Identification of CDRH3 Loops in the Naive B Cell Receptor Repertoire that Can Be Engaged by Candidate Immunogens

Olivia Swanson[1*], Joshua M. Beem[1*], Brianna Rhodes[1], Avivah Wang[1], Maggie Barr, Haiyan Chen[1], Robert Parks, Kevin O. Saunders[1,3], Barton F. Haynes[1,2,4], Kevin Wiehe[1,2#], Mihai L. Azoitei[1,2#]

[1]Duke Human Vaccine Institute, Duke University, Durham, NC

[2]Department of Medicine, Duke University, Durham, NC

[3]Department of Surgery, Duke University, Durham, NC

[4]Department of Immunology, Duke University, Durham, NC

* equal contribution

[#] Correspondence to: kevin.wiehe@duke.edu and mihai.azoitei@duke.edu

Summary

B cell lineages that are the current focus of vaccine development efforts against HIV-1, influenza or coronaviruses, often contain rare features, such as long heavy chain complementarity determining regions (CDRH3) loops. These unusual characteristics may limit the number of available B cells in the natural immunoglobulin repertoire that can respond to pathogen vaccinations. To measure the ability of a given immunogen to engage naturally occurring B cell receptors of interest, here we describe a mixed experimental and bioinformatic approach for determining the frequency and sequence of CDRH3 loops in the immune repertoire that can be recognized by a vaccine candidate. By combining deep mutational scanning and B cell receptor database analysis, CDRH3 loops were found that can be engaged by two HIV-1 germline-targeting immunogens, thus illustrating how the methods described here can be used to evaluate candidate immunogens based on their ability to engage diverse B cell lineage precursors.

Introduction

For an effective vaccine, it is important to ensure that B cell receptors exist in the naïve B cell repertoire that can be engaged and activated by a candidate immunogen. B cell activation depends on the overall frequency of the target B cell population and the affinity of the immunogen for the respective BCRs[1, 2]. Many antibodies against diverse viruses employ rare long heavy chain complementarity determining region 3 (CDRH3) loops for neutralization which may limit the number of available B cells that can be engaged by vaccination to elicit such humoral responses (**Supplementary Figure 1**). For example, HIV-1 broadly neutralizing antibodies (bnAbs) that target the V2 apex and glycan-V3 epitopes on HIV-1 Env typically contain CDRH3 loops of over 20 amino acids in length[3-9]. Similarly, bnAbs that bind to the influenza neuraminidase or hemagglutinin utilize long CDRH3 loops for the majority of their viral antigen contacts[10-13]. Recently, some isolated antibodies that neutralize both existing and emergent coronaviruses were founds to also rely on long CDRH3 loops for recognition[14-17]. Because of their potency and broad recognition of diverse viral isolates, the elicitation of these types of antibodies is currently of primary interest for HIV-1, universal influenza and pancoronavirus vaccine development efforts [2, 18-28].

The CDRH3 loop is the major antibody site involved in antigen recognition. Compared to the other five antibody CDR segments, CDRH3 loops exhibit significantly higher sequence and structural diversity, which allow them to recognize various antigens[29]. The median length of human CDRH3 loops is approximately 15 amino acids[30-32], although loops longer than 20 amino acids are expected to be present in ~13% of the BCR repertoire of an individual[33]. CDRH3 loops form through VDJ recombination, a process that involves the introduction of double-strand breaks in DNA to join the V, D, and J genes, and a break repair mechanism that adds random non-templated (N-) nucleotides at the junction sites[34-36]. Therefore, CDRH3 loops contain genetically encoded segments as well as non-templated regions. While CDRH3 loops typically

acquire point mutations during antibody affinity maturation in response to antigen stimulation, length altering insertions and deletions are rare[37, 38]. Rather, long CDR3s typically result from the specific usage of long germline encoded D and J gene segments and/or the addition of numerous N-nucleotides during VDJ recombination. In order to elicit antibody lineages that rely primarily on CDRH3 contacts for virus neutralization, it is critical to ensure that candidate immunogens can bind with high affinity to BCRs that contain CDRH3 loops related to those of target antibodies. This is particularly important when such antibodies contain unusually long CDRH3 loops that are typically rare among natural BCRs, like the ones described above against HIV-1, influenza, and coronaviruses.

Currently, natural BCRs engaged by a given immunogen are identified by labeling human PBMCs with the target molecule, followed by Florescence Activated Cell Sorting (FACS) to select B cells with the desired phenotype. The sequence of the isolated BCRs is subsequently determined and corresponding IgGs are typically produced recombinantly and characterized for binding to the target immunogen. This approach requires the ability to access and manipulate large numbers of human B cells, involves laborious and challenging experimental methods, can be biased by the selection strategy, and is expensive. To address these limitations, here we describe a mixed experimental and bioinformatic approach to identify BCRs in the natural human repertoire that contain CDRH3 loops predicted to be bound by a target immunogen (**Figure 1**). For a given antibody, our approach utilized deep scanning mutagenesis to rapidly identify possible changes in the sequence of its CDRH3 loop that are tolerated by a candidate immunogen. Using the resulting amino acid substitution profile, a public database containing the sequences of >100 million human BCRs from 10 subjects was queried to identify CDRH3 sequences predicted to be capable of binding the target immunogen (**Figure 1**). Candidate CDRH3 loops of interest were subsequently selected for experimental characterization.

Here, we applied this platform to analyze two HIV-1 Env SOSIP immunogens, CH505.M5.G458Y and CH848 10.17DT, which have been shown previously to activate in animal models precursors of the CH235 CD4 binding site and the DH270 V3-glycan HIV-1 bnAb B cell lineages [20, 23]. The mature bnAbs, DH270.6 and CH235.12, neutralize an estimated 51% and 89% respectively of circulating HIV-1 viruses, and their elicitation is the focus of HIV-1 vaccine development efforts[20, 23]. The initial step in the induction of HIV-1 bnAbs is to activate B cells whose BCRs can be subsequently matured to acquire broad neutralization activity. To this end, the 10.17DT immunogen was engineered to bind the inferred unmutated common ancestor (UCA) of the HIV-1 bnAb DH270.6. DH270 UCA contains a 20 amino acid CDRH3 loop that is the major site of interaction with the glycan-V3 epitope on HIV-1 Env. Similarly, CH505.M5.G458Y is a germline targeting immunogen that binds with high affinity to the UCA of CH235.12, an HIV-1 bnAb that targets the CD4 receptor binding site. CH235 UCA contains a 13 amino acid CDRH3 loop, which is an important, but not major, site of antigen interaction. Using our platform (**Figure 1**), we found that the natural B cell repertoire contains a high number of BCRs with CDRH3 loop sequences that should permit engagement by CH505.M5.G458Y. In contrast, B cell activation by the 10.17DT immunogen will be more limited given the restrictive sequence requirements of CDRH3 loops recognized by this molecule. Natural CDRH3 loops that are bound by the DH270.6 germline targeting immunogen 10.17DT were identified and validated, thus illustrating how our approach can be employed to evaluate the ability of vaccine candidates to engage BCRs present in the human B cell repertoire.

Results

Identification of CDRH3 loop variants recognized by germline-targeting immunogens that bind DH270.6 and CH235 bnAb precursors. Structural analysis revealed that the 20 amino acid CDRH3 loop of DH270 UCA contributes ~60% of the antibody buried surface in the 10.17DT binding complex (467Å^2 out of 799Å^2), by making significant interactions

with both the V3 loop as well as the glycan present at position N332 (**Figure 2A, Supplementary Figure 1**). In contrast, CDRH3 loop mediated interactions between the CH235 UCA antibody and the CH505.M5.G458Y immunogen are less substantial; the 13 amino acid CDRH3 loop contributes only ~30% (245Å^2 out of 858Å^2) of the total antibody buried surface at the interface (**Figure 2B, Supplementary Figure 1**). Based on these analyses, we hypothesized that 10.17DT binding to DH270 UCA would be sensitive to amino acid composition of the CDRH3 loop, while CH505.M5.G458Y would maintain high affinity interactions with diverse CDRH3 loop variants of CH235 UCA.

To determine the ability of 10.17DT and CH505.M5.G458Y to engage antibody precursors with diverse CDRH3 loops, we first developed site saturation mutagenesis libraries that sampled all single amino acid variants in the CDRH3 loops of DH270 UCA and CH235 UCA antibodies. Libraries of scFv versions of these mutated antibodies were displayed on the surface of yeast and clones that maintained binding to the target immunogen were isolated by FACS (**Supplementary Figure 2**)[39]. The DNA of selected clones was subsequently extracted and analyzed by next generation sequencing. The ability of an immunogen to bind a particular CDRH3 loop mutation was measured by determining the frequency of the respective amino acid substitution in the clones selected with the immunogen, relative to the frequency of the same substitution in the clones of the naïve, unsorted library. An increase in the presence of a mutation among the sorted clones indicated that the respective CDRH3 amino acid favors immunogen binding, while a decrease denoted that the mutation was detrimental. Using this approach, we evaluated the effect of all possible single amino acid substitution in the CDRH3 loops of DH270 UCA and CH235 UCA towards binding by 10.17DT and CH505.M5.G458Y SOSIPs, respectively.

As anticipated based on structural analysis, CH505.M5.G458Y maintained significant binding to a large number of CDRH3 substitutions in the CH235 UCA (**Figure 2C, E,**

Supplementary Figure 3A). The 13 residue CDRH3 loop of CH235 UCA is encoded by the VH1-46, D3-10*01 and JH4*02 genes, with eight residues inserted by N-nucleotide additions. At the 10 positions encompassed by the N-nucleotide and D gene regions, the immunogen recognized an average of 15.3 amino acids. In contrast, a significantly smaller number of DH270 UCA CDRH3 loop mutations resulted in antibodies that maintained 10.17DT binding (**Figure 2D, 2F, Supplementary Figure 3B**). This CDRH3 loop is encoded by VH1-2*02, D3-22*01, and JH4*02 genes, with 9 amino acids encoded by non-templated N-nucleotide additions. Because limited diversity was observed in the naïve DH270 UCA CDRH3 loop library at position 104 upon sequencing, the effect of substitution at this site on 10.17DT binding was determined experimentally. IgG antibody variants containing all 19 single amino acid substitutions were expressed recombinantly and their ELISA binding to 10.17DT was compared to that of the native DH270 UCA (**Supplementary Figure 4**). On average, only 5.3 different amino acids were tolerated by 10.17DT at each CDRH3 loop site. The largest number of functional variants was found in the V_H and J_H templated regions, where only one site tolerated fewer than six different amino acids. In contrast, the composition of the D-gene and N-addition encoded residues was more restricted, with an average of 3.5 alternative amino acids that maintained immunogen binding identified at each site. These results indicate that CH505.M5.G458Y engagement of BCRs upon vaccination will not be restricted by the amino acid composition of their CDRH3 loops, since this immunogen can bind CH235 UCA variants with highly divergent sequences in this region. In contrast, our data suggests that 10.17DT will only bind natural BCRs that have CDRH3 loops with highly conserved sequence identity to DH270 UCA.

