bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Title: Hiding in plain sight: structure and sequence analysis reveals the importance of the antibody DE loop for antibody-antigen binding.

Keywords: Antibody DE loop; Antibody structure; Antibody binding; HIV-1 bnAbs; high-throughput antibody sequences; structural bioinformatics; Protein backbone clustering

Authors: Simon P. Kelow[1,2], Jared Adolf-Bryfogle[3], Roland L. Dunbrack, Jr.[1*]

[1] Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, PA 19111 USA

[2] Dept. of Biochemistry and Molecular Biophysics, University of Pennsylvania, Philadelphia, PA 19143 USA

[3] The Scripps Research Institute, La Jolla, CA 92037 USA

* - Corresponding Author

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Antibody variable domains contain "complementarity determining regions" (CDRs), which are solvent exposed loops that form the antigen binding site. Three such loops, CDR1, CDR2, and CDR3, are recognized as the canonical CDRs. However, there exists a fourth solvent-exposed loop, the DE loop, adjacent to CDR1 and CDR2 that joins the D and E strands on the antibody v-type fold. The DE loop is usually treated as a framework region, and as such, structural and genetic studies of antibodies often ignore this loop; yet, their lengths, structures, and sequences are variable and they contact the antigen in some antigen-antibody complex structures. We analyzed all of the structures and sequences of DE loops, which we refer to as H4 and L4 in the heavy and light chain variable domains respectively, as well as searched through millions of antibody sequences from both HIV-1 infected and naïve patients to look for human DE loop sequences with interesting features. Clustering the backbone conformations of the most common length of L4 (6 residues) reveals four dominant conformations, two of which contain only κ light chains, one of which contains only λ light chains, and one of which contains both κ and λ light chains. H4 loops in mammalian germlines are all of length 8 and their structures exist in only one conformational cluster. Length-8 L4 CDRs from a subset of λ5/λ6 germlines all have a backbone conformation very similar to that of the H4 length 8 cluster. Our structural classification of the DE loop uncovers its influence on CDR1 and CDR2 conformations, which in turn affect antibody binding. Furthermore, we show that H4 sequence variability exceeds that of the antibody framework in somatically mutated sequences from naïve human high-throughput sequences, and both L4 and H4 sequence variability from λ and heavy germline sequences also exceed that of germline framework regions. Finally, we identified a variety of insertions in DE loops present in dozens of structures of broadly neutralizing HIV antibodies in the PDB, as well as antibody sequences from high-throughput sequencing studies of HIV-infected individuals, thus illuminating a possible role in humoral immunity to HIV-1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

Antibodies utilize three hypervariable loops on each variable domain to bind antigens. These three loops are referred to as complementarity determining regions or CDRs, and were first identified by their high sequence variation relative to the rest of the variable domain sequence (1). However, there is a fourth loop, referred to as the DE loop, which joins strands D and E in the immunoglobulin v-type fold (2,3) (Figure 1A). This DE loop is adjacent to CDR1 and CDR2 on the antibody heavy and light chains (Figure 1B and 1C). In the linear sequence, the DE loop sits between CDRs 2 and 3 and is encoded by V-region gene segments (4). The DE loop has been traditionally considered part of the antibody framework, so studies addressing the ability of specific DE loop residues to affect antibody binding (5–7) have addressed these residues as framework residues, and not part of a CDR-like loop. These studies made a couple of important observations about DE loop residues. Chothia and Lesk first noted interactions of the DE loop with both L1 and H1, in particular the Arg side chain at IMGT position 80 (Chothia residue 66) in the light chain that interacts with hydrophobic residues in L1 (5). Tramontano et al. noted that an Arg residue at the same position in the heavy chain (Chothia residue 71) makes hydrogen bonds to H2, stabilizing specific H2 conformations (6). Foote et al. demonstrated that antibodies lose binding affinity to target antigen upon mutation of light-chain IMGT residue Tyr87 (Chothia residue 71) to alanine, noting that this interaction mediates interaction of L1 with target antigen though a hydrogen bond between Tyr87 and Asn37 (7).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

