bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves

Rhys M. Adams[1], Justin B. Kinney[2,*], Thierry Mora[3,*], and Aleksandra M. Walczak[1,*]
[1] *Laboratoire de Physique Théorique, UMR8549, CNRS and École Normale Supérieure, 24, rue Lhomond, 75005 Paris, France*
[2] *Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungown Rd., Cold Spring Harbor, NY, 11724, USA and*
[3] *Laboratoire de Physique Statistique, UMR8550, CNRS and École Normale Supérieure, 24, rue Lhomond, 75005 Paris, France*
[*] *Equal contribution*

Despite the central role that antibodies play in the adaptive immune system and in biotechnology, much remains unknown about the quantitative relationship between an antibody's amino acid sequence and its antigen binding affinity. Here we describe a new experimental approach, called Tite-Seq, that is capable of measuring binding titration curves and corresponding affinities for thousands of variant antibodies in parallel. The measurement of titration curves eliminates the confounding effects of antibody expression and stability inherent to standard deep mutational scanning assays. We demonstrate Tite-Seq on the CDR1H and CDR3H regions of a well-studied scFv antibody. Our data sheds light on the structural basis for antigen binding affinity, and suggests a dominant role for CDR1H in establishing antibody stability. Tite-Seq fills a large gap in the ability to measure critical aspects of the adaptive immune system, and can be readily used for studying sequence-affinity landscapes in other protein systems.

I. INTRODUCTION

During an infection, the immune system must recognize and neutralize invading pathogens. B-cells contribute to immune defense by expressing receptors, called antibodies, that bind specifically to foreign antigens. The astonishing capability of antibodies to recognize virtually any foreign molecule has also been exploited by scientists for a wide variety of applications, and antibodies now play central roles in a variety of experimental techniques (immunofluorescence, western blots, ChIP-Seq, etc.). Antibody-based therapeutic drugs have also been developed for treating a wide variety of diseases [1].

Much is known about the qualitative mechanisms of antibody generation and function [2]. The antigenic specificity of antibodies in humans, mice, and most jawed vertebrates is primarily governed by six complementarity determining regions (CDRs), each roughly 10 amino acids long. Three CDRs (denoted CDR1H, CDR2H, and CDR3H) are located on the antibody heavy chain, and three are on the light chain. During B-cell differentiation, these six sequences are randomized through V(D)J recombination, then selected for functionality and for the ability to avoid recognizing host antigens. Upon participation in an immune response, CDR regions can further undergo somatic hypermutation and selection, yielding higher-affinity antibodies for specific antigens.

Many high-throughput techniques for assaying and selecting antibodies are currently available. Many methods, including phage display [3–5] and yeast display [6, 7] technologies, have been developed for artificially selecting and optimizing antibodies in vitro. More recently, advances in DNA sequencing technology have made it possible to effectively monitor antibody and T-cell receptor diversity within immune repertoires, e.g. in healthy individuals [8–15], in specific tissues [16], in individuals with diseases [17] or following vaccination [18–21].

Many questions remain about basic aspects of the quantitative relationship between antibody sequence and antigen binding affinity, such as the extent of polyspecificity [22], i.e., how many different antibody sequences bind a given antigen with a specified affinity, or the role that epistatic effects play in determining the qualitative geometry of the sequence-affinity landscape. These aspects are important for understanding the effectiveness of somatic hypermutation, and more generally for interpreting repertoire surveys and using them for diagnosis.

In recent years, a variety of "deep mutational scanning" (DMS) assays [23], have been developed for measuring protein sequence-function relationships in a massively parallel manner. The general format of DMS assays provides a promising approach for measuring the affinity of many protein sequences in a single experiment, and thus gaining insight into the sequence-affinity landscape. None of the methods yet described, however, have succeeded in providing quantitative measurements of binding affinity – i.e., dissociation constant measurements in molar units.

To enable massively parallel measurements of binding affinity for antibodies and other proteins, we have developed an assay called "Tite-Seq." Tite-Seq works by building on the capabilities of Sort-Seq, an experimental strategy that was first developed for studying transcriptional regulatory sequences in bacteria [24]. Sort-Seq combines fluorescence-activated cell sorting (FACS) with

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

high-throughput sequencing to provide massively parallel measurements of cellular fluorescence. In the Tite-Seq assay, Sort-Seq is applied to antibodies displayed on the surface of yeast cells and incubated with antigen at a wide range of concentrations. From the resulting sequence data, thousands of antibody-antigen titration curves and their corresponding dissociation constants (here denoted K_D) can be computed.

By assaying full titration curves, Tite-Seq is able to measure affinities over many orders of magnitude. Moreover, the resulting affinity values are not confounded by variation in protein expression or stability, as is the case with standard DMS assays. In fact, Tite-Seq readily provides separate measurements of protein expression level, which can serve as a proxy measurement of theromostability [25]. Finally, unlike other DMS experiments based on phage-display or ribosome-display techniques, the yeast display platform used by Tite-Seq enables simple low-throughput validation measurements [26].

We performed Tite-Seq on a protein library derived from a well-studied short chain variable fragment (scFv) antibody specific to fluorescein [6, 27]. Mutations were restricted to the CDR1H and CDR3H regions, which are known to play an important role in the antigen recognition of this scFv [27, 28] and of antibodies in general [29, 30]. The resulting affinity measurements were validated with titration curves for a handful of clones measured using standard flow cytometry.

Our measurements reveal a large and unexpected difference between the effects of mutations in CDR1H and in CDR3H. Comparing the effects of mutations with the known antibody-fluorescein co-crystal structure [31] also identifies a strong relationship between the effect that a position has on affinity and the number of molecular contacts that the residue at that position makes within the antibody.

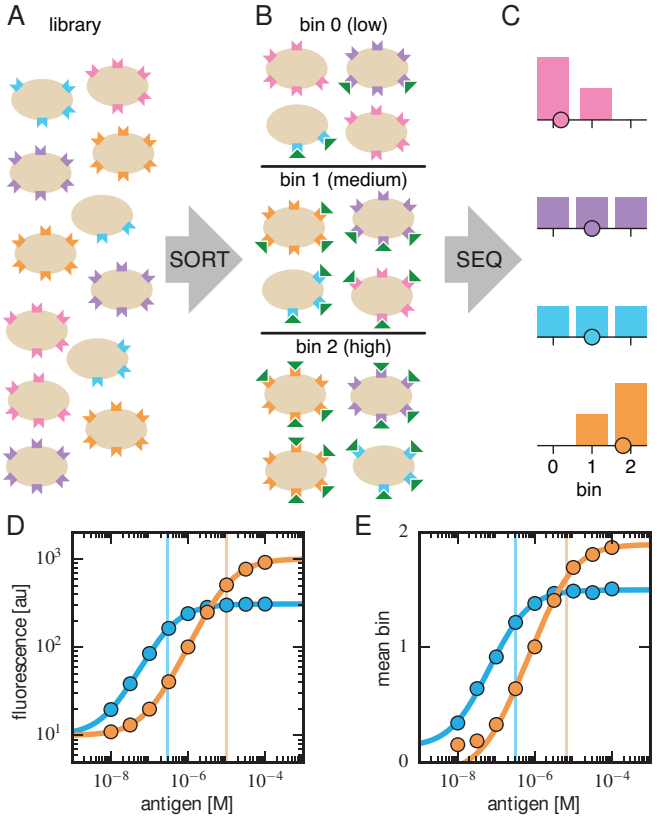

FIG. 1: **Illustration of Tite-Seq.** (A) A library of variant antibodies (various colors) are displayed on the surface of yeast cells (tan). Each cell expresses a single variant, while each variant is expressed by multiple cells. (B) The library is exposed to antigen (green triangles) at a defined concentration, cell-bound antigen is fluorescently labeled, and FACS is used to sort cells into bins according to measured fluorescence. (C) The antibody variants in each bin are sequenced and the distribution of each variant across bins is computed (histograms; colors correspond to specific variants). A summary statistic (dot), such as mean bin number, is used to quantify the average amount of bound antigen per cell. (D) Antibody-antigen K_D values (vertical lines) can be read from titration curves (solid lines) fit to measurements of average cellular fluorescence over a wide range of concentrations (dots). Such curves can be directly measured using flow cytometry, but only in a low-throughput manner. (E) Tite-Seq consists of performing the Sort-Seq experiment in panels A-C at multiple antigen concentrations, then fitting titration curves using the histogram summary statistic (e.g., mean bin number) as a proxy for average cellular fluorescence. This provides a massively parallel way to measure antibody K_D values. Tite-Seq results corresponding to the titration curves in panel E were simulated using 3 sorting bins; see Appendix A for simulation details.

II. RESULTS

A. Overview of Tite-Seq

Our general strategy is illustrated in Fig. 1. First, a library of variant antibodies is displayed on the surface of yeast cells (Fig. 1A). The composition of this library is such that each cell displays a single antibody variant, and each variant is expressed on the surface of multiple cells. Cells are then incubated with the antigen of interest, bound antigen is fluroescently labeled, and fluorescence-activated cell sorting (FACS) is used to sort cells one-by-one into multiple"bins" based on this fluorescent readout (Fig. 1B). Deep sequencing is then used to survey the antibody variants present in each bin. Because each variant antibody is sorted multiple times, it will be associated with a histogram of counts spread across one or more bins (Fig. 1C). The spread in each histogram is due to cell-to-cell variability in antibody expression, and to the inherent noisiness of flow cytometry measurements. Finally, the histogram corresponding to each antibody variant is used to compute an "average bin number" (Fig. 1C, dots), which serves as a proxy measurement for the average amount of bound antigen per cell.

It has previously been shown that K_D values can be

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

accurately measured using yeast-displayed antibodies by taking titration curves, i.e., by measuring the average amount of bound antigen as a function of antigen concentration [7, 26]. The average fluorescence y of labeled cells is expected to be related to antigen concentration via

$$y = B + A\frac{C}{C + K_D} \quad (1)$$

where A is proportional to the number of functional antibodies displayed on the cell surface, B accounts for background fluorescence, and C is the concentration of free antigen in solution. Fig. 1D illustrates the shape of curves having this form. By using flow cytometry to measure y on clonal populations of yeast at different antigen concentrations C, one can fit curves to Eq. 1 and thereby learn K_D. Such measurements, however, can only be performed in a low-throughput manner.