To validate the results of scFv library screening, a subset of DH270 UCA mutants were expressed as recombinant IgG proteins and tested for binding to 10.17DT by surface plasmon resonance (SPR) (**Figure 2G**). The four mutations tested at position W101 confirmed the

substitution profile data, with three changes that significantly affected DH270 UCA binding (W101R: $\log_2$ enrichment=-2.6, % binding of WT DH270UCA maintained=0.1%; W101G: -2.6, 4.3%; W101L: -2.6, 4.1%) and one that had a more moderate effect (W101: -0.78, 77%) as expected. Similarly, there mutations tested at positions N113 and I102D were found to significantly decrease binding as predicted (N113I: -1.7, 7.4%; N113D: -2.1, 17.2%; I102D: -4.8, 7.9%). At position 111, three mutations predicted to reduce binding also did so in recombinant IgGs (Y111W: -1.6, 61.9%; Y111H: -2.6, 30.6%; Y111N: -1.5%, 0%), although the extent of the loss measured in IgGs did not correlate with the scale of the decrease from the substitution profile data. One mutation at this site, Y111F, was found by library screening to increase binding to 10.17DT, but actually reduced immunogen binding by 50% in the recombinant IgG. At position 102, three mutations tested were expected to show DH270 UCA binding levels at least as high as 10.17DT. While this was confirmed for two of them (I102N:1.7, 128.4%; I102V:0.95, 99.1%), one mutation led to a 35% loss in binding (I102S: 0.66, 65.3%). Binding discrepancies between deep mutational libraries and recombinant IgGs are likely due to the two different platforms used to assess interactions that employ distinct antibody formats (FACS of scFvs displayed on the surface of yeast versus SPR of recombinant IgGs) and are in line with those observed in other deep scanning mutagenesis studies[40, 41]. Nevertheless, these data illustrate that high throughput screening of single site saturation libraries sufficiently recapitulates the affinity trends of IgGs containing single amino acid mutations.

Identification of CDRH3 loops in the human BCR repertoire that can be bound by 10.17DT and CH505.M5.G458Y. Next, we sought to identify CDRH3 loops in the BCR repertoire of individuals that can be engaged by 10.17DT or CH505.M5.G458Y as a way to estimate the number of B cells that these molecules may engage upon vaccination. To find similar CDRH3 sequences to those recognize by these immunogens, we searched a human BCR database developed by Briney et. al.[30] that contains ~85 million non-redundant functional

BCR sequences isolated from 10 individuals. In order to determine the frequency of DH270-like CDRH3s, we selected CDRH3 loops from the BCR database that had identical length (20 amino acids), correct D gene usage (D3-22), and had the same D gene reading frame (reading frame 2) and position within the CDRH3 as DH270 UCA. Of the 85,149,053 database sequences we analyzed, 47,975 CDRH3 loops matched our CDRH3 criteria, yielding an estimated frequency of 1 in 1,774 of BCRs with DH270-like CDRH3 properties. Next, we aligned these DH270-like CDRH3s against the CDRH3 substitution profile (**Figure 2D**) from the 10.17DT library screening to predict which CDRH3s would be compatible with 10.17DT binding. Any amino acid in the CDRH3 substitution profile with a log enrichment value below -0.2 upon 10.17DT selection was considered detrimental to 10.17DT binding and counted as a mismatch in the alignments. Of the 47,975 DH270-like CDRH3s analyzed, no sequences fully matched the amino acid variants in the substitution profile, and only one contained just one mismatch (**Supplementary Figure 5**). Therefore, the predicted frequency of BCRs expected to be bound strongly by 10.17DT is below the 1 in 85 million limit of detection set by the size of the analyzed database.

Next, we searched the BCR database for CDRH3s related to CH235 UCA that can be expected to be recognized by the target immunogen CH505.M5.G458Y. In contrast to DH270 UCA, the composition of the CDRH3 sites encoded by the D gene and recognized by CH505.M5.G458Y was not restricted in CH235 UCA (**Figure 2E**). Therefore, we did not limit the BCR database search using any D gene position or composition criteria as we did for DH270 UCA. The database search yielded 9,623,490 CDRH3s that matched the CH235 UCA CDRH3 length of 15 amino acids (11.3% of database sequences). CH505.M5.G458Y binding to CH235 UCA was not affected by the majority of the single amino acid substitutions in the CDRH3 loop (**Figure 2C**). Accordingly, we found that CDRH3 loops predicted to be bound by CH505.M5.G458Y (0 mismatches according to the substitution profile) occurred with a high frequency of 1 in 100 sequences from the human BCR database (**Supplementary Figure 5**).

These data suggested that CDRH3s compatible with CH505.M5.G458Y binding should be readily available in the naïve B cell BCR pool.

By considering different log enrichment values at which a CDRH3 substitution is deemed acceptable for immunogen binding, various sets of BCRs can be identified from the database, and the associated frequency of B cells in the immune repertoire that can be engaged by the target immunogen can be computed. While log enrichment values higher than -0.2 were considered in the analysis above, this threshold value can be lowered or increased in order to identify BCRs that are either more or less likely to be recognized by a target immunogen. This analysis can provide a measure of the sequence "distance" and the relative abundance of CDRH3 loops from the natural immune repertoire in relationships to the ideal CDRH3 loop sequence recognized by the target immunogen. CH505.M5.G458Y recognition of natural CDRH3s was predicated to remain high even when using stringent thresholds of acceptable amino acid mismatches. For example, if only amino acids with $\log_2$ enrichment of 0 or higher (corresponding to an estimate of no reduction in binding) are considered acceptable, the frequency of CH505.M5.G458Y compatible CDRH3s is ~1 in 10,000 (**Figure 3B**). In contrast, CDRH3s compatible with 10.17DT binding were only found when thresholds were lowered substantially, with estimated precursor frequencies of ~1:10 million only found when single substitutions predicted to reduce 10.17 DT binding by 40% reduction were accepted as matches (**Figure 3A**).

Because the queried BCR database contains only a small fraction of the diversity expected to be present in the immune repertoire of an individual at any one time (estimated to be 10^9-10^{12} unique BCRs)[30, 33, 34], we next wanted to determine if CDRH3s predicted to be compatible with 10.17 DT binding could be present within a set of sequences that was the same order of magnitude. To address the limitation of available repertoire depth in the BCR database, we used the computational program IGoR[42] to simulate the VDJ recombination process and to

randomly generate ~2.5 trillion CDRH3 sequences. Of these, 1.15 billion had the same length, same D gene usage and reading frame, and same D gene position as DH270 UCA. The frequency of simulated sequences that matched these criteria (~1:2,100), was similar in magnitude to the frequency of sequences with these same properties present in BCRs database (~1:1,800). Moreover, amino acid frequencies at each position in the simulated CDRH3s were remarkably similar to those observed in the natural BCR sequence database (**Figure 3D)**. Thus, these data indicate that sampling from IGoR-simulated sequences can recapitulate the distribution of CDRH3s in the human naïve BCR repertoire. Of the 1.15 billion simulated sequences with the same immunogenetics as those of the DH270 UCA CDRH3, 125 fit the 10.17 DT substitution profile with 0 amino acid mismatches ($\log_2$ enrichment <-0.2), yielding a predicted frequency of ~1 in 20 billion B cells (**Figure 3E).** The distributions of 10.17DT-compatible CDRH3 loops was highly similar in the BCR database and the IGoR generated set across different substitution threshold values, despite the large difference in the total number of sequences (~48,000 versus 1.15 billion) (**Figure 3A, C**). IGoR produced a smoother histogram due to the detection of sequences compatible with 10.17DT binding in the 1 in 10 million to 1 in 1 billion range. This sampling range was unattainable by BCR database query, thus supporting the IGoR modeling approach for accessing antibody sequences beyond those available experimentally.

We expect that immunogens that recognize only a limited number of amino acids in the N-nucleotide encoded regions of the target CDRH3 loops to subsequently engage few B cells upon vaccination due to the high diversity of the N-nucleotide encoded regions in the natural BCR repertoire. To identify which CDRH3 positions were most responsible for limiting the number of BCRs compatible with 10.17DT binding, we computed the per-position cumulative frequency within the BCR database of acceptable amino acids identified in the DH270 CDRH3 substitution profile (**Supplementary Figure 6A**). 10.17DT compatible amino acids at CDRH3

positions 5, 16, and 17 (**Figure 2F**) had the lowest cumulative frequency (<5%) among the amino acids present in 20 residues long CDRH3 sequences from the BCR database (**Supplementary Figure 6A**). This information can guide immunogen design, in order to increase the recognition of more BCRs. For example, an improved DH270 UCA targeting immunogen that could bind to CDRH3 loops containing the top three most frequent amino acids in the BCR database at positions 5, 16, and 17 would increase the frequency of recognized CDRH3s in the repertoire by a factor of ~2300 over 10.17DT (**Supplementary Figure 6B**). Such an improved immunogen would increase the estimated frequency of engaged CDRH3s from 1 in 20 billion to 1 in 8.5 million, bringing the precursor frequency of the targeted B cells into a range that is more likely to result in their competitive success in the germinal center[1].

To characterize the ability of 10.17DT to bind naturally occurring CDRH3 loops experimentally, we developed chimeric DH270 UCA antibodies where the native loop was replaced with a CDRH3 loop from the human BCR database that had the same length, same D gene usage and reading frame, and same D gene position as DH270 UCA. Natural CDRH3 loops were scored and ranked based on their sequence similarity to the DH270 UCA CDRH3 substitution profile (**Supplementary Figure 7**). The top 100 ranked antibodies were expressed together as scFvs and displayed on the surface of yeast (**Figure 3E)**. Upon sorting for two rounds with 10.17DT, no high affinity binding was observed by FACS, although some scFv sequences were enriched in the selected clones, indicative of low affinity for the antigen. Seven such chimeric antibodies were expressed and purified as recombinant IgG and their binding to 10.17DT was measured by SPR. 10.17DT bound weakly to the DH270 UCA chimeric antibodies containing natural CDRH3 loops, with binding levels less than 10% of that measured for the unmutated DH270 UCA mAb (**Figure 4**). These results were expected, since all the CDRH3 loops tested experimentally contain at least two amino acids predicted to greatly reduce 10.17DT binding based on the substitution profile.