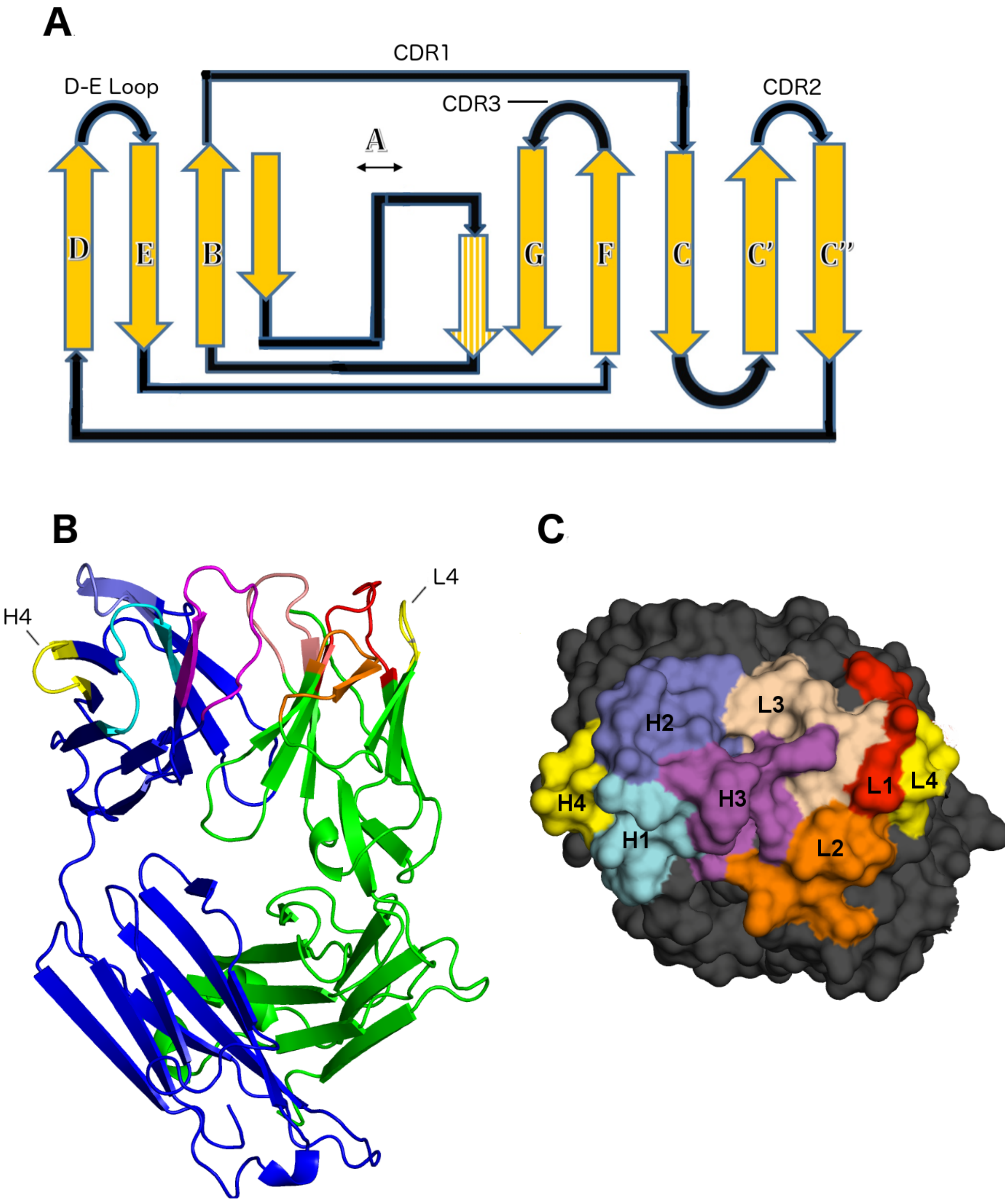

Figure 1: Position of the DE loop in antibody structures. A. V-type fold according to Bork et. al. (2). Strand A forms beta-strand pairing interactions with both strand A and strand G. **B.** Example of antibody Fab fragment (light chain in green, heavy chain in blue). **C.** Top-down view of antibody combining site. The canonical CDRs and the DE loop are marked in panel C and are represented in the same colors in panel B.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Previously we demonstrated the importance of the DE loop in redesigning an unstable anti-EGFR antibody, C10 and its affinity-matured form P2224 (8). Because the VL region of C10 appeared to be a fusion of λ3 and λ1 gene loci, introduced most likely through PCR amplification, we redesigned the antibody framework in an attempt to stabilize the antibody and prevent antibody aggregation by grafting the sequences of the λ antibody L1, L2, and L3 CDRs onto a κ framework. We observed that the λ DE loop was different in structure and sequence from a typical κ DE loop in most antibodies. Grafting the DE loop along with L1, L2, and L3 from the P2224 λ antibody onto a κ framework produced an antibody with significantly increased thermostability, which also retained P2224's binding affinity. As a control, grafting L1, L2, and L3 while keeping the host κ DE loop sequence produced an antibody with lower stability and significantly reduced affinity.

In this paper, we analyze the structures and sequences of the DE loops of heavy and light chain variable domains in the Protein Data Bank (PDB), along with a large set of sequences from multiple high-throughput antibody sequencing studies. We first define the DE loop (which we refer to as L4 on the light chain, and H4 on the heavy chain) as IMGT residues 80-87 from Ramachandran maps of residues 77-90 of heavy and light chains in the PDB. With these definitions, we expand on the observations presented by Lehmann et al. by clustering the backbone conformations of L4 and H4 loops in the structures of antibodies in the PDB to address their structural contribution to antigen binding. All human and mouse germline H4 loops are of length 8. While the vast majority of L4 loops are of length 6, with the exception of human λ5 and λ6 and mouse λ4-λ8 L4 loops, which are length 8. From a clustering of the conformations of L4 and H4, we demonstrate that L4 loops of length 6 exist in four dominant conformations, two of which only contain κ antibodies, one of which contains only λ antibodies, and one of which contains both κ and λ antibodies. The heavy chain H4 and light chain

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

L4 of length 8 share one very similar dominant conformation. We also calculate all hydrogen bond interactions between the DE loop and the CDRs, listing all of the common backbone/backbone and side-chain/backbone hydrogen bonds to backbone or side-chain atoms of L1, H1, and H2 that influence the conformation of these CDRs. We also correlate the structural features with antibody germline identity as defined by the IMGT database.