Tite-Seq allows thousands of titration curves to be measured in parallel. The Sort-Seq procedure illustrated in Fig. 1A-C is performed at multiple antigen concentrations, and the resulting average bin number for each variant is plotted against concentration. Curves of the form in Eq. 1 are fit to these proxy measurements, and K_D values are thereby inferred for each variant.

We emphasize that K_D values cannot, in general, be accurately inferred from Sort-Seq experiments performed at a single antigen concentration. Because the relationship between binding and K_D is sigmoidal, the amount of bound antigen provides a useful readout of K_D only when the concentration of antigen used in the labeling procedure is comparable in magnitude to K_D. However, single mutations within a protein binding domain often change K_D by multiple orders of magnitude. Sort-Seq experiments used to measure the sequence-affinity landscape must therefore be carried out over a range of concentrations large enough to encompass this variation.

Furthermore, as illustrated in Fig. 1D, different antibody variants often lead to substantially different levels of functional antibody expression on the yeast cell surface. If one performs Sort-Seq at a single antigen concentration, high affinity (low K_D) variants with low expression (blue variant) may bind less antigen than low affinity (high K_D) variants with high expression (orange variant). Only by measuring full titration curves can the effect that sequence has on affinity be deconvolved from sequence-dependent effects on functional protein expression.

B. Proof-of-principle Tite-Seq experiments

To test the feasibility of Tite-Seq, we used a well-characterized antibody-antigen system: the 4-4-20 single chain variable fragment (scFv) antibody [6], which binds the small molecule fluorescein with $K_D = 0.7 \pm 0.3$ nM [27]. This system was used in early work to establish the capabilities of yeast display, and a high resolution

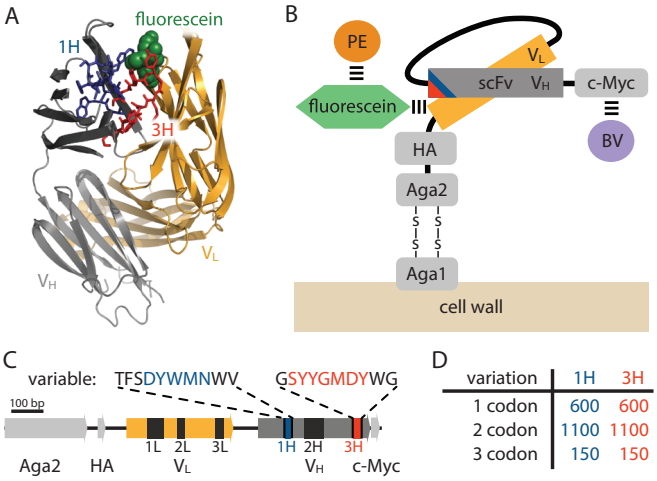

FIG. 2: **Antibody constructs used in this paper.** (A) Co-crystal structure 4-4-20 antibody; PDB code 1FLR [31]. The CDR1H and CDR3H regions are colored blue and red, respectively. (B) The scFv construct, from [6] that was used in this study. Biotinylated fluorescein was used as the antigen, and bound antigen was visualized using streptavidin-RPE (PE). The amount of surface-expressed protein was visualized separately by labeling the c-Myc epitope at the C-terminus of the scFv construct with Brilliant Violet 421 (BV). Approximate location of the CDR1H (blue) and CDR3H (red) regions within the scFv are illustrated. (C) The gene coding for this scFv construct. The six CDR regions indicated. The WT sequence of the two variable regions (10 aa each) are also shown. (D) Composition of the CDR1H and CDR3H scFv libraries.

co-crystal structure of the 4-4-20 antibody bound to fluorescein has been determined [31] (Fig. 2A). An ultra-high-affinity ($K_D = 270$ fM) variant of this scFv, called 4m5.3, has also been found [27]. In what follows, we refer to the 4-4-20 scFv from [6] as WT, and the 4m5.3 variant from [27] as OPT.

The scFv was expressed on the surface of yeast as part of the multi-domain construct described in [6] (Fig. 2B). Following [27], we used fluorescein-biotin as the antigen and labeled scFv-bound antigen with streptavidin-RPE (PE). The amount of surface-expressed protein was separately quantified by labeling the C-terminal c-Myc tag using anti-c-Myc primary antibodies and secondary antibodies conjugated to Brilliant Violet 421 (BV). See Appendix B for details on yeast display and labeling.

Two different scFv libraries were assayed (Fig. 2C): one with a 10 aa variable region encompassing CDR1H (blue), and one with a 10 aa variable region encompassing CDR3H (red). The composition of each library is summarized in Fig. 2D; each library contained the WT sequence, all 600 single-codon variants, a sample of 1100 double codons missense variance, and a sample of 150 triple codon variants. See Methods and Appendix C for details on library generation. As a negative control, we used a non-functional scFv referred to here as Δ.

Tite-Seq was carried out as follows. Approximately

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

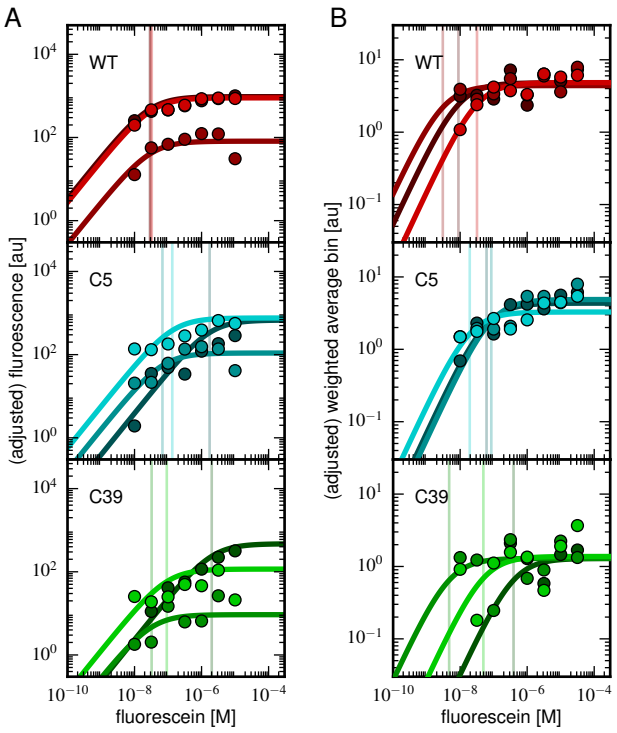

FIG. 3: **Example titration curves.** (A) Flow cytometry measurements and inferred titration curves for the WT scFv and two variant scFvs (C5 and C39). Different shades indicate results from three different replicate experiments. (B) Tite-Seq measurements and fitted curves for the same scFvs as in panel A. Different shades indicate results for different synonymous variants; not all synonymous variants are shown. In both panels, the term "(adjusted)" labeling the ordinate refers to the fact that the inferred background B was subtracted from the data points and from the fitted curves. Vertical lines show K_D values corresponding to each titration curve.

3×10^6 yeast cells expressing scFv variants in either the CDR1H or CDR3H libraries were incubated with fluorescein-biotin at one of the following concentrations: 0 M, $10^{-8.5}$ M, 10^{-8} M, $10^{-7.5}$ M, 10^{-7} M, $10^{-6.5}$ M, 10^{-6} M, $10^{-5.5}$ M, 10^{-5} M, $10^{-4.5}$ M, or 10^{-4} M. After PE labeling of bound antigen, $\sim 10^6$ cells were sorted into four bins using FACS. The bins used for these sorts are shown in Figs. S1A,B; the number of cells sorted into each bin is shown in Fig. S2A. Each bin of cells was regrown and bulk DNA was extracted. Variant CDR1H and CDR3H regions were then PCR amplified and sequenced using paired-end Illumina sequencing (see Appendix D for details). The final data set consisted of $\sim 2 \times 10^6$ sequences per bin (Fig. S2B). For each variant sequence in each bin, a "weighted average bin" value was computed as described in Appendix E.

A K_D value was then inferred for each variant DNA sequence by fitting a function of the form in Eq. 1 to these weighted average bin values. In particular, the background term B for each clone was determined from the 0 M fluorescein measurement, as well as from the average of lowest 10% of fluorescent measurements, assumed to correspond to non-functional constructs (see Appendix F for details). After inspecting the resulting titration curves, we judged that K_D values could not be confidently called below 10^{-9} M and above 10^{-5} M.

During each Tite-Seq experiment, we also measured the effects of sequence variation on scFv expression. For each scFv library, we labeled the c-Myc tag of expressed constructs with BV as described above, then sorted and sequenced cells (Figs. S1C,D and Fig. S2A,B). In what follows, we use E to denote the weighted average bin value for each variant in each library, normalized by the mean of such measurements for WT scFvs in that same library.

C. Low-throughput validation experiments

To judge the accuracy of Tite-Seq, we performed separate titration curve measurements on individual scFv clones. In addition to the WT, OPT, and Δ, we measured three clones from the CDR1H library (named C5, C7, and C39) and three clones from the CDR3H library (C45, C93, C133). Each clone underwent the same labeling procedure as in the Tite-Seq experiment, after which median fluorescence values were measured using standard flow cytometry. K_D values were then inferred by fitting titration curves of the form in Eq. 1; B was determined from data taken at 0 M fluorescein and from measurements on Δ. Three replicate titration curves were taken for each clone.

D. Tite-Seq can measure dissociation constants

Fig. 3A shows flow cytometry measurements along with fitted curves and K_D values for three scFv clones: WT, C5, and C39. Corresponding Tite-Seq data and results shown in Fig. 3B. This figure provides a sense for how precisely titration curves can be measured by each method. Although individual data points can be noisy, fitting curves to multiple data points nevertheless provide reasonably accurate measurements of affinity. The accuracy of these measurements is increased by averaging over replicates.