Overall, these data revealed that the naive B cell repertoire contains a high number of BCRs with CDRH3 loop sequence that permit engagement by CH505.M5.G458Y, while B cell activation by the 10.17DT immunogen will be significantly limited by the lack of BCRs containing CDRH3 loops expected to be bound by this immunogen. By analyzing a large collection of BCRs either sequenced from the naïve B cell repertoire or generated computationally, these methods can rapidly estimate the frequency of characteristic B cells that can be engaged by an immunogen upon vaccination.

Discussion

For the successful elicitation of target antibodies by vaccination, a candidate immunogen needs to potently activate related B cell lineages that can subsequently evolve the viral neutralizing functionality through somatic mutations. This is particularly important and needs to be explicitly considered during the immunogen design process, if target B cells are expected to occur in the naive B cell repertoire at low frequencies. Indeed, many bnAbs of interest for HIV-1, influenza or coronavirus vaccine development contain rare features such as long CDRH3 loops (**Supplementary Figure 1**). For example, bnAbs against the V2 loop and glycan-V3 loop epitopes of HIV-1, such as CH01, VRC26.25, and PG9, those that target the neuraminidase or HA stem of influenza, like 1G01 or FI6v3, or those that cross-neutralize diverse coronavirus strains, like DH1047, contain CDRH3 loops over 20 amino acids that are responsible for 50% or more of the binding interaction with the virus. As our analysis of different antibodies shows (**Supplementary Figure 1**), loops of this length typically contain long segments generated by N-nucleotide addition. Since N-nucleotides are not templated by germline V, D, or J gene segments, the resulting amino acids encoded by N-nucleotides are highly diverse across antibodies with the same immunogenetics (**Supplementary Figure 6**). Therefore, a candidate immunogen must tolerate multiple amino acids in the N-nucleotide encoded regions of the target BCRs in order to robustly activate a large number of B cells upon vaccination. This is because

the overall CDRH3 frequency in the immune repertoire can be estimated as the product of frequencies of tolerated amino acids at each individual position in the CDRH3 loop. Therefore, each additional CDRH3 position in which immunogen recognition is restricted to only a few amino acids leads to a rapid reduction in the overall frequency of natural CDRH3 loops that can be engaged by vaccination. Our study thus supports the explicit design of immunogens that can accommodate diverse amino acids in the N-nucleotide encoded regions of BCRs related to those of target antibody lineages as an immunogen design strategy to elicit antibodies containing long CDRH3 loops.

Typically, the frequency of B cells that can be activated by a given immunogen is estimated by FACS analysis of natural B cells isolated from uninfected individuals. Cells are selected based on their ability to selectively bind a target immunogen and their BCRs are subsequently identified and characterized. This approach is laborious, expensive, may be biased by the particular cell labeling and sorting strategy, and requires access to large number of isolated B cells, especially when the target BCRs are expected to be present at low frequency naturally. In comparison, the approach described here relies on bioinformatic analysis of human BCR sequences from either publicly available databases or simulated from models of VDJ recombination, using sequence identification criteria determined experimentally. For a given antibody-antigen pair, we first employed deep mutational scanning to determine the single amino acid variants in the CDRH3 loop that maintained antigen binding. Based on the resulting data, we then found human BCRs containing CDRH3 loops related to the target antibody and that are expected to be recognized by a particular immunogen. This method allows the identification of CDRH3 loops that are significantly different in sequence than that of the target antibody, yet are still expected to be bound by the given immunogen. In order to sample the immune repertoire at sequence depths beyond those available in databases of isolated human BCRs, we demonstrated that IGoR simulations can accurately model natural CDRH3 loop

sequence diversity. These synthetic sequences can then be analyzed to determine the frequency of rare BCRs that may be engaged by a target immunogen. With this approach, we found that BCRs containing CDRH3 related to that of CH235 UCA, and that are predicted to be engaged by the targeting immunogen CH505.M5.G458Y, are abundant in the natural repertoire. Therefore, we expect that vaccination with CH505.M5.G458Y will engage and activate CH235.12 precursors with diverse CDRH3 loops. In contrast, 10.17DT recognition of DH270.6 precursors is likely restricted to antibodies with CDRH3 loops that contain the same D gene and are highly similar in sequence to the DH270 UCA CDRH3. Three or less amino acid variants were tolerated by 10.17DT at 14 of the 20 CDRH3 residues, and few D gene substitutions preserved immunogen binding. We identified and characterized several naturally occurring CDRH3 loops that are bound by 10.17DT, albeit with low affinity, when transplanted onto the backbone of DH270 UCA. Antibodies with CDRH3 loops of the same length and contain the same D gene in the same reading frame at the identical position as in DH270 UCA are present with high frequency in the natural repertoire (~1:1700). However, 10.17DT recognition of these loops is limited by their amino acid composition in N-nucleotide encoded regions. Indeed, at 9 such sites of DH270 UCA, 10.17DT recognizes only 4.2 amino acids on average. This example of highly restricted amino acids requirements in the CDRH3 for 10.17DT recognition emphasizes that B cell frequency definitions that rely simply on immunogenetic features of CDRH3s like length, D gene usage and D gene position, without regard to the specific amino acid composition of the CDRH3, particularly in the randomly distributed N-nucleotide regions, could lead to gross overestimations of the frequencies of B cells that can be engaged by candidate immunogens. Our library screening based approach should lead to better estimates of the frequencies of compatible CDRH3s by considering the contribution to binding of each amino acid at each CDRH3 position. Additionally, the data from the library screening approach can be used to inform an immunogen design strategy for increasing the number of B cells that

are activated by optimizing the immunogen to tolerate more diverse amino acids in the CDRH3 loop regions encoded by N-nucleotide addition.

The CDRH3 centric approach used in this study has some limitations in estimating the frequency of B cells targeted by a given immunogen. It is possible that some CDRH3 loops identified as favorable by our method could be part of BCRs that contain other molecular features that prevent immunogen binding, such as incompatible VH or VL genes. In addition, the deep scanning mutagenesis approach evaluates the effect on immunogen binding of only one amino acid change in the context of the WT CDRH3 of the target antibodies. However, as shown in this study, many of the BCRs identified contain multiple amino acids changes relative to the target CDRH3. These loops may contain novel structural and molecular features where the additive effect of individual mutations may no longer predict the overall binding propensity to the target immunogen. Nevertheless, in this study we were able to identify and validate experimentally CDRH3 loops that showed binding to 10.17DT, yet had up to 45% different amino acids than those of the DH270 UCA CDRH3. On the other hand, it is also possible that our approach may omit BCRs that contain CDRH3 loops with slightly different immunogenetics than those of target antibody, but that could still be activated by the target immunogens and subsequently mature into antibodies with the desired reactivity. For example, it is possible that 10.17DT may recognize BCRs that contain CDRH3 loops of different lengths than that of DH270 UCA and that these BCRs could develop into mature antibodies with similar function to those from the DH270 lineage either by acquiring indels or by yet unknown evolutionary pathways.

While in this study we focused our analysis on naïve B cell BCR CDRH3 loop sequences, our method is generalizable, and other BCR features important for immunogen recognition, such as variable heavy or light chain gene templates or specific acquired somatic mutations, can be used as search criteria.

Main Figures

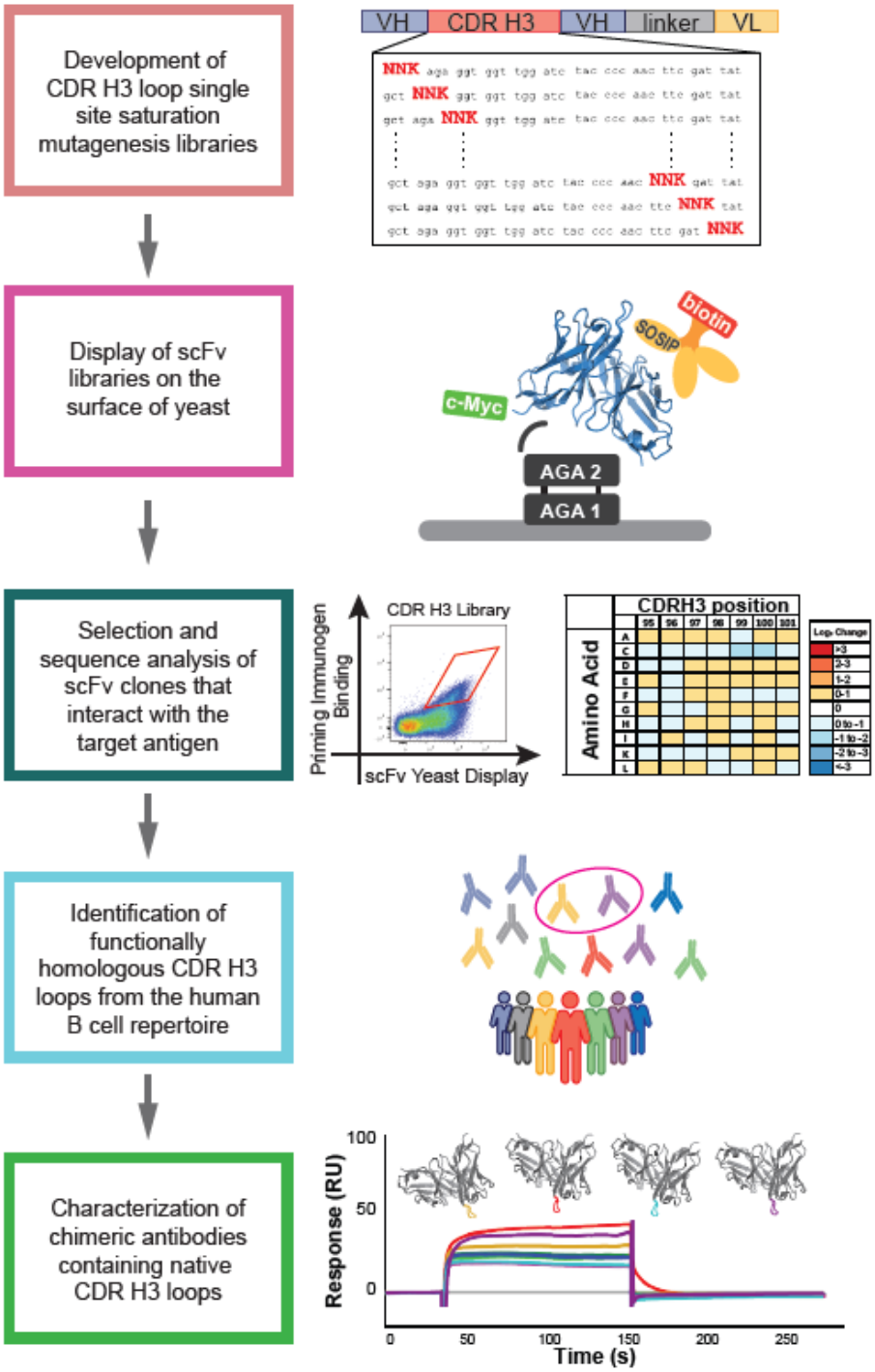


Figure 1. Combined experimental and bioinformatic workflow to characterize the ability of candidate immunogens towards the binding of diverse CDRH3 loops from the natural B Cell Receptor (BCR) repertoire.