Finally, we show an analysis of the structures and sequences of antibodies related to broadly neutralizing antibodies (bnAbs) which neutralize HIV-1 in human patients (9–45). From the structures of HIV-1 bnAbs, we identify structures with insertions in DE loops on both the light and heavy chain which contact the antigen gp120, and compare the binding contribution of the DE loops with insertions to the rest of the CDRs in these antibodies. From sequencing studies of HIV-1 infected individuals, we identify insertions and deletions, hypersomatic mutation, and frameshift mutations in and around the DE loop region for the light and heavy chain. The insertion and frameshift mutations are not observed in a large set of naïve antibodies, and thus may represent a mechanism in humoral immunity to HIV-1.

Results

Clustering of canonical length L4 and H4 structures

To define the regions of structural variability in both H4 and L4, we plotted the φ and ψ dihedrals of the D and E strands and the residues in between for all heavy and light chains of antibodies in the PDB (Figure 2). We found that IMGT residues 77-79 and 86-90 nearly uniformly occupy the beta region of the Ramachandran map, while there is some variability in residues 80-82 and 85 of light chains and residues 81-85 of the heavy chain. So that the starting and ending residues are opposite each other in the beta strands and that the definition of H4 and L4 represent the same regions in

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

both domains, we define the DE loop as IMGT residues 80-87. Kabat and Chothia number the H4 region as residues 71-78 and L4 loops of length 6 as residues 66-71 (L4 loops of length 8 would require insertion codes, such as 68A, 68B). In the rest of this paper, we number the residues in the DE loops from 1 to N for DE loops of length N, such that L4 loops of length 6 are numbered 1-6, and L4 and H4 loops of length 8 are numbered 1-8. A mapping of our residue numbering to those of IMGT, Kabat, and Chothia is presented in Table 1 and 2.

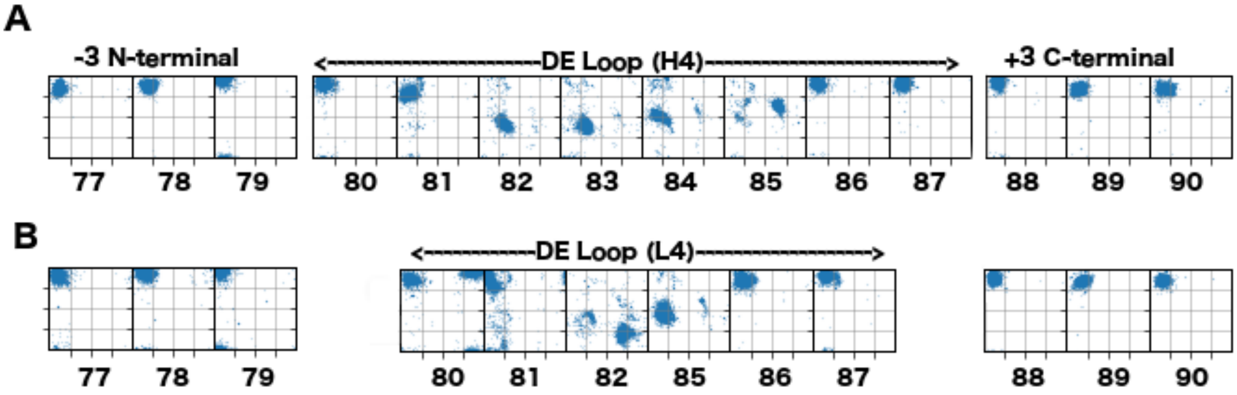

Figure 2. Ramachandran plots for part of the D strand, DE loop, and part of the E strand (IMGT residues 77-90). A. Phi (x-axis) and psi (y-axis) for residues in H4, and the 3 anchor residues before and after the loop. IMGT residue number provided at the bottom of each panel. **B.** Phi/psi for residues in L4, and the 3 anchor residues before and after the loop.

We clustered the structures of L4 loops with germline lengths 6 and 8 and H4 loops of length 8 using a maximum dihedral angle metric described in Materials and Methods. In this work, we used a density-based clustering algorithm, DBSCAN (46) to identify and remove outliers and to identify common conformations within the data (see Methods). Table 3 provides a summary of the L4 and H4 clusters, specifying their gene, consensus Ramachandran conformation, consensus sequence, number of PDBs with that length of DE loop, fraction of PDB chains which share that cluster identity, number of unique sequences, and the average φ and ψ dihedral values for each residue in the DE loop for that cluster.

Table 1. Map between various numbering schemes are residue within length 6 light chain DE loops.