Fig. 4A compares the K_D values measured by Tite-Seq to those measured by flow cytometry for WT and the five library clones found to have K_D values within the range of detection (10^{-9} M to 10^{-5} M). For the CDR3H clone C45, which is not plotted, Tite-Seq yielded $K_D \geq 10^{-5}$ M whereas flow cytometry gave $K_D = 10^{-5.23 \pm 0.39}$ M. As expected, our three K_D measurements for OPT (CDR1H Tite-Seq, CDR3H Tite-Seq, and flow cytometry) were all at or below the detection boundary of 10^{-9} M.

Error bars on flow cytometry K_D values were computed using the average variance observed in replicate

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

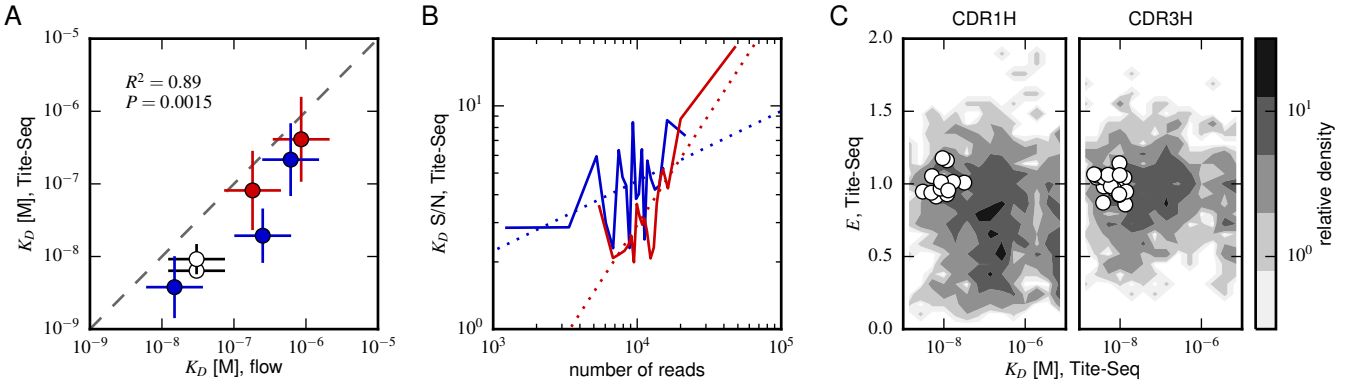

FIG. 4: **Accuracy and precision of Tite-Seq.** (A) Tite-Seq measurements of K_D for the WT scFv (white), three variant CDR1H clones (blue), and two variant CDR3H clones (red). Error bars on flow K_D values are the same for all data points; they show the average mean squared error computed computed using three replicate measurements for each clone. Error bars on Tite-Seq K_D values were estimated using the regression fit in panel B. (B) Estimated signal-to-noise ratios (S/N) of Tite-Seq measurements as a function of the number of sequence reads. Solid lines indicate S/N values computed for CDR1H (blue) and CDR3H (red) variants within twenty equally populated bins along the abscissa. Dotted lines indicate linear regression fits. (C) Density plots of K_D and E measurements for variants in the CDR1H (left) and CDR3H (right) libraries. White dots indicate measurements for synonymous mutants of the WT scFv (18 for CDR1H, 15 for CDR3H). K_D values $\geq 10^{-5}$ M and $\leq 10^{-9}$ M are not shown.

measurements. The computation of error bars on Tite-Seq K_D values was more involved: measurements on synonymous variants were used to estimate the signal-to-noise ratio (S/N) as a function of the number of sequence reads; this dependence of S/N on read number was then used to estimate error bars on Tite-Seq K_D values. The S/N values observed for each of the two libraries are illustrated in Fig. 4B. Details on error estimates are provided in Appendix F.

Fig. 4A reveals a strong correspondence between the Tite-Seq and flow cytometry K_D values. There is a detectable bias, but it is small compared to the 100-fold range of K_D values measured. The robustness of Tite-Seq is further illustrated by the consistency of K_D values measured for the WT scFv in the two different libraries: $K_D = 10^{-8.20\pm0.05}$ M for CDR1H, $K_D = 10^{-8.04\pm0.21}$ M for CDR3H. These measurements are also consistent with the flow cytometry measurement of $K_D = 10^{-7.52\pm0.40}$ M. We note that the Tite-Seq K_D measurements for the WT scFv are about a factor of 10 larger than the previously measured value of $K_D = 0.7 \pm 0.3$ nM [27]. This is likely due to differences in buffers: we used a TBS buffer with more salt (700 mM NaCl) than the PBS buffer used by [27] (150 mM NaCl), a difference that is expected to increase K_D.

The necessity of performing K_D measurements over a wide range of antigen concentrations is illustrated in Fig. S3. At each antigen concentration used in our Tite-Seq experiments, the enrichment of scFvs in the high-PE bins correlated poorly with the K_D values inferred from full titration curves. Moreover, at each antigen concentration used, a detectable correlation between K_D and enrichment was found only for scFvs with K_D values close to that concentration.

Fig. S4 suggests a possible reason for the weak correlation between enrichment in high-PE bins and K_D values. We found that, at saturating concentrations of fluorescein ($2\,\mu$M), cells expressing the OPT scFv bound twice as much fluorescein as cells expressing the WT scFv. This difference was not due to variation in the total amount of displayed scFv, which one might control for by labeling the c-Myc epitope as in [32]. Rather, this difference in binding reflects a difference in the fraction of displayed scFvs that functioned properly. Yeast display experiments performed at a single antigen concentration cannot distinguish such differences in scFv functionality from differences in scFv affinity.

E. Differing effects of mutations in CDR1H and CDR3H

In both the CDR1H and CDR3H libraries, a large fraction of mutations (41% for CDR1H, 58% for CDR3H) increased K_D above our detection limit of 10^{-5} M. A much smaller fraction of variants (2.2% for CDR1H, 5.4% for CDR3H) had K_D values below our measurement threshold of 10^{-9} M. Both of these results suggest that binding affinity is more sensitive to variation in CDR3H than to variation in CDR1H, a finding consistent with the conventional understanding of antibody function [29, 33]. This difference, however, is somewhat modest.

A much more striking difference between CDR1H and CDR3H was observed in how mutations affected expression of scFvs on the yeast cell surface. This difference can be seen in Fig. 4C, which provides a global view of the K_D and E values measured by Tite-Seq. This landscape shows that expression is more sensitive to mu-

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

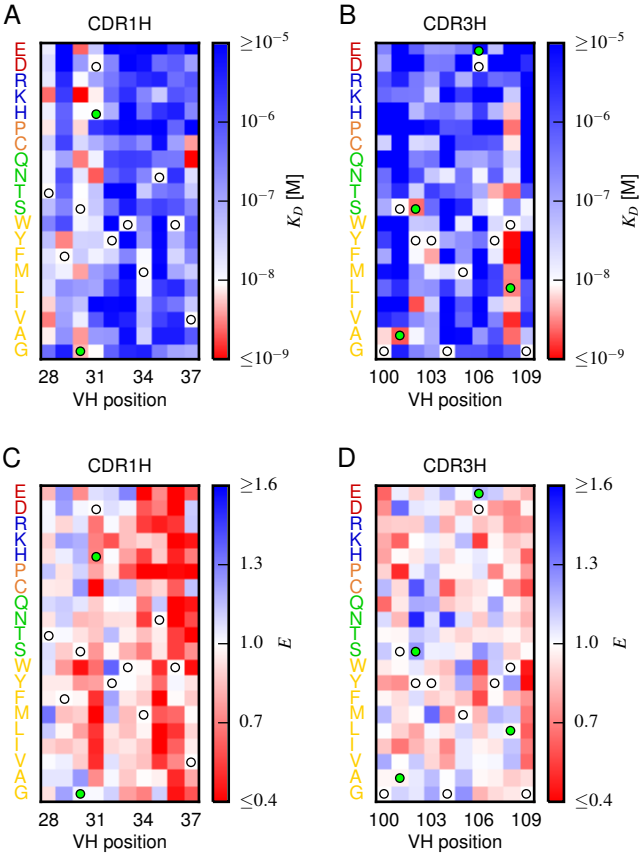

FIG. 5: **Effects of substitution mutations.** Heatmaps show the measured effects on affinity (A,B) and expression (C,D) of all single amino acid substitutions within the variables regions of CDR1H (A,C) and CDR3H (B,D). White dots indicate residues of the WT scFv. Green dots indicate non-WT residues in the OPT scFv.

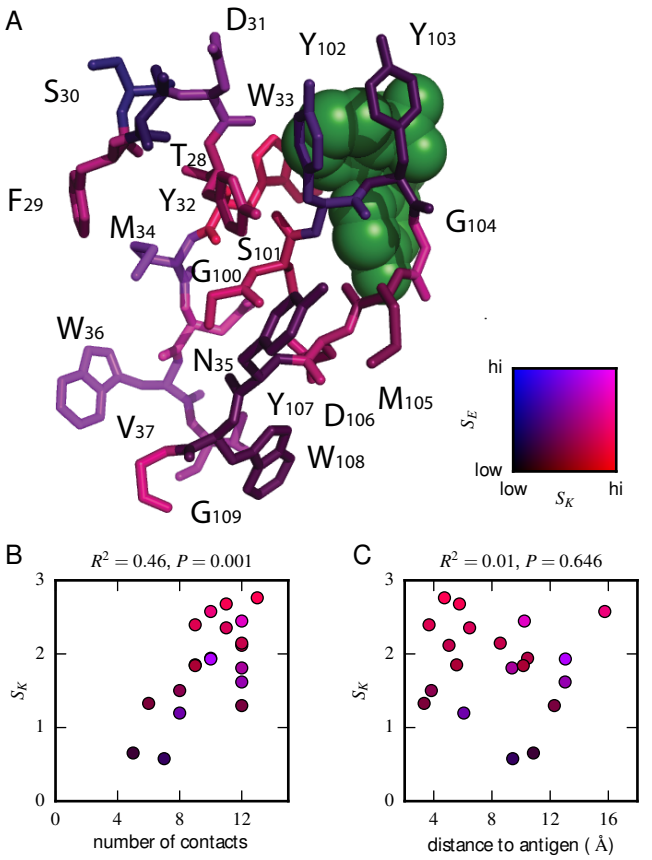

FIG. 6: **Structural context of mutational effects.** (A) Crystal structure [31] of the two variable regions of the WT antibody in complex with fluorescein (green). Each residue (CDR1H: positions 28-37; CDR3H: positions 100-109) is colored according to the S_K and S_E values computed for that position. These variables, S_K and S_E, respectively quantify the sensitivity of K_D and E to amino acid substitutions at specified positions; see Eqs. 2 and 3 for definitions. (B,C) For each position in the CDR1H and CDR3H variable regions, S_K is plotted against either (B) the number of contacts the WT residue makes within the protein structure, or (C) the distance of the WT residue to the fluorescein antigen.

tations in CDR1H than in CDR3H. For instance, both CDR1H and CDR3H showed similar numbers of variants with $E > 1.25$ (5.8% vs. 5.9%, respectively), but CDR1H showed 65% more variants with $E < 0.75$ (56% vs. 34%), and three times more variants with $E < 0.5$ (31% vs. 10%).