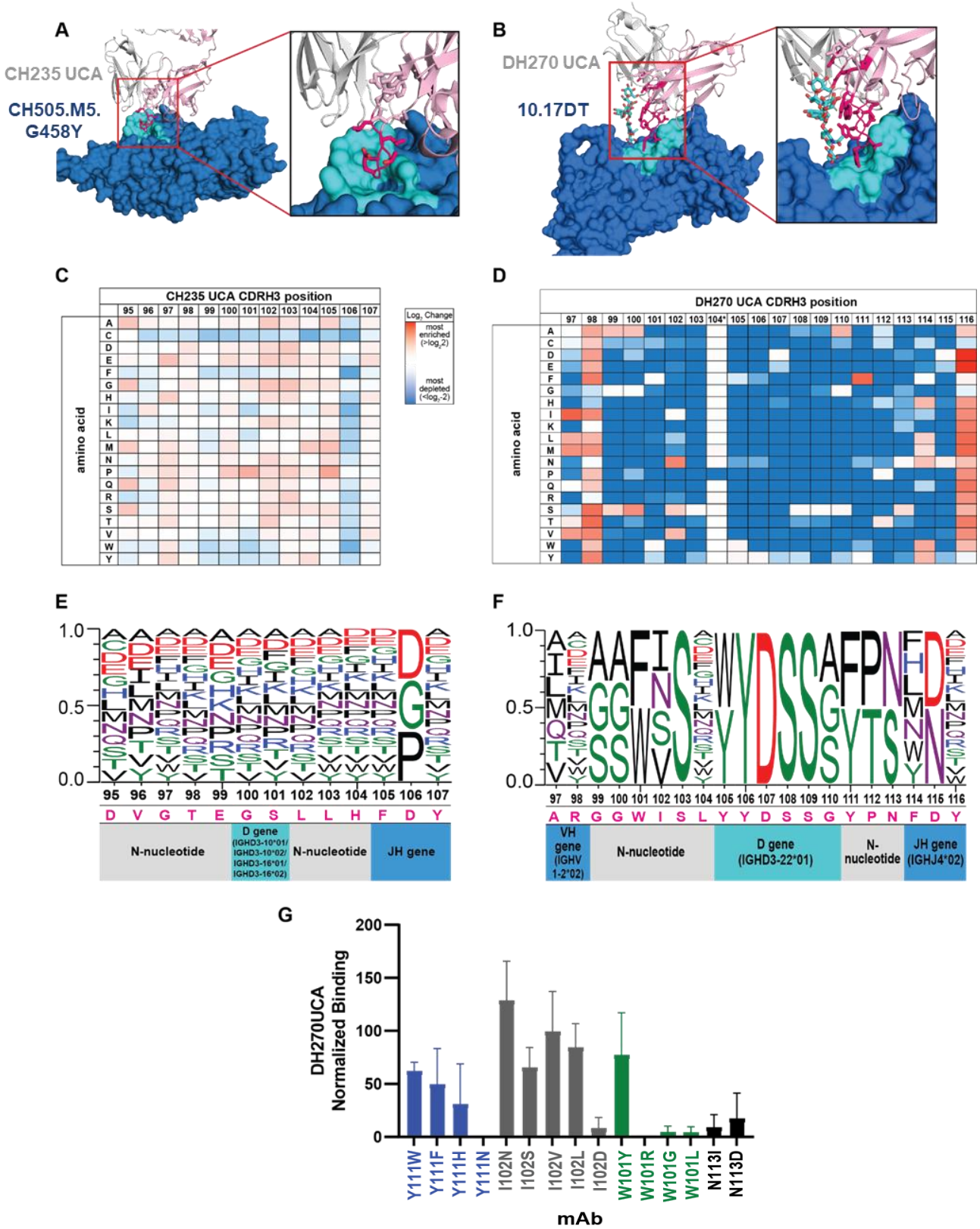


Figure 2. Immunogen recognition of CDRH3 loop variants in the unmutated common ancestors (UCAs) of HIV-1 bnAbs CH235.12 and DH270.6. Structural representation of the

CH505.M5.G458Y (**A**) and 10.17DT (**B**) immunogens (*dark blue*) in complex with CH235UCA and DH270UCA respectively (heavy chain: *pink*; light chain: *grey*). CDRH3 residues are shown in magenta sticks and the residues they contact on the immunogens are indicated in cyan. Tables of $\log_2$ enrichment ratios describing the effect of each amino acid substitution in the CDRH3 loops of CH235UCA (**C**) and DH270UCA (**D**) towards target immunogen binding. Values are plotted from highly enriched (*red*) to significantly depleted (*blue*) and represent the average of two independent experiments. **E, F.** Logo plots depicting CDRH3 loop amino acid variants that maintained antibody binding ($\log_2$(enrichment)>-0.2) to the target immunogen. **G.** DH270UCA3 IgGs containing the single mutation indicated on the x-axis were expressed recombinantly and their binding to 10.17DT was measured by ELISA and SPR.

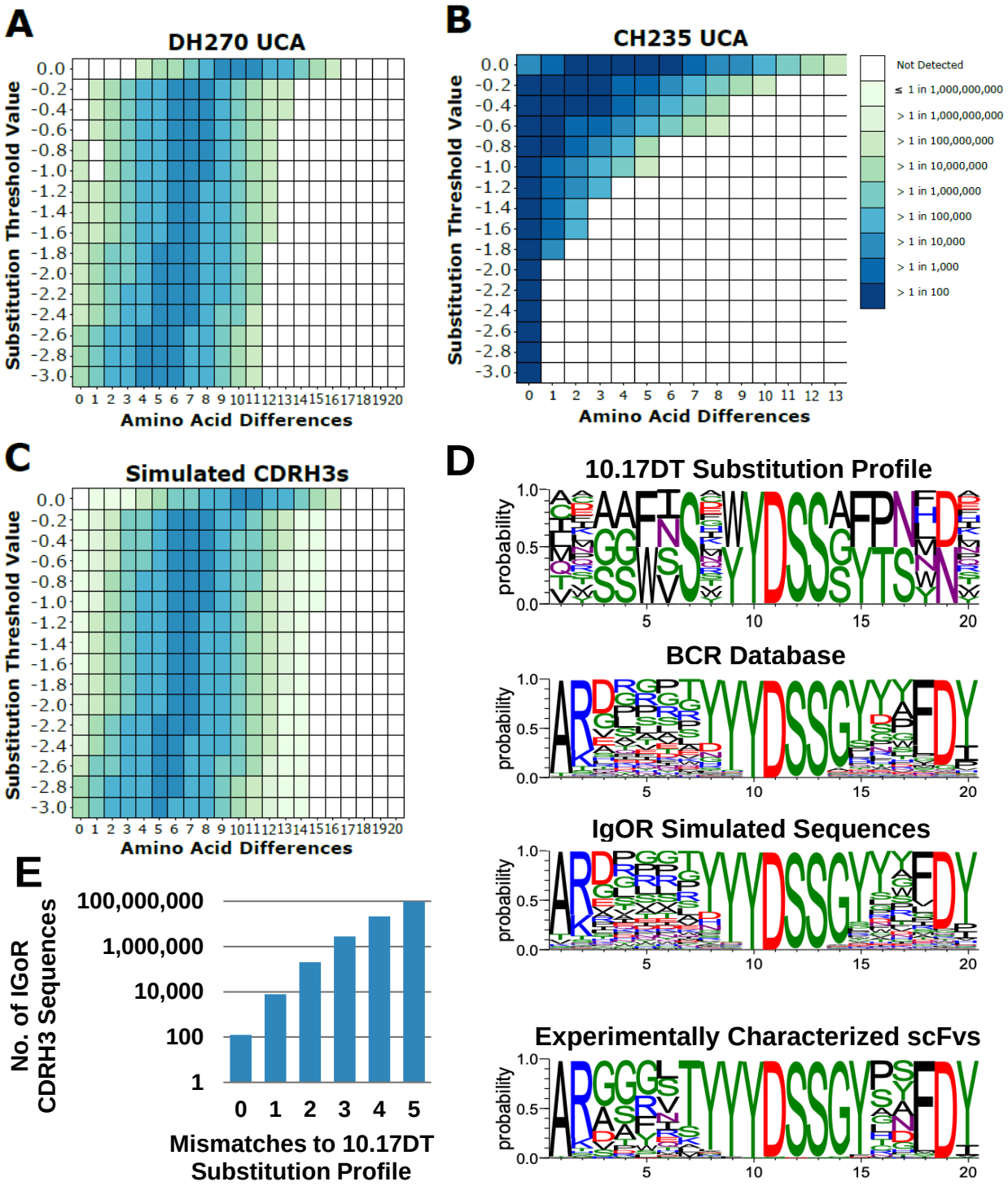


Figure 3. Frequency of CDR H3 loops predicted to bind 10.17DT and CH505.M5.G458Y. Heatmaps displaying the frequencies of **A**) 10.17DT compatible (DH270 UCA-like) and **B**) CH505.M5.G458Y.GNTI- compatible (CH235 UCA-like) CDRH3s in the Briney et al. BCR sequence database. Each heatmap shows the calculated CDRH3 loop frequency for a given number of amino acid differences (*x-axis*) at varying degrees of substitution threshold values (*y-axis*) relative to the respective CDR H3 substitution profile. The substitution threshold value is

the cutoff of log fold change in binding above which a specific amino acid at a given position in the CDRH3 amino acid substitution profile is considered compatible with immunogen binding. Higher substitution threshold values correspond to a stricter definition of what constitutes a matching amino acid in the CDRH3 substitution profile. **C.** Frequency of IgOR simulated CDRH3s that match the amino acid substitution profile after filtering for loops that have the same length, D gene, D gene reading frame, and D gene position as the CDRH3 of DH270UCA. **D.** Sequence logos for: DH270 UCA substitution profile based on the scFv screening library; CDRH3s from the BCR database of 20 amino acids in length and that utilize the IGHD3-22 gene with the core “YDSS” motif starting at position 10; Simulated dataset of CDRH3 sequences of 20 amino acids in length and IGHD3-22 gene usage with the core “YDSS” motif starting at position 10; and the 100 chimeric DH270 UCA antibodies expressed as scFvs on the surface of yeast. **E.** The number of DH270 UCA-like CDRH3 loop sequences simulated with IGoR and separated by the number of mismatches they contain relative to the 10.17DT binding profile.