Number in DE loop	IMGT	Chothia/Kabat	AHo
1	80	66	82
2	81	67	83
3	82	68	84
4	85	69	87
5	86	70	88
6	87	71	89

Table 2. Map between various numbering schemes are residue within length 8 heavy chain DE loops

Number in DE loop	IMGT	Chothia/Kabat	AHo
1	80	71	82
2	81	72	83
3	82	73	84
4	83	74	85
5	84	75	86
6	85	76	87
7	86	77	88
8	87	78	89

Table 3. DE loop canonical families

Gene	Cluster	Rama. string	Consensus sequence	# PDB chains	Percent chains	Unique seqs	1	2	3	4	5	6	7	8
κ	L4-6-1	EBEABB	GSGTD{FY}	3,302	61.1	68	120, 170	-164, 158	74, -100	-118, -11	-120, 122	-132, 154		
λ/κ*	L4-6-2	BBEABB	{KR}SGTT{AYF}	1,056	19.5	65	-144, 137	-146, 116	63, -120	-99, 9	-119, 133	-117, 148		
κ	L4-6-3	BBAABB	GS{EG}T{DA}F	106	2.0	6	157, 165	-90, -137	-94, -26	-125, -16	-123, 125	-130, 152		
λ	L4-6-4	PBEABB	LIGD{RK}A	116	2.1	7	-111, 128	-123, 110	66, -125	-99, 14	-128, 160	-88, 147		
λ/κ	L4-6-noise	-	-	759	14.0	83								
λ5/λ6	L4-8-1	BBAAALBB	ID{SRD}SSNSA	67	100.0	6	-128, 137	-121, 100	-63, -32	-66, -33	-103, 3	54, 43	-135, 157	-113, 151
-	L4-8-noise	-	-	-	0.0	-								
H	H4-8-1	BBAAALBB	-	4,041	93.0	507	-143, 149	-122, 110	-65, -30	-67, -34	-102, 2	53, 47	-129, 141	-113, 144
H	H4-8-noise	-	-	291	7.0	120								

Properties and frequencies of L4 and H4 structural families and noise clusters from the DE loop clustering results for each length of light chain and heavy chain DE loop. ϕ,ψ values (in degrees) are given for residues 1 through 6 of L4-6 clusters and 1 through 8 of L4-8-1 and H4-8-1 clusters. Ramachandran map regions are: A=alpha-helix region; B = beta sheet region; P = polyproline II helix region; E = epsilon region (lower right of Ramachandran map); L=alpha-left region.

* Cluster L4-6-2 is composed of 68% λ chains and 32% κ chains

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

For length-6 L4 structures, we observed four different clusters that are primarily related to antibody gene. Figure 3A displays the φ/ψ plots for the L4 length-6 clustering, where each row represents a cluster, and each column represents a residue within L4. This figure colors the data points by framework identity (κ in blue, λ in magenta) and shows the noise data as the bottom-most row. The four clusters partition out primarily according to gene, yielding the following clusters (ordered by decreasing size): a primary κ gene cluster (L4-6-1), a mixed λ/κ cluster (L4-6-2) that contains majority λ structures, a secondary κ (L4-6-3) cluster, and a secondary λ cluster (L4-6-4).

The primary difference between the two biggest clusters, L4-6-1 and L4-6-2, is the amino-acid identity and Ramachandran conformation of the first residue. In the germlines of all human κ light chains and nearly all mouse κ light chains, the first residue of the DE loop is glycine. In human and mouse λ germlines, the first residue is (in order of most common to least common): Lys, Ser, Ile, Asp, Leu, Arg, or Thr in 81 out of 84 human and mouse IMGT alleles and Gly in only 3 human alleles (all IGLV9-49). In L4-6-1, the first residue is in an epsilon conformation (φ=119.8°, ψ=170.2°), consistent with a Gly in κ antibodies. In L4-6-2, the first residue is in a beta conformation (φ=-144.9°, ψ=137.2°), consistent with the λ non-Gly residues. 252 out of 333 κ structures in L4-6-2 (75%) contain somatic mutations at the first residue position from Gly to Arg, Ala, Glu, and Gln in decreasing order of frequency.

L4-6-3 is an all-κ cluster that differs from all-κ L4-6-1 at positions 2 and 3, such that L4-6-1 has average (φ_2,ψ_2=-164.1°,157.5°; φ_3,ψ_3=73.6°,-100.0°) and L4-6-3 has average (φ_2,ψ_2=-90.4°,-137.0°; φ_3,ψ_3=-93.6°,-26.0°). In the L4-6-3 structures, 18 of 106 (17%) chains (from 3 PDB entries) have somatic mutations at position 3 from Gly to Glu, which moves residue 3 from an epsilon conformation to an alpha conformation (with a

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

compensating change at residue 2). The remaining structures have germline sequences, including one structure with a germline Arg residue at residue 3. The all-λ cluster L4-6-4 resembles L4-6-2, except it has a shift in φ_1 and ψ_2 of about 30° compared to L4-6-2. This is due to germline-encoded hydrophobic residues (Leu, Ile) at the first two residues of the DE loop in the L4-6-4 sequences (e.g. Mouse IGLV1*01, IGLV2*01; human IGLV7-43*01, IGLV7-46*01).