These differences between CDR1H and CDR3H can also be seen in Fig. 5, which shows the measured effect of each substitution mutation on affinity and expression. Figs. 5A,B reveal that CDR1H and CDR3H exhibit similar substitution-affinity landscapes. In both regions, the large majority of substitutions increase K_D, and mutations that decrease K_D tend to cluster at a small number of residue positions (CDR1H positions 28, 31, and 37, and CDR3H positions 104 and 108). Figs. 5C and 5D, which show the substitution-expression landscape, differ more substantially. Specifically, substitutions at positions 31 and 34-37 in CDR1H affect expression, on average, much more than do mutations at any of the CDR3H positions.

To further validate Tite-Seq affinity measurements, we examined positions in the high affinity mutant OPT that differ from WT and that lie within the variable regions tested here. As illustrated in Fig. 5A,B, five of the six OPT-specific mutations are neutral or reduce K_D. The remaining mutation, D106E, has previously been shown to require co-mutation with S101A in order to increase affinity [28].

F. Structural correlates of the sequence-affinity landscape

Next we asked whether the sensitivity of affinity and expression values to the identity of the residue at each variable position could be understood from a structural perspective. To quantify this sensitivity at each position

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

i, we computed two quantities:

$$S_K^i = \sqrt{\left\langle \left(\log_{10} K_D^{ia} - \log_{10} K_D^{\mathrm{WT}}\right)^2 \right\rangle_a}, \quad (2)$$

$$S_E^i = \sqrt{\left\langle \left(E^{ia} - E^{\mathrm{WT}}\right)^2 \right\rangle_a}. \quad (3)$$

Here, K_D^{WT} and E^{WT} respectively denote the dissociation constant and expression level measured for the WT scFv, K_D^{ia} and E^{ia} denote analogous quantities for the scFv with a single substitution mutation of amino acid a at position i, and $\langle\cdot\rangle_a$ denotes an average computed over the 19 non-WT amino acids at that position.

Fig. 6A shows the WT CDR1H and CDR3H variable regions in complex with fluorescein. Each residue is colored according to the S_K and S_E values computed for its position. As expected from Figs. 4C and 5C,D, the positions that strongly affect expression are located in CDR1H. To get a better understanding of what aspects of the structure might govern affinity, we plotted S_K values against two other quantities: the number of amino acid contacts made by the WT residue within the antibody structure (Fig. 6B), and the distance between the WT residue and the antigen (Fig. 6C). We found a strong correlation between S_K and the number of contacts, but no significant correlation between S_K and distance to antigen. S_E did not correlate significantly with either of these variables (see Fig. S5).

III. DISCUSSION

We have described a massively parallel assay, called Tite-Seq, for measuring the sequence-affinity landscape of antibodies. Although Tite-Seq was demonstrated here in the context of antibody-antigen binding, this assay can be used to study any receptor-ligand binding system that is compatible with yeast display.

As one application, we found that Tite-Seq measurements can help elucidate the structural basis of protein sequence-function relationships. For the fluorescein-binding scFv antibody studied here, we observed that the number of contacts that an amino acid has with other residues was highly correlated with how strongly mutations to that amino acid affected antigen-binding affinity. By contrast, we found no significant correlation between the distance of an amino acid from the antigen and the magnitude of mutational effects on affinity. This suggests that protein residues primarily affect binding affinity indirectly, via interactions with other residues, rather than through direct interactions with the antigen.

We also found differing effects between CDR1H and CDR3H. As expected, we found that variation in CDR3H had a (modestly) larger effect on affinity than variation in CDR1H. Somewhat unexpectedly, we also found that variation in CDR1H affects protein expression much more strongly than variation in CDR3H. Surface expression in the yeast display system provides a proxy readout of thermostability. Our observation therefore suggests that CDR1H plays a much larger role than CDR3H in stabilizing antibody structure. Together, these two findings suggest a biochemical rational for why CDR3H is more likely than CDR1H to be mutated in functioning receptors [33], why it acquires mutations earlier than other CDRs [33], and why variation in CDR3H is often sufficient to establish antigen specificity [29].

The range of affinities measured in our Tite-Seq experiments include a substantial fraction of the physiological range relevant to affinity maturation ($\sim 10^{-6}$ M to 10^{-10} M) [34–36]. Experimental details made it difficult for us to measure affinities of $\lesssim 10^{-9}$ M, but this limitation might be alleviated with improvements to our labeling protocol. We therefore suggest that future Tite-Seq experiments might be useful for mapping the sequence-affinity trajectories of antibodies during the affinity maturation process.

Many DMS experiments for measuring protein sequence-affinity relationships have been described in recent years [23], including methods that combine Sort-Seq with yeast display [32]. Such approaches apply a selection procedure to a library of variant proteins, then use the enrichment of each protein variant as a proxy measurement of the affinity. In particular, Reich et al. [32] recently showed that such measurements can provide approximate values for the relative rank-order of K_D values (at least when the variable protein is an unstructured peptide).

Tite-Seq differs fundamentally from prior DMS experiments in that full titration curves, not enrichment statistics, are used to measure binding affinities. The measurement of titration curves provides three major advantages. First, this determines absolute K_D values (in units of concentration), not just rank-order values. Second, because ligand binding is a sigmoidal function of affinity, DMS experiments performed at a single ligand concentration will become less sensitive the more receptor K_Ds differ from ligand concentration. Mutations within a protein's binding domain, however, often change K_D by multiple orders of magnitude. Titration curve measurements integrate information over a range of concentrations large enough to measure these effects.

Finally, protein sequence determines not just ligand-binding affinity, but also the amount of surface-displayed protein and the fraction of this protein that is functional. As demonstrated in our experiments, these confounding effects can strongly distort yeast display affinity measurements made at a single antigen concentration. The titration curves measured by Tite-Seq, however, are able to disambiguate these effects. More generally, changing a protein's amino acid sequence can be expected to change multiple biochemical properties of that protein. Our work emphasizes the importance of designing massively parallel assays appropriately in order to disentangle these properties so that measurements of the specific activity of interest can be obtained.

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

IV. METHODS

Variant CDR3H and CDR1H regions were generated using microarray-synthesized oligos (LC Biosciences, Houston TX. USA). These were inserted into the 4-4-20 scFv of [6] using cassette-replacement restriction cloning as in [24], thereby yielding $\gtrsim 10^8$ transformants. EYB100 yeast [6] were transformed with scFv-expressing plasmids using standard methods, yielding providing $\gtrsim 10^5$ transformants. Prior to labeling, OPT- and Δ-displaying yeast were spiked into the CDR3H and CDR1H libraries to provide internal positive and negative controls. Further experimental details are provided in Appendices B-D. The used oligos are reported in Appendix G.

Acknowledgements. We would like to thank Jacklyn Jansen, Amy Keating, Lother Reich, and Bruce Stillman for helpful discussions. We would also like to thank Dane Wittrup for sharing plasmids and yeast strains. RMA, TM and AMW were supported by European Research Council Starting Grant n. 306312. JBK was supported by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

[1] Chan AC, Carter PJ (2010) Therapeutic antibodies for autoimmunity and inflammation. *Nat. Rev. Immunol.* 10:301–316.

[2] Murphy KP, Travers P, Walport M (2008) *Janeway's Immunobiology* (Garland Science), 7 edition.

[3] Smith GP (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228:1315–1317.

[4] Vaughan TJ, et al. (1996) Human antibodies with subnanomolar affinities isolated from a large non-immunized phage display library. *Nat Biotech* 14:309–314.

[5] Schirrmann T, Meyer T, Schütte M, Frenzel A, Hust M (2011) Phage display for the generation of antibodies for proteome research, diagnostics and therapy. *Molecules* 16:412.

[6] Boder ET, Wittrup KD (1997) Yeast surface display for screening combinatorial polypeptide libraries. *Nature Biotechnology* 15:553–557.

[7] Gai SA, Wittrup KD (2007) Yeast surface display for protein engineering and characterization. *Current Opinion in Structural Biology* 17:467–473.

[8] Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science (New York, NY)* 324:807–810.

[9] Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:4099–4107.

[10] Robins HS, et al. (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Science translational medicine* 2:47ra64–47ra64.

[11] Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America* 107:5405–5410.

[12] Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* 109:16161–16166.

[13] Zvyagin IV, et al. (2014) Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proc Natl Acad Sci U S A* 111:5980–5985.

[14] Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM (2014) Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci U S A* 111:9875–9880.

[15] Elhanati Y, et al. (2015) Inferring processes underlying B-cell repertoire diversity. *arXiv:1212.3647 [q-bio.QM]*.

[16] Madi A, et al. (2014) T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* 24:1603–1612.

[17] Parameswaran P, et al. (2013) Convergent antibody signatures in human dengue. *Cell host & microbe* 13:691–700.

[18] Jiang N, et al. (2013) Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Sci. Transl. Med.* 5:171ra19–171ra19.

[19] Vollmers C, Sit RV, Weinstein Ja, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 110:13463–8.

[20] Laserson U, et al. (2014) High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci.* 111:4928–4933.

[21] Galson JD, Pollard AJ, Trück J, Kelly DF (2014) Studying the antibody repertoire after vaccination: Practical applications. *Trends Immunol.* 35:319–331.