A

Antibody	CDRH3 Sequence	Total Mutations	Unfavorable Mutations
DH270 UCA	A R G G W I S L Y Y D S S G Y P N F D Y	-	-
188	. . A . F . T Y . . . . . C G . . .	6	2
650a	. . A S R V T Y . . . . . A D . . .	8	3
650b	. . . S R V T Y . . . . . A D . . .	7	3
651a	. . A P R . T Y . . . . . L . .	6	3
713c	. . . . G V T Y . . . . . T S . . .	6	2
713d	. K S A Y . T Y . . . . . S Y L . .	9	3
650	. . . S F V M Y . . . . . Y S . . .	7	2

B

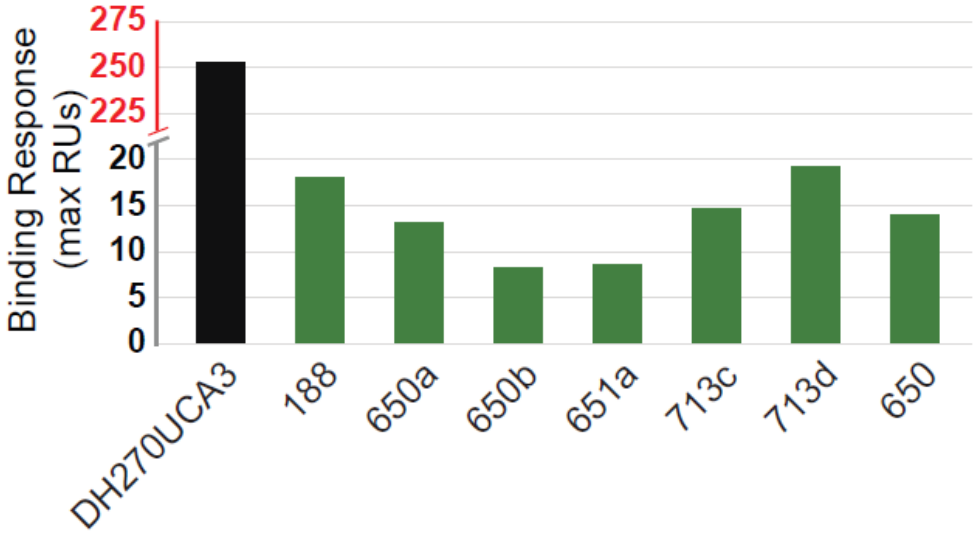
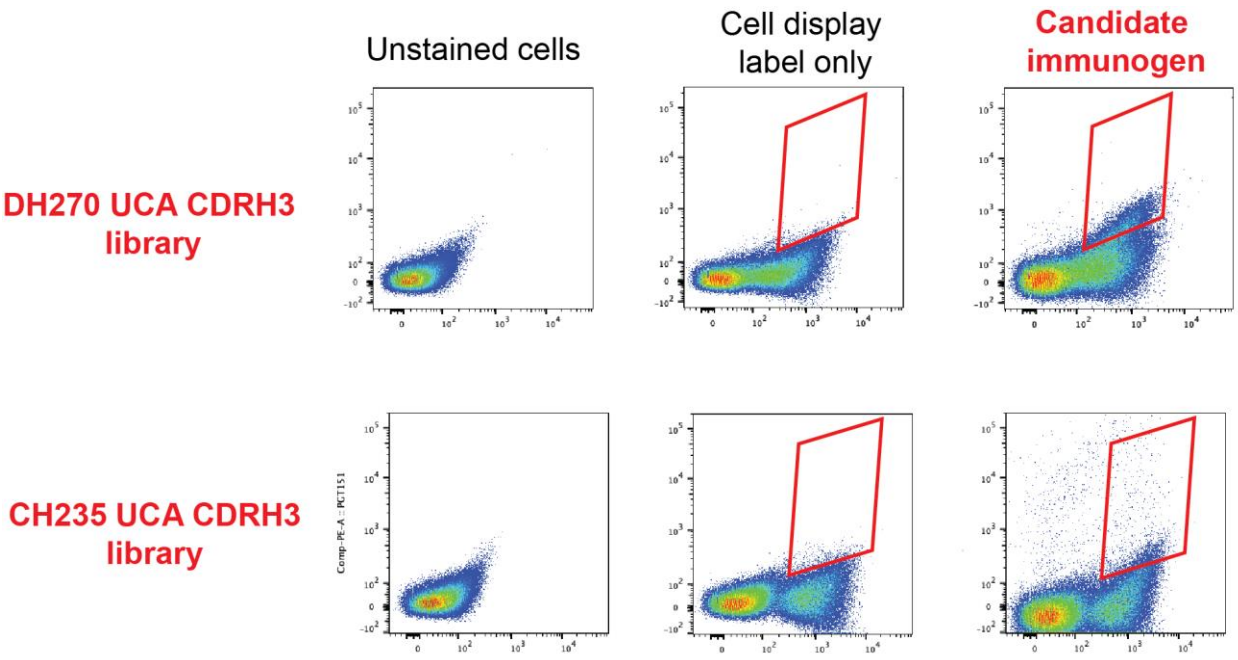


Figure 4. 10.17DT recognition of natural CDRH3 loops. A. Seven natural CDRH3 loops predicted to be recognized by 10.17DT were selected from a BCR database for additional characterization. Sequence alignments with the CDRH3 loop of DH270UCA3 highlight the sequence changes expected to be neutral (*black*) and detrimental (*magenta*) towards 10.17DT binding. CDRH3 residues encoded by the D gene are shown in *blue*, while those encoded by N-nucleotides are in *red* **B.** DH270UCA3 chimeric antibodies where the WT CDRH3 loop has been replaced by those described in (**A**) were expressed recombinantly and assessed for binding to 10.17DT by SPR**.**

Supplementary Figures

Antibody	Virus	Viral Protein Target	PDB ID	CDRH3 length	HC BSASA	LC BSASA	BSASA Total	BSASA CDRH3	% BSASA CDRH3
EDE2 A11	Dengue/Zika	E	4UTB	26	512	59	571	438	77%
HEPC3	HCV	E2	6MEI	19	953	68	1022	580	57%
CH235	HIV	Env CD4 bs	6UDA	13	613	275	889	275	31%
CH03	HIV	Env V1V2	5ESV	26	385	0	385	263	68%
VRC26.25	HIV	Env V1V2	6VTT	38	1115	0	1115	1115	100%
PG9	HIV	Env V2	3U4E	30	586	0	586	586	100%
BG18	HIV	Env V3	6DFH	23	395	326	721	389	54%
DH270	HIV	Env V3	6UM5	20	410	59	469	242	52%
CH65	Influenza	HA (RBS)	5UGY	19	554	296	850	396	47%
FI6V3	Influenza	HA (stem)	3ZTJ	22	542	320	863	532	62%
1G01	Influenza	NA	6Q23	21	780	251	1031	841	82%
DH1047	SARS-CoV-2	RBD	7LD1	24	546	238	784	380	48%

Supplementary Figure 1. **Multiple antibodies against diverse viruses contain long CDRH3 loops that are essential for viral recognition**. Antibody/antigen structural complexes describing a wide variety of antibody bound to viral proteins were analyzed using PDBePISA to determine the buried solvent accessible surface area (BSASA) at the binding interface in units of $Å^2$. Immunogenetics annotation programs (Cloanalyst, JoinSolver and Partis) independently inferred VDJ gene usage and N-nucleotide addition regions. We were unable to delineate the N-nucleotide, D and J regions for antibodies without publicly available heavy chain nucleotide sequences (denoted by ND in the table). BSASA binding contributions for residues in the light chain (LC) heavy chain (HC), and CDRH3 loop are reported for each antibody.



Supplementary Figure 2. **FACS selection of DH270 UCA and CH235 UCA CDRH3 single-site saturation libraries for binding to the target immunogens**. 10.17DT and CH505.M5.G458Y scFv libraries were displayed on the surface of yeast and labeled for cell surface display (x-axis) and binding to the candidate immunogen (y-axis) . DH270 and CH235 UCA libraries were sorted with 300nM concentration of biotinylated SOSIP target immunogens 10.17DT and CH505.M5.G458Y respectively, and all binding clones were recovered for downstream DNA isolation and NGS analysis.

A

CH235 UCA log 2 (fold change) towards CH505.M5.G458Y binding

	99	100	101	102	103	104	105	106	107	108	109	110	111
	D	V	G	T	E	G	S	L	L	H	F	D	Y
A	0.8	0.2	0.1	0.2	0.0	0.2	0.1	0.7	0.4	0.2	0.6	-0.6	0.2
C	0.0	-0.9	-0.7	-0.7	-1.3	-1.1	-0.9	-0.8	-0.6	-1.7	-1.1	-1.7	-0.3
D	0.0	-0.1	0.0	0.4	0.1	0.2	0.3	0.5	0.7	0.1	0.3	0.0	0.2
E	0.2	-0.1	0.8	0.3	0.0	0.5	0.3	0.7	0.6	0.3	0.3	-0.4	0.5
F	-0.4	-0.3	0.1	0.1	-0.7	-0.5	-0.6	-0.3	-0.2	0.1	0.0	-1.9	-0.3
G	0.8	-0.4	0.0	0.2	0.1	0.0	0.4	0.8	0.8	0.4	0.4	-0.2	-0.1
H	-0.2	-0.2	0.6	0.0	-0.1	0.1	-0.2	0.5	0.4	0.0	0.1	-0.9	0.3
I	-0.6	0.3	-0.1	0.1	-0.4	0.2	-0.8	-0.2	0.0	-0.2	0.4	-1.6	0.1
K	-0.8	-0.3	-0.2	0.0	0.0	0.0	0.1	0.5	0.3	0.0	0.1	-1.1	0.2
L	0.0	0.1	0.2	0.0	-0.3	0.2	-0.6	0.0	0.0	0.1	0.4	-1.0	0.1
M	0.7	0.0	0.4	0.2	-0.3	0.4	-0.2	0.1	0.1	0.7	0.9	-1.0	0.1
N	-0.1	0.2	0.4	0.1	0.1	0.2	0.1	0.5	0.3	0.0	0.4	-0.5	0.1
P	-0.3	0.3	0.5	0.2	0.0	0.7	1.1	0.5	0.7	0.3	1.2	-0.1	-0.1
Q	0.8	-0.3	0.6	0.2	-0.2	0.2	0.0	0.4	0.3	-0.1	0.3	-0.5	0.2
R	-0.6	-0.4	0.2	0.0	0.0	0.1	0.0	0.3	0.6	0.0	-0.1	-0.9	-0.1
S	0.7	-0.2	0.3	0.1	-0.1	0.3	0.0	0.5	0.4	0.4	0.4	-0.6	0.2
T	0.1	0.1	0.4	0.0	0.0	0.3	-0.2	0.6	0.1	-0.1	0.3	-0.7	0.2
V	0.1	0.0	0.0	0.0	-0.3	-0.1	-0.5	0.2	0.1	0.2	0.1	-0.9	0.3
W	-0.3	-0.3	0.3	-0.3	-0.8	-0.5	-0.9	-0.4	0.1	0.1	-0.1	-1.7	-0.2
Y	-0.3	0.0	0.5	0.1	-0.5	-0.6	-0.6	-0.5	0.2	0.4	-0.2	-0.7	0.0
*	-4.5	-4.7	-4.1	-4.7	-2.7	-4.1	-4.6	-2.5	-2.2	-4.4	-4.2	-3.8	-2.5