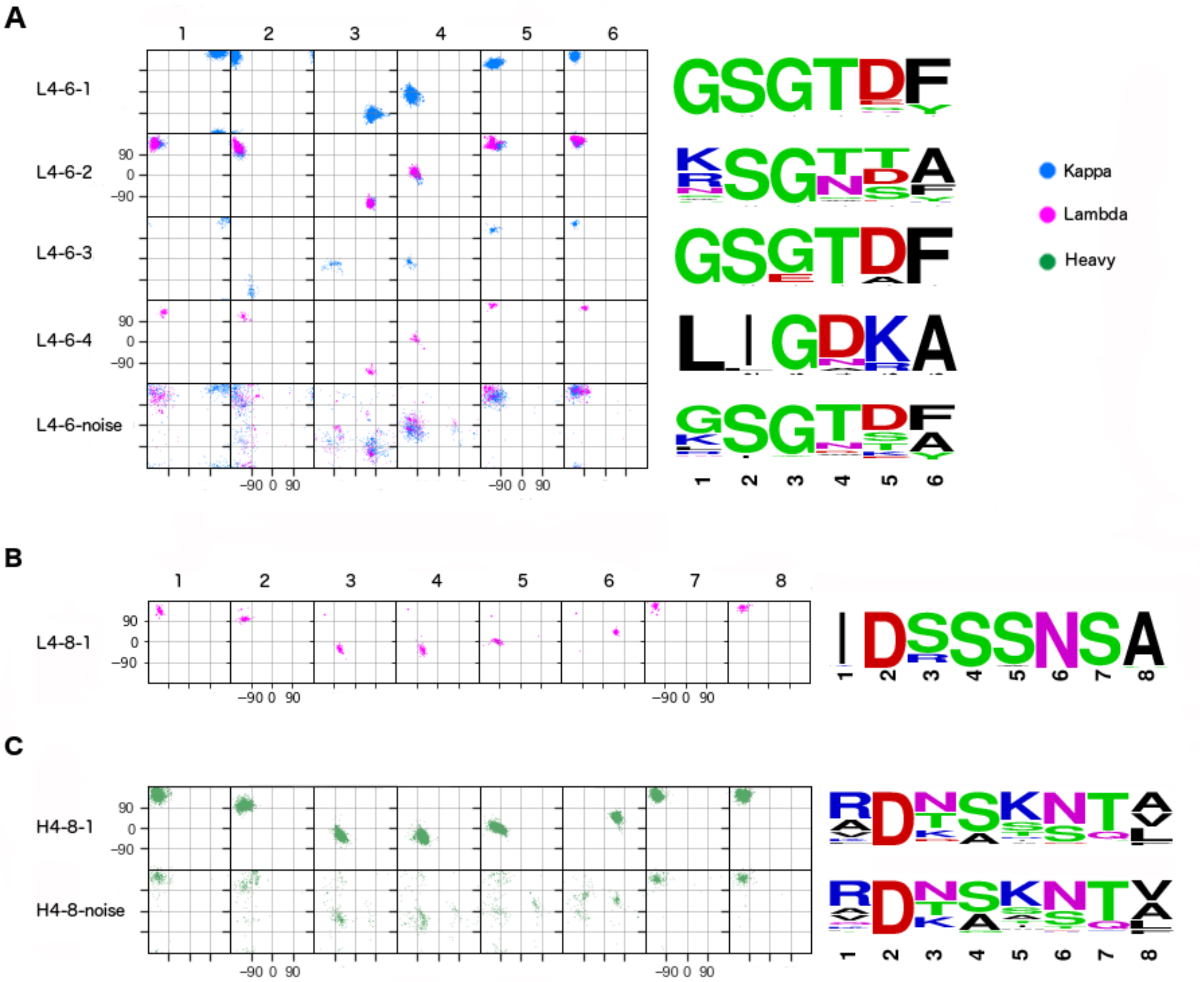

Figure 3: Canonical families of L4 and H4. Phi/Psi plots for each residue in the DE loop for each of the L4 and H4 DE loop clusters with their respective sequence logos. **A.** L4 length 6 loops. **B.** L4 length 8 loops. **C.** H4 length 8 loops.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Across the four clusters, the changes in backbone conformation may be viewed structurally as a hinge motion away from the variable domain of the antibody, with L4-6-1 being closest to the domain, and L4-6-4 the farthest away from the domain. Figure 4A shows representative structures of L4-6-1, L4-6-2, L4-6-3, and L4-6-4 DE loops superposed by alignment to the stems of the DE loop (-3 C-terminal, +3 N-terminal).