[22] Wucherpfennig KW, et al. (2007) Polyspecificity of T cell and B cell receptor recognition. *Seminars in immunology* 19:216–224.

[23] Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods* 11:801–807.

[24] Kinney JB, Murugan A, Callan CG, Cox EC (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. USA* 107:9158–9163.

[25] Shusta EV, Kieke MC, Parke E, Kranz DM, Wittrup KD (1999) Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. *Journal of Molecular Biology* 292:949–956.

[26] VanAntwerp JJ, Wittrup KD (2000) Fine affinity discrimination by yeast surface display and flow cytometry. *Biotechnology progress* 16:31–37.

[27] Boder ET, Midelfort KS, Wittrup KD (2000) Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proceedings of the National Academy of Sciences of the United States of America* 97:10701–10705.

[28] Midelfort KS, et al. (2004) Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *Journal of Molecular Biology* 343:685–701.

[29] Xu JL, Davis MM (2000) Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Im-*

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

munity 13:37–45.

[30] Hoet RM, et al. (2005) Generation of high-affinity human antibodies by combining donor-derived and synthetic complementarity-determining-region diversity. *Nat Biotech* 23:344–348.

[31] Whitlow M, Howard AJ, Wood JF, Voss EW, Hardman KD (1995) 1.85 A structure of anti-fluorescein 4-4-20 Fab. *Protein engineering* 8:749–761.

[32] Reich LL, Dutta S, Keating AE (2015) SORTCERY-A High-Throughput Method to Affinity Rank Peptide Ligands. *Journal of Molecular Biology* 427:2135–2150.

[33] Liberman G, Benichou J, Tsaban L, Glanville J, Louzoun Y (2013) Multi Step Selection in Ig H Chains is Initially Focused on CDR3 and Then on Other CDR Regions. *Frontiers in immunology* 4:274.

[34] Batista FD, Neuberger MS (1998) Affinity dependence of the b cell response to antigen: A threshold, a ceiling, and the importance of off-rate. *Immunity* 8:751–759.

[35] Foote J, Eisen HN (1995) Kinetic and affinity limits on antibodies produced during immune responses. *Proceedings of the National Academy of Sciences of the United States of America* 92:1254–1256.

[36] Roost HP, et al. (1995) Early high-affinity neutralizing anti-viral igg responses without further overall improvements of affinity. *Proceedings of the National Academy of Sciences of the United States of America* 92:1257–1261.

Appendix A: Tite-Seq simulations

For Fig. 1 in the main text, we simulated data at nine antigen concentrations ($C = 10^{-9}$ M, $10^{-8.5}$ M, ..., 10^{-4} M) for two hypothetical scFvs: one with $K_D = 3 \times 10^{-7}$ M, $A = 300$, $B = 10$, and one with $K_D = 10^{-5}$ M, $A = 1000$, $B = 10$. For each clone and at each concentration, fluorescence signals were simulated for 1000 cells by multiplying the y quantity in Eq. 1 by a factor of $\exp(\eta)$ where η is a normally distributed random number. Fig. 1D shows the resulting flow cytometry measurements, computed by taking the median of these simulated fluorescence signals. Curves of the form Eq. 1 were fit to these data by minimizing the square deviation between predicted $\log_{10} y$ values and $\log_{10}$ flow cytometry measurements. Fig. 1E shows Tite-Seq measurements simulated by sorting these 1000 cells into three bins defined by the following fluorescence boundaries: $(0, 30)$ for bin 0, $(30, 300)$ for bin 1, and $(300, \infty)$ for bin 2. The mean bin number was then computed for each clone at each concentration, and curves of the form $\log_{10} y$ were fit to mean bin values using least squares optimization.

Appendix B: Yeast display

1. Innoculation aliquots for yeast display

To ensure consistency in yeast display population sizes, we pre-aliquoted yeast cells with consistent densities. For clones, we picked colonies and grew them in sc -trp 2% glucose liquid media. For libraries, we resuspended 1 ml·OD from our $-80°$ C library stocks into sc -trp 2% glucose liquid media to ensure library diversity. We grew liquid cultures until their OD600 absorption measured between 0.9 and 1.1. We then stored yeast aliquots with 10% glycerol at $-80°$ C. For clones, aliquots were meted out in 0.2 ml·OD aliquots. For libraries, aliquots were meted out in 1.5 ml· OD aliquots.

2. Yeast display induction and labeling

We added cell aliquots stored at $-80°$ C to SC -trp + 2% glucose media until the OD600 equaled 0.05. We kept clonal populations above 0.2 ml·OD, and library populations above 3 ml·OD. After 8 hours, we added 25% volume of YPD, centrifuged the cells at 3000 rpm for 8 minutes, aspirated the media, and resuspended cells at 0.05 OD600 SC -trp +2% galctose media. scFv production was induced in galactose for approximately 16 hours at 22° C. We then added 25% volume YPD, and spun down the cells at 3000 rpm for 8. Cells were washed with 100 μl of ice cold TBS-BSA. TBS-BSA is made from 1 ml 100 mg/ml BSA added to 9 ml TBS (473 ml of water, 25 ml of 1 M Tris, 2.92 g of NaCl adjusted to pH 8 with NaOH). Cells were then centrifuged, and resuspended in labeling solution containing 0 M or $10^{-8.5} - 10^{-4}$ M biotinylated fluorescein, or 1.4 μg/ml c-Myc rabbit antibody, per 0.2 ml·OD of cells. Fluorescein volumes were kept at high enough levels to ensure ten times as many fluorescein molecules as scFv molecules, assuming 100,000 scFv per cell. Cells were shaken at 700 rpm for 1 hour at room temperature during labeling. Cells were then centrifuged, and the supernatant was drained. Cells were then twice washed with 1.5 ml ice cold TBS-BSA. Cells were centrifuged at 3,000 rpm for 8 minutes and the supernatant was aspirated. The cells were then suspended in 112.5 μl/(ml·OD cells) of a secondary labeling solution having 4 μg/ml streptavidin R-PE or 0.8 μg/ml BV421 anti-rabbit donkey antibody. Secondary labeling occurred for 30 minutes at 4C at 300 rpm. Cells were then centrifuged at 3,000 rpm for 8 minutes, resuspended in ice cold TBS, and prepared for FACS or flow cytometry analysis.

3. Flow cytometry and FACS gating strategies

We measured titration curves by either flow cytometry or FACS sorting for Tite-Seq using similar gating strategies. For flow cytometry, cells were first filtered by their forward scatter channels (fsc) and side scatter channels (ssc). scFv expression was then measured by average BV421 channel values corresponding to c-Myc epitopes on cell surfaces, excluding extreme values. For K_D measurements, we further gated cells by their fluorescein and PE signals. We removed cells with high fluorescein fluorescence, since fluorescein's fluorescence signal is quenched by antibody binding, meaning that non-quenched signal indicates non-specific fluores-

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cein binding. Concurrently, we also removed extreme PE values. We then measured titration curves from average PE channel values, corresponding to fluorescein bound to scFv antibodies. For Tite-Seq, we first filtered cells by fsc and ssc channels. For expression measurements, we sorted cells into one of four gates along the BV421 channel. For Tite-Seq titration curves, we furthered filtered cells by their fluorescein and PE signals. Cells were then sorted into one of four gates by their PE signal values.

Appendix C: Generation of scFv libraries

We obtained custom microarray oligonucleotides from LC Sciences to generate a library of mutant CDR1H (μ1H) and CDR3H (μ3H) domains. We created separate libraries for the CDR 1H and 3H regions by performing PCR on these oligonucleotide libraries. CDR3H mutants in the LC sciences microarray library were amplified by PCR using oRAL10 and oRAR10 primers to create μ3H sequences. CDR1 library sequences (μ1H) were created by amplifying the LC Sciences microarray library by PCR using oRAL11 and oRAR11 primers.

Non-mutated single CDR regions were created by performing PCR on the pJK36 plasmid (Fig. S6A) with 1H_2F and 1H_1R primers to create the 1H primer, and 3H_2R and 3H_1F primers to create the 3H primer. By mixing non-mutated (1H or 3H) primers with library primers (μ1H or μ3H) and performing PCR on the pJK36 plasmid we created 360 bp long mutagenized scFv insertion sequences, iRA10, with one or both CDR regions mutated. The iRA10 sequence with added Bsal cut sites were made into iRA11 sequences libraries by performing PCR with oRA10 and oRA11 primers. The iRA11 sequences have complementary sticky DNA to the plasmid backbone pRA10 (Fig. S7A) easing the integration of the sequence into the plasmid.

Having created library sequences, iRA11, that have sticky single stranded DNA at their ends after BsaI digest, we created a target plasmid for the iRA11 library sequences. The target plasmid pRA10 (Fig. S7A) was designed to have the same features as the pJK36 plasmid, where the CDR1H and 3H scFv domains were replaced with a cassette containing the toxic ccdB gene the chloramphenicol marker, and surrounded by BsmBI digest sites. pRA10, was created from Gibson cloning iRA12, iRA13, and iRA14 sequences together. iRA12 contained the BsmBI digest sites, ccdB toxic gene, and chloramphinicol drug resistance markers. iRA12 was created by PCR from pJK14 (Fig. S6B) from [24], which contains the ccdB cassette, with oRA12 and oRA13. iRA13 was created by performing PCR on pJK36 with oRA15 and oRA16 (Fig. S6A). iRA14 was created by performing PCR on pJK36 with oRA14 with oRA17 (Fig. S6A). We transformed DB3.1 E. coli, which are ccdB resistant, with the assembled pRA10 plasmid (Fig. S7A), grew them in ampicillin and chloramphinicol, and then extracted the pRA10 plasmids from *E. coli* using the Qiagen plasmid mini-prep kit.