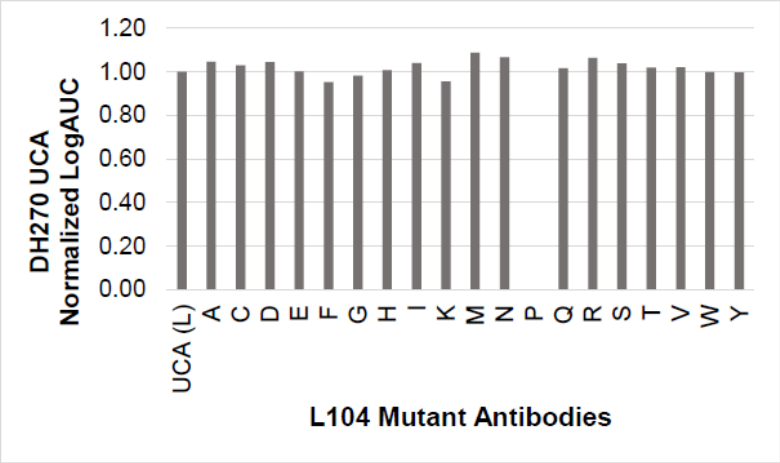
B

DH270 UCA log 2 (fold change) towards 10.17DT binding

	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116
	A	R	G	G	W	I	S	L	Y	Y	D	S	S	G	Y	P	N	F	D	Y
A	0.0	1.2	0.8	0.9	-6.0	-4.1	-3.5		-5.1	-4.8	-3.9	-1.7	-2.3	1.0	-5.9	-0.5	-3.4	-4.1	-3.1	0.7
C	-0.2	0.7	-1.5	-1.5	-1.4	-2.3	-2.6		-4.0	-3.4	-5.4	-2.9	-5.5	-1.6	-3.2	-0.7	-3.9	-0.7	-4.2	-0.6
D	-1.0	1.4	-3.6	-1.1	-4.8	-4.8	-5.0		-3.8	-4.8	0.1	-6.6	-6.2	-3.6	-4.5	-5.2	-2.1	-2.8	0.1	3.0
E	-1.7	1.1	-4.5	-4.0	-6.1	-2.6	-6.0		-5.8	-5.8	-0.9	-4.5	-5.8	-5.4	-3.1	-5.8	-5.3	-2.4	-1.0	3.0
F	-3.8	1.8	-5.5	-3.9	0.0	-2.8	-4.0		-0.5	-1.8	-5.4	-3.9	-4.7	-3.9	2.1	-5.0	-3.8	0.0	-4.0	0.2
G	-1.7	-0.6	0.0	0.0	-2.6	-5.2	-5.0		-5.5	-6.2	-3.9	-3.1	-2.3	0.1	-4.5	-3.7	-1.7	-4.9	-3.8	-0.3
H	-4.0	-1.0	-5.2	-5.3	-4.5	-4.1	-5.0		-3.8	-3.4	-2.3	-4.5	-5.2	-5.2	-2.6	-0.8	-1.8	1.0	-2.6	1.4
I	2.2	1.6	-4.6	-5.6	-5.0	0.0	-4.4		-4.6	-4.5	-4.6	-6.3	-7.2	-5.1	-3.9	-6.3	-1.7	-1.1	-4.5	1.6
K	-3.8	0.0	-5.2	-4.9	-4.9	-5.9	-5.4		-5.4	-4.5	-6.1	-5.2	-6.0	-4.8	-3.9	-5.4	-1.2	-4.1	-4.3	1.8
L	1.1	1.0	-4.5	-5.4	-2.7	-1.1	-5.3		-5.3	-5.5	-5.4	-5.3	-4.1	-5.6	-4.7	-4.2	-4.6	0.6	-4.7	2.1
M	1.5	1.5	-5.8	-7.5	-5.4	-2.3	-5.7		-5.2	-3.5	-4.9	-4.5	-6.0	-4.6	-4.7	-6.1	-3.5	0.4	-4.6	2.1
N	-2.5	-0.1	-5.0	-4.8	-5.2	1.7	-4.4		-1.9	-1.7	-2.9	-4.1	-4.6	-4.1	-1.5	-2.0	0.1	0.9	0.3	0.9
P	-2.6	0.9	-4.8	-1.4	-5.6	-5.1	-3.8		-6.4	-6.0	-5.2	-4.2	-4.0	-4.5	-4.0	0.1	-5.2	-5.6	-3.8	1.6
Q	0.0	0.7	-6.0	-11.1	-4.6	-4.2	-4.4		-5.1	-5.6	-4.5	-6.3	-6.3	-7.6	-5.2	-3.5	-4.7	-0.9	-5.0	1.9
R	-4.7	-0.1	-3.2	-3.5	-2.6	-2.2	-4.9		-5.3	-5.1	-5.9	-5.2	-4.9	-3.5	-5.2	-3.9	-5.0	-5.1	-5.0	1.5
S	-0.8	2.0	0.6	1.6	-4.0	0.7	0.1		-5.0	-4.3	-5.7	0.1	0.1	0.1	-4.0	-2.8	0.5	-3.3	-2.9	1.0
T	0.3	2.0	-3.8	-2.0	-4.9	-3.9	-2.5		-5.8	-5.5	-4.9	-1.9	-2.3	-6.1	-4.1	0.3	-2.1	-5.5	-3.9	1.8
V	1.4	1.7	-2.6	-2.9	-5.1	1.0	-4.7		-4.9	-5.4	-2.9	-4.8	-5.8	-2.9	-4.4	-3.6	-4.3	-3.3	-2.6	1.6
W	-4.3	0.9	-5.0	-5.0	0.1	-4.3	-4.7		0.2	-3.8	-6.1	-5.4	-6.4	-2.9	-1.6	-5.3	-4.4	1.2	-4.7	1.0
Y	-4.3	1.7	-4.5	-4.3	-0.8	-6.2	-1.4		0.2	0.1	-0.9	-0.7	-5.1	-1.0	0.1	-5.1	-2.1	1.0	-1.4	-0.2
*	-3.8	-1.8	-5.4	-4.2	-3.0	-5.4	-4.8		-2.5	-1.6	-5.3	-5.2	-1.0	-6.5	-1.5	-5.8	-4.9	-4.5	-5.1	-1.9

Supplementary Figure 3. **scFv library $\log_2$ (fold change) matrices**. **A.** Log fold change of scFv residues tolerated at each position of CH235 UCA CDRH3 after selection with immunogen CH505.M5.G458Y. **B.** Log fold change of scFv residues tolerated at each position of DH270 UCA CDRH3 after selection with immunogen 10.17DT. Wild type residues shown in yellow. * Stop codon.

L104 Mutant Mab	UCA Normalized
UCA (L)	1.00
A	1.05
C	1.03
D	1.04
E	1.00
F	0.95
G	0.98
H	1.01
I	1.04
K	0.96
M	1.09
N	1.07
P	0.00
Q	1.02
R	1.06
S	1.04
T	1.02
V	1.02
W	1.00
Y	1.00



Supplementary Figure 4. **Binding characterization of DH270 UCA mutant antibody variants at position 104**. Single site mutated antibodies containing every possible substitution residue at position 104 of DH270 UCA were recombinantly produced and tested for binding to 10.17DT target immunogen. Results are reported as logAUC, normalized to DH270 UCA binding signal.