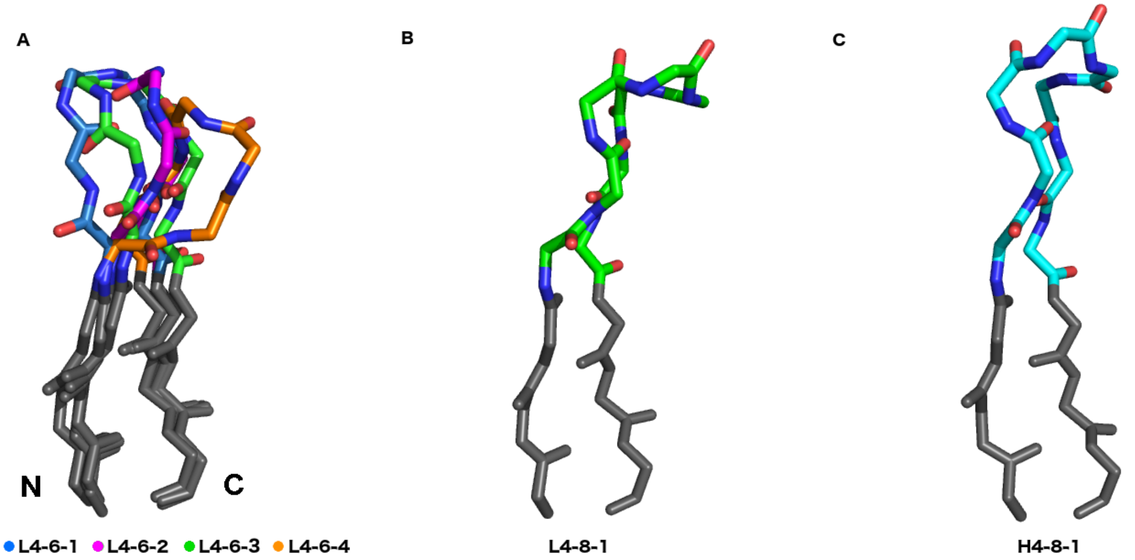

Figure 4: Structures of all DE loop canonical length clusters. A. A sticks representation of the antibody light-chain DE loop (L4-6) backbone is shown where L4-6-1 (PDB 4ebqL, blue) sits closest to the antibody domain, L4-6-3 (1mjuL, green) hinges slightly away and flips the second carbonyl of the DE loop backbone about 180 degrees relative to the other clusters, L4-6-2 (4unuA, magenta) hinges further away from the domain than L4-6-1 and L4-6-3, and L4-6-4 (5xctB, orange) sits the furthest away from the domain. The stems of the DE loop are colored dark gray. **B.** Same representation as in (A), but for the sole L4-8-1 cluster. **C.** Same representation as in (A), but for the sole H4-8-1 cluster.

For the 67 length-8 L4 structures in the PDB, which are related to a small number of λ germlines in humans, mice, rats, rabbits, and macaques, we observed a single cluster (Figures 3B and 4B) representing 8 unique sequences. No structures were placed into noise by the DBSCAN algorithm, indicating a low level of structural variance. Out 49 PDB entries containing light chains with length-8 DE loops, 17 of them are involved in Bence-Jones homodimers associated with light-chain amyloidosis (47).

For canonical length 8 H4 structures, clustering with DBSCAN produced

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

a single cluster, which we refer to as H4-8-1 (Figures 3C and 4C). Any other clusters generated in the clustering step had fewer than 4 unique sequences. The H4-8-1 cluster has 507 unique sequences, exhibiting far greater sequence variation than any of the L4-6 clusters. The conformation of length 8 H4 structures is structurally very similar to that of length 8 L4 structures as shown in structural alignment of the two clusters by the CDR4 stem (Figure 5).

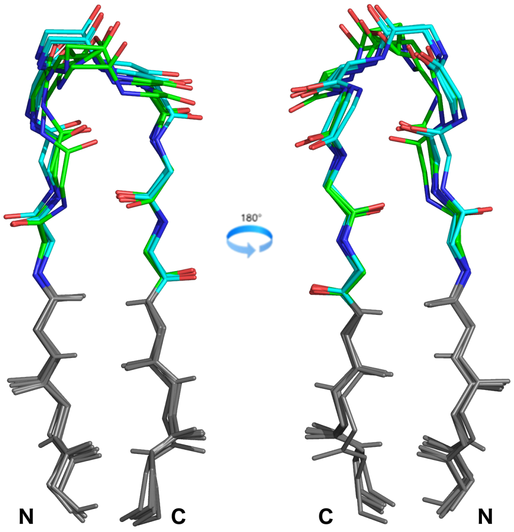

Figure 5. Comparison of H4-8-1 and L4-8-1 structures. Superposition of high-resolution heavy chain H4-8-1 structures (cyan, PDB chains: 2x1qA, 4qyoB, 2vxvH), and L4-8-1 structures (green, PDB chains: 1cd0A, 2w01A, 3h0tA) aligned by the stem of CDR4 (colored in gray) show structural homology between the two clusters.

Relationship between CDR4 conformation and conformations of CDR1 and CDR2

To describe the relationship between various L4 conformations with CDR1 and CDR2 conformations, we first calculated the occurrence of each L4-6 cluster given the various common L1 clusters (Table 4). Three κ L1 clusters are more than 98% L4-6-1: L1-10-1, L1-15-1, and L1-17-1. All of the remaining κ L1 clusters are 82-93% L4-6-1. For most of these the secondary cluster is L4-6-2, indicating a tendency for residue 1 in the corresponding germlines to mutate from Gly to another residue type. For the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

λ germlines, all of the L1 clusters except L1-14-1 are 100% L4-6-2. L1-14-1 is associated with L4-6-4, because the germlines in this cluster contain length-14 L1 loops and amino acids Leu-Ile in positions 1-2 of the DE loop.