To introduce our mutagenized CDR1H and/or CDR3H sequences back into an scFv expressing plasmid, we replaced the ccdB gene in pRA10 with the iRA11 insertions containing our library mutations. To do this we digested both the iRA11 library sequences and the pRA10 plasmid to create sticky DNA sequences that could be easily ligated together (Fig. S7 B). We digested the pRA10 backbone with BsmBI and iRA11 with BsaI. Digesting pRA10 with BsmBI had the additional effect of excising the ccdB and chloramphenicol resistance genes from the plasmid. The BsmBI and BsaI digests were purified using Qiagen's QIAquick PCR purification kit and mixed together at molar ratios of 1:2.5, respectively. The BsmBI and BsaI digest mixtures were ligated with T4 DNA ligase and then dialyzed to remove salts. Electrocompetant DH10B *E. coli* were electroporated with 50 μl of ligated DNA. *E. coli* were grown for 1 hour in SOB media, then ampicillin was added and *E. coli* were grown overnight. Because ccdB is toxic to DH10B cells, only plasmids that were successfully cut and reliagated were able to grow. We used Qiagen's plasmid mini-prep kit to extract amplified pRA11 libraries. We then chemically transformed EBY100 *S. cerevisiae* cells for each of the pRA11 libraries (Fig. S7B). Finally, we induced scFv expression in the transformed *S. cerevisiae* libraries. We FACS filtered cells which displayed both HA and c-Myc domains in roughly equal amounts to eliminate *S. cerevisiae* with indels in the scFv domain of their plasmids. From this filtered library we isolated and sequenced 144 clones. We then chose 6 clones to confirm Tite-Seq measurements.

Appendix D: Tite-Seq implementation

Each library was spiked with 0.625% pJK37 and Δ controls and then subjected to 12 FACS rounds. The first FACS separated cells based on receptor expression approximately evenly into different bins based on average c-Myc labeling. The other 11 FACS separated cells were grown in different fluorescein concentrations approximately evenly based on the amount of bound fluorescein. The affinity FACS gates were drawn to have roughly log-linear mean values. Cells were sorted into 1 ml 2X YPAD media contained in a rounded 5 ml polypropylene tube. Cells were regrown overnight in sc -trp 2% glucose. 25 ml·OD of cell populations were spun down and resuspended in 200 μl 0.5 mm glass beads, 200 μl of Phenol/chloroform/isoamyl alcohol and 200 μl of yeast lysis buffer (982 ml water, 10 ml 1 M NaCl, 2 ml Triton X-100, 1 ml 1 M Tris pH 8.0, 0.1 ml 1 M EDTA, 5 ml 20%SDS). The cell/phenol/glass bead/yeast lysis mixture was vortexed for 30 minutes. 200 μl water was added. We then spun the tubes down five times, removing the aqueous layer and adding different media between each centrifugation of cells in the following order: 2 x 200 μl of Phenol/chloroform/isoamyl alcohol, 2 x 200 μl

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of chloroform/isoamyl alcohol, 1 ml ice cold 100% EtOH. We then spun the tubes down, aspirated the EtOH, and added 250 μl 70% ice cold EtOH. We again spun the tubes down, aspirated the EtOH, and resuspended purified DNA in 100 μl of IDTE.

We then performed PCR on our phenol purified DNA with bar coded primers L2AF_XX and L2AF_XX (Fig. S7B). Here the XX refers to 1 of 48 possible barcode combinations (see section 4 of Appendix G). We used variable length, 7-10 bp, bar-codes to increase sequence variability during DNA sequencing. We then pooled these sequences and performed a second round of PCR with Illumina adapter primers PE1v3ext and PE2v3. PCR products from the single library CDR1H and CDR3H FACS + 2.5% molarity random PhiX DNA submitted separately for Illumina Hi-Seq analysis.

We assigned sequence origins according to the 7-10 bp barcodes. Correspondingly to the barcode length we used the 8-11th to 14th bp to confirm the orientation of the sequence (i.e. containing CDR1H or CDR3H region). We categorized each read solely as a function of its CDR1H and CDR3H region. We excluded sequences that we did not design, and those with mismatching bar-codes.

Appendix E: Signal reconstruction from sequence data

Here we describe how to reconstruct the fluorescence signal of a large number of variants from sequence data following FACS sorting.

After processing, the data can be summarized by the number of times a given scFv variant s_i was matched to a sequence read in a given FACS bin $j = 0, \ldots, 3$. We call this number $M_j(s_i)$. Each sequence s_i was assigned a probability distribution $p_j(s_i)$ over bin index j, represented by histograms in Fig. 1D, and defined as

$$p_j(s_i) = \frac{k_j M_j(s_i)}{\sum_{j'} k_{j'} M_{j'}(s_i)}. \tag{E1}$$

During PCR, sequence counts differed from the number of cells sorted. To correct for this we scaled the number of sequences by k_j to match the number of sorted cells. The k_j scaling factor is the ratio of the number of cells sorted into a bin, N_j, to the total number of reads from that bin,

$$k_j = \frac{N_j}{\sum_i M_j(s_i)}. \tag{E2}$$

The fluorescence of each sequence (represented by circles in Fig. 1C) was then estimated as

$$F(s_i) = \sum_{j=1}^{4} w_j p_j(s_i), \tag{E3}$$

where w_j is the estimated mean fluorescence in bin j, taken to be $\propto V^j$, where V is the fold difference between the upper and lower fluorescence bounds deliminting each bin. This method of averaging was used instead of the median or the geometric mean, because the fluorescence was typically bimodal and these alternatives would have given too much weight to the lowest peak of nonfluorescent cells.

This procedure was applied to the FACS-sorted sequence data sorted using antigen labeling, for each concentration $C = 0$ M, $10^{-8.5}$ M, 10^{-8} M, …, 10^{-4} M of the antigen. The reconstructed Tite-Seq signal, calculated using Eq. E3 for each C, is denoted by $y_{\text{Tite-Seq}}(s_i, C)$. We applied the same procedure to the data sorted using scFv labeling, yielding the expression signal $E_{\text{Tite-Seq}}(s_i)$.

To estimate the error made on the Tite-Seq measurement $F(s_i)$, we used variations across synonymous mutants. We grouped the sequences s_i into groups of equal size, ranked according to their read count, $M(s_i) = \sum_j M_j(s_i)$, and indexed each group by the their mean read count M. In each group, we calculated the mean square difference $\sigma^2_{\text{emp}}(M) = \langle (F(s_i) - \bar{F}(a_i))^2 \rangle_M$, where the average $\langle \cdot \rangle_M$ was taken over all sequences in the group indexed by M, and $\bar{F}(a_i)$ is the average signal over synonymous mutants with the same amino-acid sequence a_i. Because $\sigma_{\text{emp}}(M)$ was still a noisy function of M, we approximated it by a power law,

$$\sigma_{\text{emp}}(M) \approx \sigma_{\text{model}}(M) = \alpha M^{\beta}, \tag{E4}$$

where α and β were fitted using mean square error minimization in logarithmic space. In the case of antigen-labeled data, we pulled data from all antigen concentration C together to obtain the empirical error $\sigma_{\text{emp}}(M)$.

Appendix F: Inferring K_D from titration curves

The antigen fluorescence was assumed to follow a non-cooperative Hill function:

$$y(s_i, C) = B(C) + A(s_i) \frac{C}{C + K_D(s_i)}, \tag{F1}$$

where $y(s_i, C)$ is the antigen fluorescence, C is the antigen concentration, $B(C)$ is the background autofluorescence, A a scaling factor, and $K_D(s_i)$ the dissociation constant of scFv s_i. The fluorescence $y(s_i, C)$ can be estimated directly from the mean of flow cytometry fluorescence data in a single clone, or using $y_{\text{Tite-Seq}}(s_i, C)$ in the Tite-Seq experiment as explained above.

The autofluorescence was estimated as $B(C) = y_0(C) - y_0(0) + y(s_i, 0)$, where $y_0(C)$ is the fluorescence of clones with no affinity for the antigen, at concentration C. Note it may depend on C because of non-specific binding. For flow-cytometry measurements, we measured y_0 directly as the mean fluorescence of a strain expressing a non-productive antibody. For Tite-Seq measurements, we estimated it as the average of the lowest 10% of $y_{\text{Tite-Seq}}(s_i, C)$ of all sequences in the library, reasoning

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that these 10% are nonfunctional variants. The sequence-dependent offset $y(s_i,0) - y_0(0)$ was added to enforce $y(s_i,0) = B(0)$.

The dissociation constant was inferred by minimizing the following objective function with respect to K_D and C for each sequence s_i:

$$O(K_D, A; s_i) = \sum_C \frac{\left[\ln y(s_i,C) - \ln\left(B(C) + A\frac{C}{C+K_D}\right)\right]^2}{\sigma(s_i,C)^2} \tag{F2}$$

where the sum on C runs over values of the fluorescein concentrations used in the experiments. We set $\sigma(s_i,C) = 1$ in the case of flow cytometry measurements, and to $\sigma(s_i,C) = \sigma_{\text{model}}(M(s_i,C))$ in the case of Tite-Seq measurements, where $\sigma_{\text{model}}(M)$ is given by Eq. E4, and $M(s_i,C)$ is the read count of sequence s_i at antigen concentration C.

The parameters were found by scanning combinations of $K_D \in [10^{-10}M, 10^{-3}M]$, and $A \in [A_0, 10A_0]$ for Tite-Seq or $A \in [A_0, 100A_0]$, where A_0 was set to enforce positive values of A. Tite-Seq fits generally had lower ranges of fluorescence than flow cytometry measurements due to the binning process, justifying the different ranges of Tite-Seq fits. Some antigen concentrations were not used in the fit because of abnormal fluorescence distributions. These were $C = 10^{-8.5}$ M and $C = 10^{-4}$ M for Tite-Seq, and $C = 10^{-8.5}$ M, $C = 10^{-4.5}$ M and $C = 10^{-4}$ M for flow-cytometry measurements.

To estimate the error on K_D, we used a similar procedure to the one explained in the previous section for F. We grouped sequences into bins according to their total read counts over all fluorescein concentrations and FACS bins, $M_T(s_i) = \sum_C \sum_j M_j(s_i,C)$, and calculated the error as $\sigma^2_{K_D}(M_T) = \langle (K_D(s_i) - \bar{K}_D(a_i))^2 \rangle_{M_T}$, where $\bar{K}_D(a_i)$ is the average K_D over synonymous mutants with amino-acid sequence a_i. Since this function was noisy, we fitted it to a power-law analogously to Eq. E4. The signal-to-noise ratio was then calculated as $\text{Var}(K_D(s_i))/\sigma^2_{K_D}(M_T)$, where the variance was taken over all variants in the library.

