL21(DE3) after reaching mid log phase by induction with isopropyl β-D-1-thiogalactopyranoside (IPTG; 0.5 mM final concentration) for three hours at 37 °C. After pelleting, cells were lysed using BPER protein extraction reagent (Pierce) and proteins were purified by nickel affinity chromatography (Qiagen).

Both IgG(Fc) binders were cloned into pJD3 and assayed for function in alkaline phosphatase-based detection assays. Binding over background was determined using 96-well polystyrene plates saturated with IgG(Fc) or BSA. Purified e10Fn3-AP fusions (5 μg/mL in TBS, 0.05% Tween 20, 0.1% BSA) were incubated with blocked wells overnight at room temperature, washed, and developed with *p*-nitrophenyl phosphate (PNPP) (Sigma) (1 mg/mL in 1 M diethylamine, pH 9.8) for one hour. The mock sandwich assay was performed by using streptavidin saturated 96-well polystyrene plates to immobilize various quantities of biotinylated IgG(Fc).

eFn-anti-IgG1 was cloned into pAO9-btn and expressed for luminescence-based enzyme-linked binding assays in comparison to commercially available anti-human IgG(Fc) (Pierce). 96-well plates were saturated with IgG(Fc), whole IgG, IgA (Rockland Immunochemicals) or 293T lysate. In each well, we incubated a solution of either enzymatically-biotinylated (BirA, Avidity) eFn-anti-IgG1 (100 ng/mL) coupled with streptavidin-HRP (1:50,000 dilution; 25 ng/mL) (Pierce) or anti-human IgG(Fc)-HRP (1:40,000) in TBS plus 0.05% Tween 20 and 0.1% BSA was incubated for one hour. The concentrations used for both eFn-SA-HRP and anti-IgG(Fc) were empirically determined to produce the best signal to noise under these conditions (data not shown). After washing, samples were developed using SuperSignal ELISA Pico (Pierce) and measured on an LMax luminometer (Molecular Devices).

Five MBP-binding e10Fn3s were cloned into pAO5-btn and expressed for comparison in luminescence based detection assays. Enzymatically biotinylated e10Fn3s (1 μg/mL) were added together with SA-HRP (25 ng/mL) in TBS plus Tween 20 and BSA to wells saturated with MBP or BSA. After incubating for one hour, wells were washed and samples were developed as described above. eFn-anti-MBP1-SA-HRP binding was compared to monoclonal anti-MBP (1:10,000) (NEB) which was bound for one hour, washed and incubated with anti-mouse-HRP (1:20,000).

eFn-anti-IgG1 binding affinity. To determine the affinity of eFn-anti-IgG1 for IgG, we performed an ELISA-based assay in which monomeric, biotinylated protein was bound prior to conjugation with SA-HRP. Polystyrene wells were coated with 20 ng of IgG overnight and then blocked with 2% BSA. Various concentrations of eFn-anti-IgG1 were incubated for 6 hours at room temperature, then washed three times, and bound with SA-HRP (50 ng/mL) for 30 minutes. After washing three times, samples were developed using SuperSignal ELISA Pico. The K_D was determined by non-linear regression (Prism, Graphpad Software).

Protein G competition assay. Plates were coated with 50 ng of IgG, and eFn-anti-IgG1-SA-HRP was assayed for detection as described above (100 ng/mL e10Fn3 plus 25 ng/mL SA-HRP) except with various quantities of protein G (Pierce) added as competitor.

eFn-anti-MBP SPR. Surface plasmon resonance (Biacore T100) was used to determine binding constants for eFn-anti-MBP1 and eFn-anti-MBP2. Biotinylated e10Fn3 proteins were immobilized onto an SA chip and various concentrations of MBP were flowed at a rate of 50 μL/min for 2 minutes in TBS plus 0.05% Tween 20 at 25 °C. Binding constants were determined using the Biacore evaluation software.

e10Fn3-based western blotting. Protein samples were separated by SDS PAGE and transferred to nitrocellulose, which was blocked with 5% milk in TBS plus 0.1% Tween 20. Blots were probed for one hour with biotinylated eFn-anti-IgG1 (200 ng/mL) plus SA-HRP (25 ng/mL) or eFn-anti-MBP1 (200 ng/mL) plus SA-HRP (50 ng/mL) in TBS plus 0.1% Tween 20 and 5% milk. After washing, blots were developed by enhanced chemiluminescence.

REFERENCES

[1] C. A. Olson, R. W. Roberts, *Protein Sci* **2007**, *16*, 476.
[2] G. Cho, A. D. Keefe, R. Liu, D. S. Wilson, J. W. Szostak, *J Mol Biol* **2000**, *297*, 309.
[3] W. L. DeLano, M. H. Ultsch, A. M. de Vos, J. A. Wells, *Science* **2000**, *287*, 1279.
[4] A. Koide, C. W. Bailey, X. Huang, S. Koide, *J Mol Biol* **1998**, *284*, 1141.
[5] C. A. Olson, J. D. Adams, T. T. Takahashi, H. Qi, S. M. Howell, T. T. Wu, R. W. Roberts, R. Sun, H. T. Soh, *Angew Chem Int Ed Engl*.
[6] C. A. Olson, H.-I. Liao, R. Sun, R. W. Roberts, *ACS Chem. Biol.* **2008**, *3*, 480.
[7] R. Liu, J. E. Barrick, J. W. Szostak, R. W. Roberts, *Methods Enzymol* **2000**, *318*, 268.
[8] C. Suzuki, H. Ueda, E. Suzuki, T. Nagamune, *J Biochem* **1997**, *122*, 322.
[9] B. H. Muller, C. Lamoure, M. H. Le Du, L. Cattolico, E. Lajeunesse, F. Lemaitre, A. Pearson, F. Ducancel, A. Menez, J. C. Boulain, *Chembiochem* **2001**, *2*, 517.
[10] R. B. Kapust, D. S. Waugh, *Protein Sci* **1999**, *8*, 1668.