Table 4. Occupancy of the co-occurrence of L1/L4 pairs from structures in the PDB

L1 cluster	Kappa/Lambda	# chains	L4-6-1	L4-6-2	L4-6-3	L4-6-4
L1-10-1	K	155	99.4	0.6	-	-
L1-11-1	K	1337	90.0	9.5	0.5	-
L1-11-2	K	418	91.6	8.4	-	-
L1-12-1	K	151	93.0	3.0	4.0	-
L1-12-2	K	101	90.0	9.0	1.0	
L1-15-1	K	186	98.4	-	0.4	-
L1-16-1	K	506	82.0	1.8	16.2	-
L1-17-1	K	277	99.3	0.4	0.4	-
L1-11-3	λ	118	-	100.0	-	-
L1-13-1	λ	199	-	100.0	-	-
L1-14-1	λ	111	-	-	-	100.0
L1-14-2	λ	145	-	100.0	-	-

For each L1 cluster, the distribution among the L4 clusters is provided in percent (excluding the noise cluster).

Second, we have calculated all hydrogen bonds between CDR4 and CDR1 or CDR2. Supplementary Table 1 shows all hydrogen bonds calculated between the DE loop and CDR1 or CDR2 for each CDR1 and CDR2 cluster of L1, H1, and H2 (there are no characteristic hydrogen bonds between L4 and L2 with an occupancy over 60%). The hydrogen bonds are grouped by structures with the same amino acid at the same position within the DE loop in the case of hydrogen bonds involving side-chain atoms.

Hydrogen bonds between the DE loop and CDR1 and CDR2 partition into the following primary categories: (1) backbone-backbone hydrogen bonds shared across several CDR1 clusters; (2) backbone-backbone hydrogen bonds unique to specific DE loop/CDR1 pairings; (3) hydrogen bonds between side-chain atoms in the DE loop and backbone atoms at positions shared across several CDR1 clusters; (4) hydrogen bonds between DE loop side-chain atoms and CDR1/CDR2 backbone atoms that are specific

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

for some CDR clusters/lengths; and (5) hydrogen bonds between DE loop backbone atoms and CDR1 side-chain atoms that occur in L1 loops longer than 14 residues. Figure 6 shows examples from selections of hydrogen bonds between DE loop atoms and CDR1 atoms.