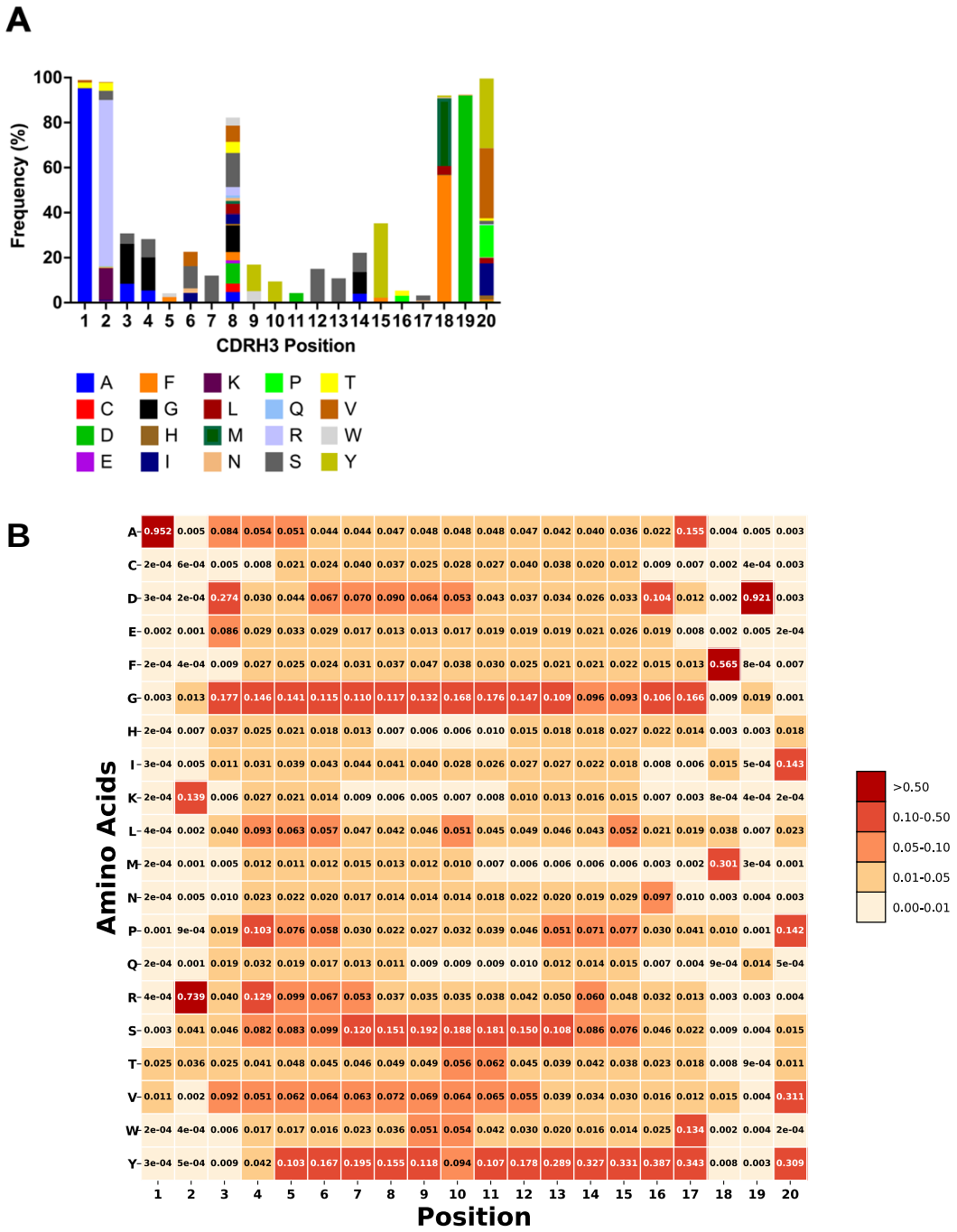
A

Subject	Total Sequences	Number of Mismatches			
		0	1	2	3
316188	1900705	0	0	2	4
326650	6293163	0	0	4	20
326651	21296713	0	1	3	58
326713	29924389	0	0	3	50
326737	3607616	0	0	0	3
326780	6205036	0	0	0	7
326797	5231004	0	0	0	14
326907	2765819	0	0	0	2
327059	8231433	0	0	2	15
D103	2712449	0	0	0	2
Total	85149055	0	1	14	175
Mean	8816832.7	0	0.1	1.4	17.5

B

Subject	Total Sequences	Number of Mismatches			
		0	1	2	3
316188	1900705	10144	37176	60290	55617
326650	6293163	36255	123980	191255	170085
326651	21296713	137111	452280	667683	560597
326713	29924389	242883	752695	1055954	851037
326737	3607616	18751	66658	109941	105155
326780	6205036	41011	138510	204896	178671
326797	5231004	28756	99795	157769	147596
326907	2765819	13280	49304	82497	79685
327059	8231433	57800	176557	252871	211151
D103	2712449	14624	51400	81101	75644
Total	85149055	600615	1948355	2864257	2435238
Mean	8816832.7	60061.5	194835.5	286425.7	243523.8

Supplementary Figure 5. **Number of natural CDRH3 predicted to be recognized by the candidate immunogens**. **A.** The total number of sequences isolated together with the number of CDRH3 sequences that contain 0, 1, 2, or 3 mismatches relative to the DH270 UCA CDRH3 substitution profile recognized by 10.17DT. **B.** The total number of sequences isolated together with the number of CDRH3 sequences that contain 0, 1, 2, or 3 mismatches relative to the CH235 UCA CDRH3 substitution profile recognized by CH505.M5.G458Y.



Supplementary Figure 6. Positional Frequency of 10.17DT Compatible Amino Acids in the BCR Database. A. The amino acid frequency at each position in 20 amino acid long CDRH3 sequences from the BCR database estimated to be compatible with 10.17DT binding based on a $\log_2$ enrichment of greater than -0.2 in the DH270 UCA3 CDRH3 substitution profile. Positions 5, 16, and 17 are highly restricted relative to the amino acids observed in these positions in the BCR database. **B**. The frequency of all amino acids at each position in 20 amino acid long

CDRH3s in the BCR database. Immunogens that can bind and tolerate alternative high frequency amino acids at position 5 (e.g. Gly, Tyr, and Ser), at position 16 (e.g. Tyr, Gly, and Asp), and at position 17 (e.g. Tyr, Gly, and Ala) would greatly improve the overall frequency of compatible CDRH3s in the BCR repertoire.

PTID	Sequence	Mismatch	Score
316188	ARAGFITYYYDSSGYCGFDY	2	19.7451016
326650	ARASRVTYYYDSSGYADFDY	3	19.8713869
326650	ARGSRVTYYYDSSGYADFDY	3	20.19291
326797	ARSGWSTYYYDSSGYHGFDY	3	20.3358721
326713	ATASRLTYYYDSSGYPAFDY	3	20.7481279
326651	ARAPRITYYYDSSGYPNLDY	3	21.0307618

Supplementary Figure 7. The sequence and score of the 100 natural CDRH3 sequences that best match the substitution profile of DH270 UCA towards 10.17DT binding. Chimeric DH270 UCA antibodies containing these CDRH3 loops were expressed as an scFv library on the surface of yeast and assayed for binding to 10.17DT immunogen.

Acknowledgements

We would like to thank Sravani Venkatayogi for bioinformatics support. We would like to thank Duke University and OIT Research Computing for providing computational resources and data storage through the Duke Compute Cluster. We are grateful for experimental support and sample analysis from the Duke Cancer Institute Flow Cytometry facility, the Duke Human Vaccine Institute Flow Cytometry core facility, the Duke Human Vaccine Institute Biomolecular Interaction Analysis (BIA) core facility and the Duke Human Vaccine Institute Viral Genetic Analysis core facility.

Methods

Plasmids and DNA synthesis. CDRH3 single site mutagenesis libraries were synthesized on a BioXp system (CodexDNA). For CH235 UCA CDRH3 libraries, heavy chain residues 95-107 encoded by the D gene, the N-nucleotide addition, as well as the first three amino acids of the JH gene (amino acid residues 95-107) were allowed to sample all the 20 amino acids. For DH270 UCA CDRH3 libraries, heavy chain residues 97-116 encoded by the last two amino acids of the VH gene, the D gene, the N-nucleotide additions, and the first three residues of the JH gene were allowed to sample all the 20 amino acids. Genes encoding the antibody heavy and light chains were commercially synthesized and cloned into pcDNA3.1 vector (GenScript). DNA primers for sequencing and insert amplification were ordered from IDT.

Development and screening of scFv libraries on the surface of yeast. *Library design and synthesis.* Single site saturation mutagenesis libraries were synthesized as described above. Library scFvs were then amplified with Q5 polymerase and purified by agarose gel extraction and PCR cleanup (Qiagen) as per the manufacturer's protocol.

Library transformation into S. Cerevisae. *S. cerevisiae* EBY100 yeast cells were transformed to express the CDRH3 single site saturation mutagenesis libraries as previously described[39, 43, 44]. Cells were transformed by electroporation with a 3:1 ratio of 12µg scFv library DNA and 4µg pCTCON2 plasmid digested with BamHI, SalI, NheI (NEB). The typical sizes of the transformed libraries, determined by serial dilution on selective plates, ranged from 2-6x10^7. Between 60% and 80% of the sequences recovered from the transformed libraries were confirmed to contain full length, in-frame genes by Sanger sequencing (Genewiz). Yeast libraries were grown in SDCAA media (Teknova) supplemented with pen-strep at 30C and shaking (225 rpm).

Library screening by FACS. scFv expression on the surface of yeast was induced by culturing the libraries in SGCAA (Teknova) media at a density of 1x10^7 cells/mL for 24-36 hours. Cells were washed twice in ice cold PBSA (0.01M sodium phosphate, pH 7.4, 0.137M sodium chloride, 1g/L bovine serum albumin) and incubated for 1 hour at 4C with 300nM biotinylated immnunogens expressed as HIV-1 Env SOSIPs (10.17DT for the DH270 library and CH505_M5_G458Y for the CH235 library). Cells were then washed twice with PBSA, resuspended in secondary labeling reagent α c-myc:FITC (ICL) and streptavidin:R-PE (Sigma-Aldrich) and incubated at 4C for 30 minutes. Cells were washed twice with PBSA after incubation with the fluorescently labeled probes and sorted on a FACS-DiVa (BD). Double positive cells for PE and FITC were collected and expanded for one week in SDCAA media supplemented with pen-strep before DNA isolation from selected clones. FACS data was analyzed with Flowjo_v10.6 software (Becton, Dickinson & Company). All clones selected by FACS were expanded, and their DNA was extracted (Zymo Research) for analysis by Next Generation Sequencing (Illumina) and Sanger sequencing (Genewiz).

Sequence analysis of isolated library clones. ScFv encoding plasmids were recovered from yeast cultures by miniprep with the Zymoprep yeast plasmid miniprep II kit

(Zymo Research) as previously described[39]. Isolated DNA was transformed into NEB5α *E. coli* (NEB) and the DNA of individual bacterial colonies was isolated (Wizard Plus SV Minipreps, Promega) and analyzed by Sanger sequencing. To prepare for Next Generation Sequencing, the scFv insert from isolated plasmids was amplified by PCR using Q5 polymerase (NEB). DNA samples were prepped and run using the Illumina MiSeq v3 reagent kit following manufacturer's protocols. Illumina sequencing returned an average of 21.6 million reads per sample, of which an average of 20.7 million mapped to the scFv amplicon. Sequencing data was processed using Geneious Prime and in-house scripts to compute the amino acid frequency and distribution.

Natural human BCR repertoire analysis. We downloaded locally the BCR sequence database developed by Briney et al., which was constructed from deep sequencing 10 individuals' BCR repertoires[30]. We deduplicated all VDJ sequences per subject and discarded any VDJ sequences that contained premature stop codons. The resulting dataset was comprised of 85,149,053 sequences. Frequency matrices of the naïve (N) and sorted experimental libraries (S) were created using the amino acid counts for each position as determined by NGS sequencing. A log fold change matrix was subsequently calculated using $F_{ij}=\log_2(S_{ij}/N_{ij})$ for each substitution *i* at each position *j*. The curated BCR sequence database was queried first for sequences with the same CDRH3 length, same D gene, D gene reading frame and position in the case of DH270 UCA, or by same CDRH3 length in the case of CH235 UCA. Sequences that met these criteria were then measured against the CDRH3 amino acid substitution profile of their target antibody by considering any amino acid with a log fold change of -0.2 or higher (corresponding to 13% reduction in binding or less compared to the UCA) as a match (Supplementary Figure 3). Additionally, heatmaps were generated by varying the binding thresholds used to define a match/mismatch and computing the frequency of sequences within the input dataset that occurred over a range of distances from 0 to N mismatches where N is the length of the CDRH3. For production of chimeric DH270 UCA antibodies with the most DH270-

like CDRH3s from the BCR database, we first selected CDRH3 sequences with the same length and then scored each sequence by summing the inverse of the fold change value over each position in the CDRH3 and selecting the top 100 scoring CDRH3s for experimental characterization.