Appendix G: Oligos used

1. library oligos

oRAL10: TTCTGAGGAGACGGTGACTGAGGTTCCTTG
oRAR10: TGAAGACATGGGTATCTATTACTGTACG
oRAL11: CAGTCCTTTCTCTGGAGACTGGCG
oRAR11: ATGAAACTCTCCTGTGTTGCCTCTGGATTC
Oligo 1H: GTGTTGCCTCTGGATTC ACTTTTAGTGACTACTGGATGAACTGGGTC CGCCAGTCTCCAGA AGGAGAGTTTCAT
Oligo 3H: GTGACTGAGGTTCCTTG ACCCCAGTAGTCCATACCATAGTAAGAACC CGTACAGTAATAGATACCCAT
Blue denotes CDR1 mutations, red denotes CDR3 mutations.

2. Non-mutated library oligos

3H1F: TTCTGAGGAGACGGTGACT
3H2R: TGAAGACATGGGTATCTATTACTGTAC
1H2F: CAGTCCTTTCTCTGGAGACTG
1H1R: ATGAAACTCTCCTGTGTTGCCT

3. Plasmid oligos

oRA10:GCATATCTAAGGTCTCGTTCTGAGGAGACGGTGAC
oRA11:GCCGATTGTTGGTCTCCATGAAACTCTCCTGTGTTGC
oRA12:GAAATAAGCTTTTGTTCTGGAGACGGCAATTGCTAG
oRA13:AACCTGGGAGGCCCATGAAGAGACGTTCCACGC
oRA14:AGACAAGCTGTGACCCGAAAGGGCCTCGTGATA
oRA15:CACGAGGCCCTTTCGGGTCACAGCTTGTCTGTA
oRA16:CTAGCAATTGCCGTCTCCAGAACAAAAGCTTATTTCTGAA
oRA17:CTAGCGTGGAACGTCTCTTCATGGGCCTCCCAGG

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4. Hi-throughput sequencing oligos

Illumina PE1v3ext: AATGATACGGCGACCACCGAGATCT ACACTCTTTCCCTACACGACG
Illumina PE2v3: AAGCAGAAGACGGCATACGAGATCGGT CTCGGCATTCCTGCT
L1AF:ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNAGTCTTCTTCAGAAATAAGC
L1AR:CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNGCTTGGTGCAACCTG

NNNN denotes a variable length nucleotide barcode. Forward barcodes were paired identically to reverse barcodes. The barcodes are: 'TAAGTGGCGC', 'CCATGCCAC', 'AAGCATCA', 'TCTGAGC', 'AACTAAGCCT', 'CGGATTCAG', 'CCAGCGCT', 'GACGTCA', 'ATACAACAGC', 'CGACTAGGT', 'TACGATCG', 'CCTAATG', 'ACATATCGAC', 'GTCGATGTG', 'TGACGGTC', 'ATGGTAC', 'GTACATCCTT', 'ACTTGGTTG', 'ATTGCAGG', 'CGCCATC', 'AGACGTATCC', 'GAAGGTCCG', 'TCATGTGA', 'CAATTGG', 'TTGGCGGCAT', 'TCCATCCTG', 'TGGCGCAT', 'ACCTTCG', 'TATTGCAATC', 'CACCACGAT', 'TTGTAGGC', 'AACTGAC', 'CGTCGCTACC', 'GAGTCTTGG', 'GATGCCTC', 'TGAACCT', 'AAGGTCTCGT', 'GGTAGTTCC', 'CGATAATC', 'CTCATGT', 'CTCTTCCAAG', 'TGTGAATAC', 'TCTCGCCA', 'CCTCGGA', 'ACGCCTGCTC', 'GCTCTGGAC', 'TAACTTGC', 'GCTTGAT', 'ACTGTATGTC', 'GCGACAGAT', 'ACGGAGTG', 'ACGTATA', 'CCAGACGCGT', 'CTTGACAGA', 'GAACAGGT', 'CGCTGCA', 'GGCTTGGAGT', 'AGTTCTACA', 'CATGCGAG', 'TGTCCAA', 'GCGAAGACCT', 'GGCGCTTAA', 'GCAGTTAG', and 'TCAGTAT'

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

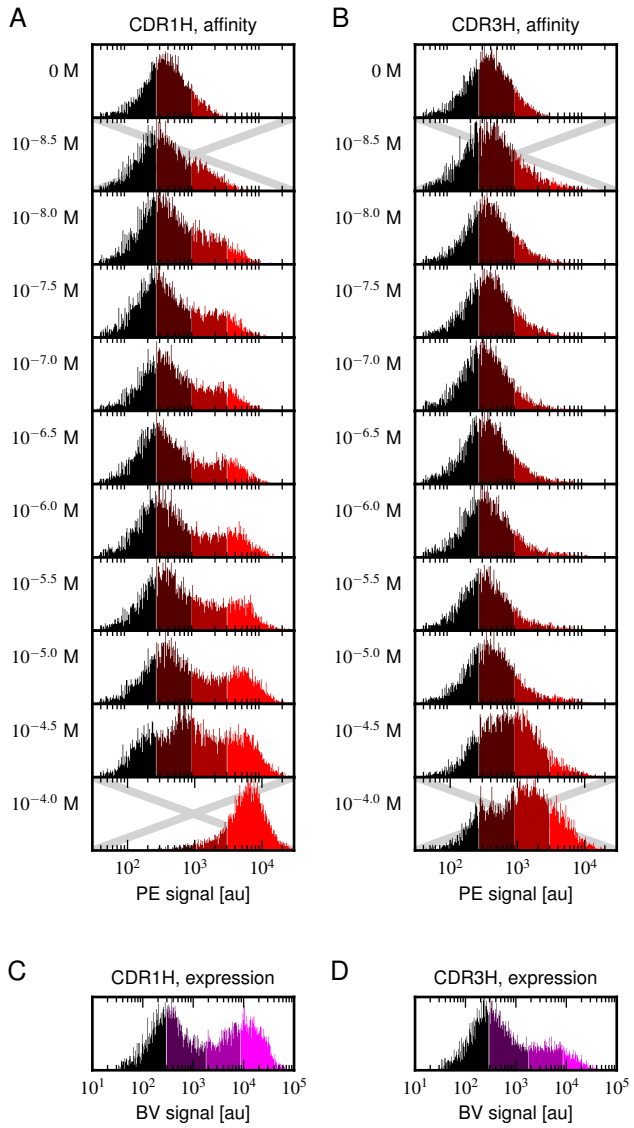

FIG. S1: **FACS gates used for Tite-Seq.** (A,B) Fluorescence gates used to sort (A) CDR1H and (B) CDR3H libraries based on the PE fluorescence readout of bound antigen. Colors indicate the fluorescences gates used for bins 0, 1, 2, and 3 (arrayed from left to right). Each sample was sorted for $\sim$ 20 min; the number of cells sorted into each bin is shown in Fig. S2A. An "X" in the background indicates that Tite-Seq data at that fluorescein concentration was ignored during the inference of titration curves. These data were ignored due to irregularities in the florescence distributions of one or both of the scFv libraries. Such irregularities were likely caused by variation in the labeling procedure. (C,D) Fluorescence gates used to sort (C) CDR1H and (D) CDR3H libraries based on the BV fluorescence readout of expression.

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

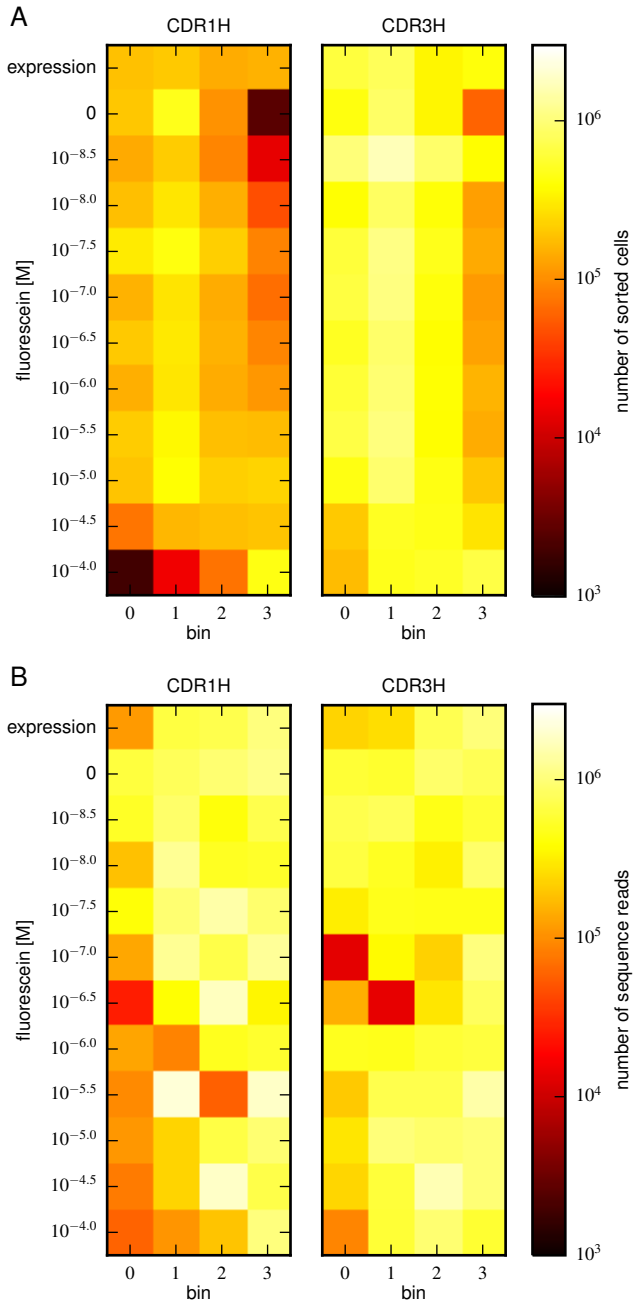

FIG. S2: **FACS counts and read counts.** (A) The number of PE-labeled cells sorted into each bin at each fluorescein concentration during Tite-Seq measurements, together with the number of BV-labeled cells sorted into each bin during Sort-Seq expression measurements. (B) The number of filtered Illumina sequence reads obtained for each bin of sorted cells shown in panel A.