CDRH3 sequence simulation. We employed the IGoR program to simulate VDJ recombination and generate synthetic CDRH3 sequences[42]. VDJ sequences were generated without using the IGoR error model and thus represent unmutated sequences. As sequences were generated, we stored only CDRH3s that matched the criteria for being DH270 UCA-like: same length, same D gene, same D gene reading frame and same D gene position. This resulted in $5x10^{12}$ total generated sequences with $3.3x10^{11}$ CDRH3 sequences of the same length as DH270 UCA and $1.15x10^{9}$ CDRH3 sequences with the same length, same D gene, same D gene reading frame and same D gene position.

Antibody expression and purification. Antibodies were expressed and purified as previously described[23, 39]. Briefly, 100mL cultures of Expi293F cells at a density of $2.5x10^{6}$ cells/mL were transiently transfected with 50µg each heavy and light chain encoding plasmids and Expifectamine (Invitrogen) per manufacturer's protocol. Five days after transfection, cell culture media was cleared of cells by centrifugation, and the supernatant was filtered through a 0.8µm filter (Nalgene). Clarified supernatant was incubated with Protein A beads (ThermoFisher) over night at 4C, washed with 20mM Tris supplemented with 350mM NaCl (pH=7), followed by elution with a 2.5% Glacial Acetic Acid Elution Buffer and subsequent buffer exchange into 25mM Citric Acid supplemented with125mM NaCl (pH=6). IgGs expression was confirmed by reducing SDS-PAGE analysis and quantified by measuring absorbance at 280nM (Nanodrop 2000).

Recombinant HIV-1 Env SOSIP production. SOSIP envelopes were produced recombinantly as previously described[23]. Briefly, Freestyle 293 cells were transfected with

293Fectin complexed with envelope-expressing DNA and furin-expressing plasmid DNA. After 6 days, SOSIPS were purified via PGT145 affinity chromatography with subsequent size exclusion chromatography. Trimeric HIV-1 Env fractions were pooled, flash-frozen, and stored at -80C in 10mM Tris pH 8, 500mM NaCl buffer.

Binding analysis by Surface Plasmon Resonance. 10.17DT binding of DH270 UCA antibody variants containing single site mutations (Figure 2E) as binding of DH270 UCA chimeric antibodies containing CDRH3 loops identified from the natural BCR database (Figure 4) was analyzed by Surface Plasmon Resonance on a BIAcore 3000 instrument (GE Healthcare). Biotinylated SOSIP Env 10.17DT was immobilized on a CM5 chip coated with streptavidin. Antibodies were then injected at different concentrations to determine the maximum (Rmax) binding response to 10.17DT. Background binding levels to 10.17DT obtained by buffer alone and by a control non-HIV antibody (palivizumab) were subtracted from the measured response. For measurements of the single mutant DH270 UCA3 antibodies (Figure 2E), 3000-4000 RUs of 10.17DT were captured and antibodies were injected at concentrations of 100nM and 1µM. For measurements of the DH270 UCA chimeric antibodies (Figure 4), 2000-3000 RUs of 10.17DT were captured and antibodies were injected at concentrations of 20µM and 2µM. Data to determine the Rmax value was analyzed using the BIAevaluation 4.1 software (GE Healthcare). Rmax values were reported as relative signal to that obtained by the interaction of WT DH270 UCA antibody to 10.17DT.

Binding analysis by ELISA. Binding to 10.17DT by DH270 UCA single mutant antibodies at position L104 (Supplementary Figure 4) and by a subset of DH270 UCA3 antibody variants containing single site mutations (Figure 2E) was measured by ELISA. The target antibody was captured on 96 well plates overnight, and then incubated with serial dilutions (1:3) of 10.17DT SOSIP, starting at a concentration of 100µg/mL. Binding was detected by incubation with biotinylated antibody PGT151, that is specific Env trimers, followed by addition of

streptavidin-HRP. Plates were read at 450nm on a SpectraMax 384 PLUS reader (Molecular Devices). The logarithm of the area under the curve (LogAUC) was calculated using Prism 8.

References

1. Abbott, R.K. et al. Precursor Frequency and Affinity Determine B Cell Competitive Fitness in Germinal Centers, Tested with Germline-Targeting HIV Vaccine Immunogens. *Immunity* **48**, 133-146.e136 (2018).
2. Dosenovic, P. et al. Anti-HIV-1 B cell responses are dependent on B cell precursor frequency and antigen-binding affinity. *Proc Natl Acad Sci U S A* **115**, 4743-4748 (2018).
3. Bonsignori, M. et al. Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. *Journal of virology* **85**, 9998-10009 (2011).
4. Bonsignori, M. et al. Staged induction of HIV-1 glycan-dependent broadly neutralizing antibodies. *Sci Transl Med* **9** (2017).
5. Doria-Rose, N.A. et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509**, 55-62 (2014).
6. Lorenzi, J.C.C. et al. Neutralizing Activity of Broadly Neutralizing anti-HIV-1 Antibodies against Primary African Isolates. *Journal of virology* **95**, e01909-01920 (2020).
7. Steichen, J.M. et al. HIV Vaccine Design to Target Germline Precursors of Glycan-Dependent Broadly Neutralizing Antibodies. *Immunity* **45**, 483-496 (2016).
8. Walker, L.M. et al. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* **326**, 285-289 (2009).
9. Walker, L.M. et al. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477**, 466-470 (2011).
10. Corti, D. et al. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333**, 850-856 (2011).
11. Joyce, M.G. et al. Vaccine-Induced Antibodies that Neutralize Group 1 and Group 2 Influenza A Viruses. *Cell* **166**, 609-623 (2016).
12. Stadlbauer, D. et al. Broadly protective human antibodies that target the active site of influenza virus neuraminidase. *Science* **366**, 499-504 (2019).
13. Whittle, J.R. et al. Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proc Natl Acad Sci U S A* **108**, 14216-14221 (2011).
14. Li, D. et al. In vitro and in vivo functions of SARS-CoV-2 infection-enhancing and neutralizing antibodies. *Cell* **184**, 4203-4219.e4232 (2021).
15. Martinez, D.R. et al. A broadly cross-reactive antibody neutralizes and protects against sarbecovirus challenge in mice. *Sci Transl Med* **0**, eabj7125 (2021).
16. Starr, T.N. et al. SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature* **597**, 97-102 (2021).
17. Tortorici, M.A. et al. Broad sarbecovirus neutralization by a human monoclonal antibody. *Nature* **597**, 103-108 (2021).
18. Corbett, K.S. et al. Design of Nanoparticulate Group 2 Influenza Virus Hemagglutinin Stem Antigens That Activate Unmutated Ancestor B Cell Receptors of Broadly Neutralizing Antibody Lineages. *mBio* **10**, e02810-02818 (2019).
19. Impagliazzo, A. et al. A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. *Science* **349**, 1301-1306 (2015).
20. LaBranche, C.C. et al. Neutralization-guided design of HIV-1 envelope trimers with high affinity for the unmutated common ancestor of CH235 lineage CD4bs broadly neutralizing antibodies. *PLOS Pathogens* **15**, e1008026 (2019).
21. Lin, Y.R. et al. HIV-1 VRC01 Germline-Targeting Immunogens Select Distinct Epitope-Specific B Cell Receptors. *Immunity* **53**, 840-851 e846 (2020).

22. Martinez, D.R. et al. Chimeric spike mRNA vaccines protect against Sarbecovirus challenge in mice. *Science* **373**, 991-998 (2021).
23. Saunders, K.O. et al. Targeted selection of HIV-specific antibody mutations by engineering B cell maturation. *Science* **366** (2019).
24. Steichen, J.M. et al. A generalized HIV vaccine design strategy for priming of broadly neutralizing antibody responses. *Science* **366** (2019).
25. Tian, M. et al. Induction of HIV Neutralizing Antibody Lineages in Mice with Diverse Precursor Repertoires. *Cell* **166**, 1471-1484 e1418 (2016).
26. Wohlbold, T.J. et al. Vaccination with soluble headless hemagglutinin protects mice from challenge with divergent influenza viruses. *Vaccine* **33**, 3314-3321 (2015).
27. Yassine, H.M. et al. Hemagglutinin-stem nanoparticles generate heterosubtypic influenza protection. *Nat Med* **21**, 1065-1070 (2015).
28. Saunders, K.O. et al. Neutralizing antibody vaccine for pandemic and pre-emergent coronaviruses. *Nature* **594**, 553-559 (2021).
29. Regep, C., Georges, G., Shi, J., Popovic, B. & Deane, C.M. The H3 loop of antibodies shows unique structural characteristics. *Proteins* **85**, 1311-1318 (2017).
30. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D.R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393-397 (2019).
31. Wu, T.T., Johnson, G. & Kabat, E.A. Length distribution of CDRH3 in antibodies. *Proteins: Structure, Function, and Bioinformatics* **16**, 1-7 (1993).
32. Zemlin, M. et al. Expressed Murine and Human CDR-H3 Intervals of Equal Length Exhibit Distinct Repertoires that Differ in their Amino Acid Composition and Predicted Range of Structures. *Journal of Molecular Biology* **334**, 733-749 (2003).
33. DeKosky, B.J. et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* **113**, E2636-2645 (2016).
34. Arnaout, R. et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* **6**, e22365-e22365 (2011).
35. Bassing, C.H., Swat, W. & Alt, F.W. The Mechanism and Regulation of Chromosomal V(D)J Recombination. *Cell* **109**, S45-S55 (2002).
36. Schatz, D.G. & Ji, Y. Recombination centres and the orchestration of V(D)J recombination. *Nature Reviews Immunology* **11**, 251-263 (2011).
37. Briney, B.S., Willis, J.R. & Crowe, J.E., Jr. Human Peripheral Blood Antibodies with Long HCDR3s Are Established Primarily at Original Recombination Using a Limited Subset of Germline Genes. *PLoS One* **7**, e36750 (2012).
38. Ivanov, I.I. et al. Development of the Expressed Ig CDR-H3 Repertoire Is Marked by Focusing of Constraints in Length, Amino Acid Use, and Charge That Are First Established in Early B Cell Progenitors. *The Journal of Immunology* **174**, 7773 (2005).
39. Swanson, O. et al. Rapid selection of HIV envelopes that bind to neutralizing antibody B cell lineage members with functional improbable mutations. *Cell Reports* **36** (2021).
40. Starr, T.N. et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e1220 (2020).
41. Chan, K.K. et al. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261-1265 (2020).
42. Marcou, Q., Mora, T. & Walczak, A.M. High-throughput immune repertoire analysis with IGoR. *Nat Commun* **9**, 561 (2018).
43. Benatuil, L., Perez, J.M., Belk, J. & Hsieh, C.M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng Des Sel* **23**, 155-159 (2010).

44. Chao, G. et al. Isolating and engineering human antibodies using yeast surface display. *Nature Protocols* **1**, 755-768 (2006).