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

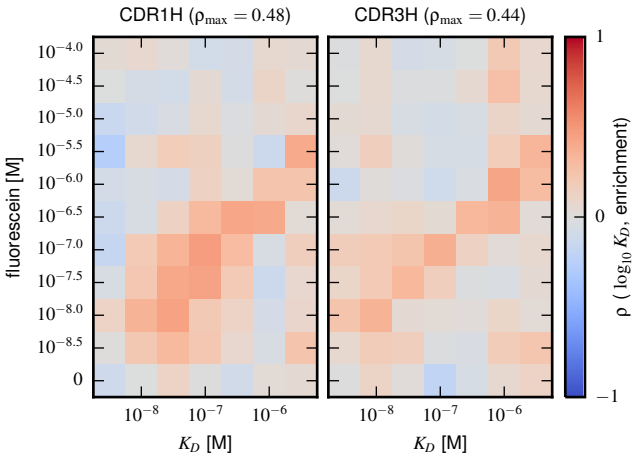

FIG. S3: **Enrichment correlates poorly with K_D.** To assess how well simple enrichment calculations might reproduce the K_D values measured by Tite-Seq, we did the following calculation. For each of the two libraries (CDR1H and CDR3H), we partitioned scFvs into seven groups based on their measured K_Ds (columns). For each group at each antigen concentration (rows), we then computed the enrichment of each scFv in the high PE bins (bins 2,3) relative to the low PE bins (bins 0,1). In these enrichment calculations, the number of counts in each bin was re-weighted to accurately reflect the fraction of library cells falling within the fluorescence range of that bin. This figure shows the resulting Spearman rank correlation (ρ) between enrichment and log K_D values computed for each scFv group at each antigen concentration. In both libraries, we see that correlation values above background (which can be assessed from the values in the 0 M fluorescein row) only occur close to the diagonal, i.e., when K_D is close to the fluorescein concentration used. Even then, the maximum correlation value ($\rho_{\max}$) observed for each library remains below 0.5, indicating the general inability to determine K_D from enrichment statistics.

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

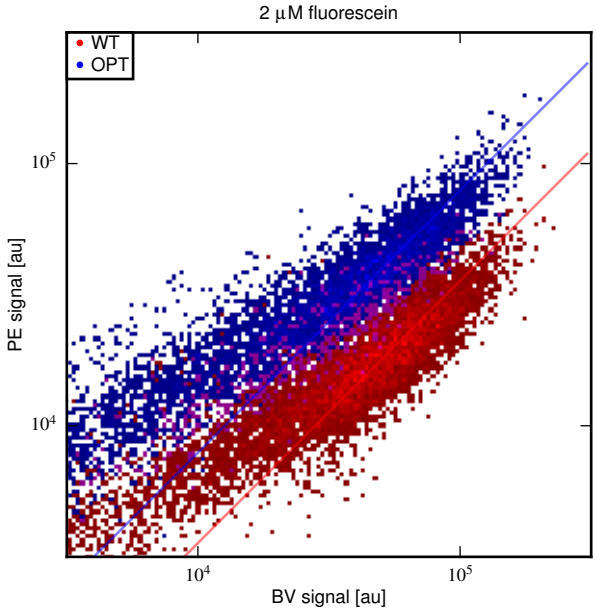

FIG. S4: **Different fractions of displayed OPT and WT scFvs are functional.** 2D flow cytometry histograms showing both OPT- and WT-expressing cells labeled with PE and BV after incubation at 2 μM fluorescein. At this fluorescein concentration, nearly all functional WT and OPT scFvs are bound. Regression lines (fixed to have slope 1) were fit to data points with BV signal between $10^{4.5}$ and 10^5. The vertical shift of the OPT data relative to the WT data indicates a factor of 2.03 ± 0.07 difference (computed from four replicate experiments) in the amount labeled antigen. This difference is not due to a difference in the number of surface-displayed scFvs, as this would cause the OPT and WT clouds to lie along the same diagonal. Rather, this difference between WT and OPT is due to variation in the fraction of surface-displayed scFvs that are functional.

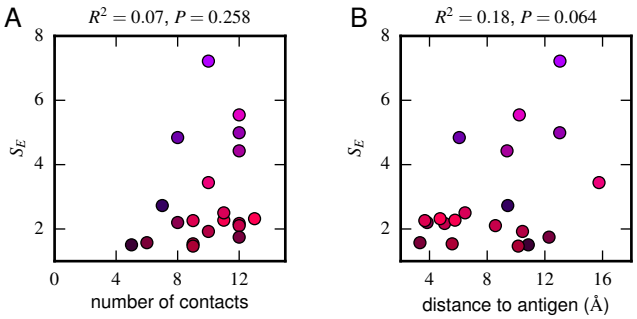

FIG. S5: **The effect of amino acid variation on expression does not have a clear structural basis.** For each amino acid position in the CDR1H and CDR3H variable regions, S_E is plotted against either (A) the number of contacts or (B) the distance to the fluorescein molecule. Points are colored as in Fig. 6. No significant correlation is observed in either plot.

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

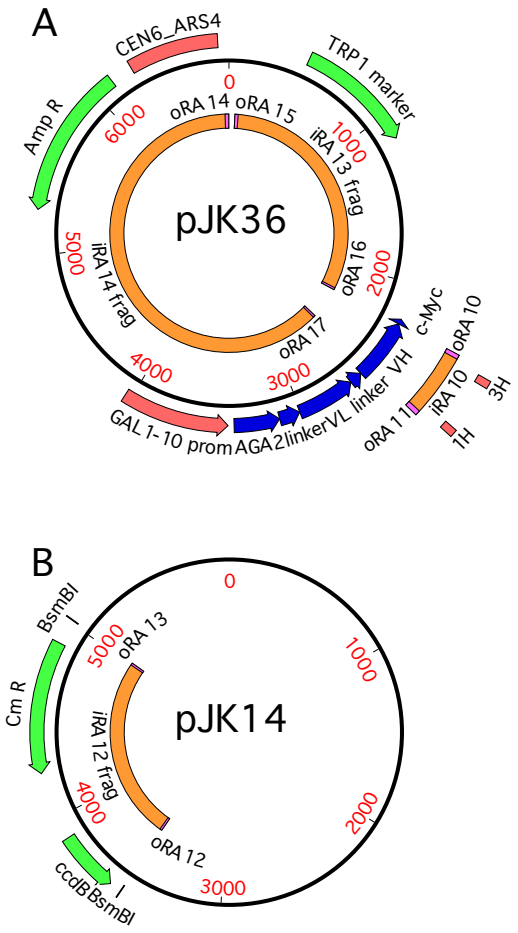

FIG. S6: (A) We performed PCR on pJK36 using a microarray library of oligonucleotides to generate iRA10 libraries of CDR1, 3H variants. We performed PCR on pJK36 to generate the iRA13 and iRA14 sequences. (B) We performed PCR on pJK14 to generate the iRA12 insertion sequence. Cloning fragments excluding extension sequences are shown in orange, primer binding sites in purple, genes in green, non-coding elements in red, and scFv domains in blue. Plasmid pJK36 is identical to plasmid pCT302 from, and plasmid pJK37 is identical to plasmid pCT-4M5.3; both of these were reported in [27] and were kindly provided by Dane Wittrup. The pJK14 plasmid was reported in [24].

bioRxiv preprint doi: https://doi.org/10.1101/036335; this version posted January 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

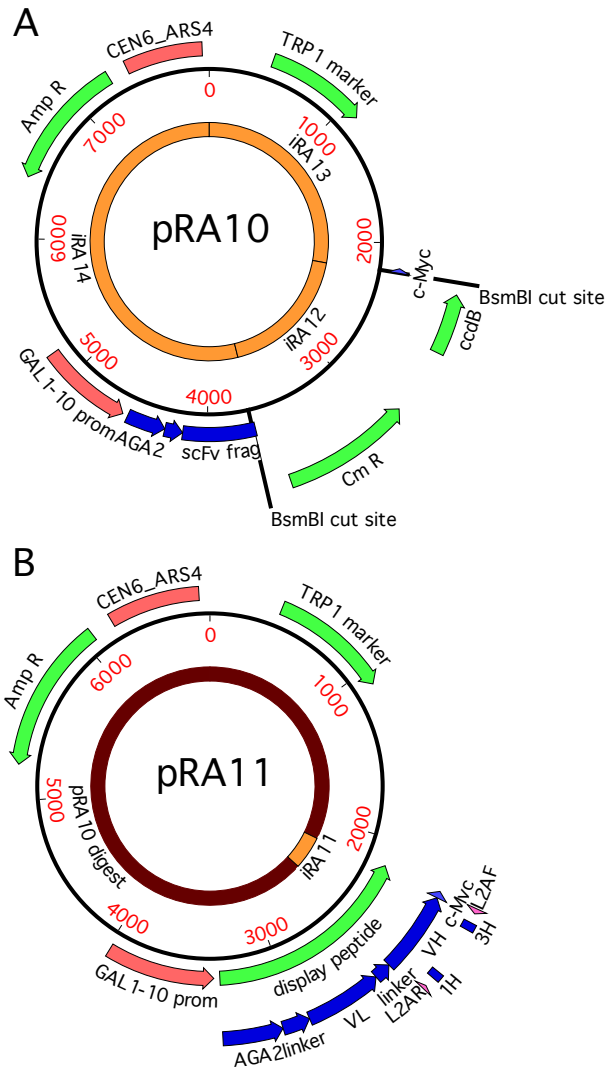

FIG. S7: (A) We digested the pRA10 backbone with BsmBI and then ligated the fragment with BsaI digested iRA11 fragments to create (B) pRA11 plasmid libraries with mutated CDR1H or 3H domains. Cloning fragments excluding extension sequences are shown in orange, primer binding sites in purple, genes in green, non-coding elements in red, plasmid digest in brown, and scFv domains in blue.