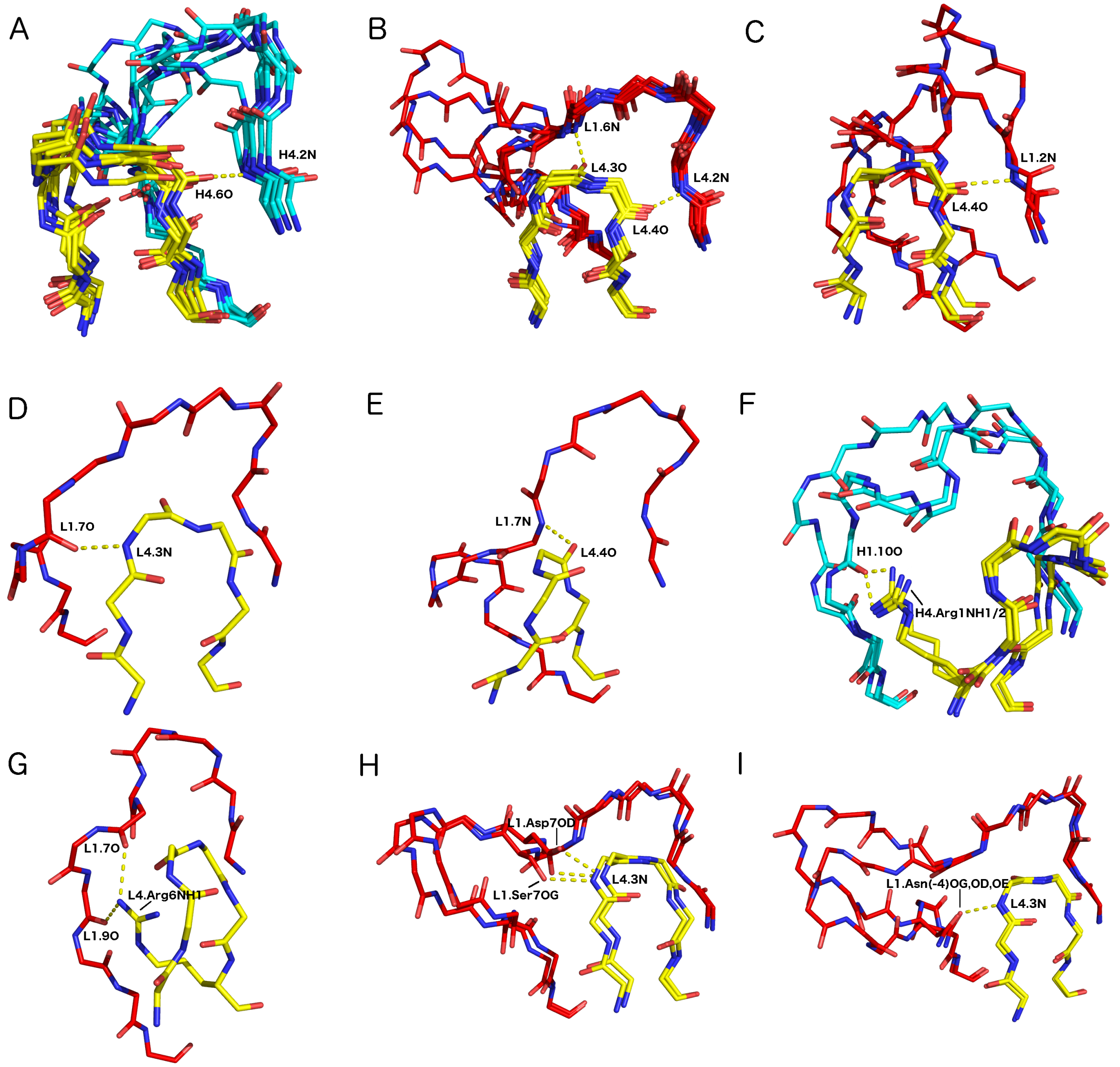

Figure 6: Various characteristic hydrogen bonds between the DE loop and CDR1. A-C. Shared backbone/backbone hydrogen bonds to H1 (H1-13-1, H1-13-2, H1-13-3, H1-13-4, H1-13-7, H1-14-1, and H1-15-1) and L1 (L1-10-1, L1-11-1, L1-11-2, L1-11-3, L1-12-1, L1-12-2, L1-13-1, L1-14-2, L1-15-1, L1-16-1, and L1-17-1) clusters. **D-E.** Unique backbone-backbone hydrogen bonds observed in in L1-11-1, and L1-14-1 clusters. **F**. Shared side-chain/backbone hydrogen bond between H4 and several (H1-13-1, H1-13-2, H1-13-3, and H1-13-4) clusters. **G**. Unique side-chain/backbone hydrogen bond to two different L1 backbone carbonyls founds only in association with L1-12-3. **H-I**. Rare L1 side-chain to DE loop

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

backbone hydrogen bonds which occur in association with longer length (L1-15-1, L1-16-1, and L1-17-1) L1 clusters.

Regardless of DE loop conformation, DE residue 4 in length 6 L4 and length-8 H4 forms a backbone-backbone hydrogen bond to the second residue's backbone nitrogen in CDR1 (counting L1 residues immediately after the Cys of the disulfide bond; Figure 6A and 6C) for the vast majority of L1 clusters (L1-14-1 excluded). This hydrogen bond is part of the beta sheet containing the C-terminal segment of CDR4 and the N-terminal strand of CDR1. On the light chain, most DE loop structures also have a backbone-backbone hydrogen bond between DE residue 3 and residue 6 of CDR1 (L1-12-3, L1-13-1, L1-14-1, and L1-14-2 excluded). Structures that have both of these hydrogen bonds have very similar conformations between the residues that are hydrogen bonded, even amongst a diverse set of L1 lengths (Figure 6B). Conversely, L4-6-4 structures do not have a hydrogen bond between DE residue 3 and L1, because the backbone of L4 is further from L1 than in other clusters, making the carbonyl oxygen of DE residue 3 unavailable. Instead, the backbone of L4 residue 4 is in place to make a hydrogen bond to residue 7 in L1-14 loops (Figure 6E). The need for this specific interaction is suggested in Table 4, as the conformation L1-14-1 occurs exclusively with L4-6-4, because the sequences are encoded in the same germlines. As noted above, L4-6-4 requires aliphatic residues in the first two residues of L4. These are only present in human and macaque IGLV7 and IGLV8 and mouse IGLV1 and IGLV2 sequences.

Beyond backbone-backbone hydrogen bonds correlated with the arrangement of the L4 backbone atoms, we note several particular side-chain/backbone hydrogen bonds that occur uniquely with L1 conformations. For example, residue R6 hydrogen bonds to the backbone carbonyl of L1-13-1 in 14/16 chains, whereas when the DE residue 6 is Lys, but is still paired with L1-13-1, the hydrogen bond occupancy is only 19%. As noted in

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

previous studies (5,48), the OH atom of the Tyr6 side chain of the DE loop forms a hydrogen bond to the backbone nitrogen atom of residue 8 in length-11 L1 CDRs, flipping its conformation from L1-11-1 (predominantly Phe6) to L1-11-2 (predominantly Tyr6). This hydrogen bond forms in 90% of structures of L1-11-2 with a Tyr residue at position 6 of L4. When this residue is Phe6 instead, this hydrogen bond is lost, and the structure of L1 is L1-11-1, and a unique hydrogen bond instead forms between the backbone nitrogen of DE residue 3 to the carbonyl of L1-11 residue 7 (Figure 6D). In similar fashion, we note a new hydrogen bond of the side chain of R6 in L4-6-2 to the carbonyl backbone oxygen atoms of residues 7 and 9 of L1-12-3 structures, creating a highly stable hydrogen bond network. This is an example where the exclusive occurrence of a L1/L4 pair is associated with a unique contact between L1 and L4.

For L1-15-1 and L1-17-1, the carbonyl oxygen of DE residue 1 of L4 is not only hydrogen bonded to a backbone nitrogen atom in L1, but the backbone nitrogen atom of DE residue 1 is also hydrogen bonded to various side-chain oxygen atoms of residue 7 in L1-15-1 (Asp, Ser, or Thr; Figure 6H), or residue 14 (Ser or Asn; Figure 6I) in L1-17-1. Taken together, these results demonstrate that L1/L4 pairs often entail highly specific interactions, facilitated by the L4 cluster-specific arrangement of the L4 backbone, and side-chain atoms in different L4 clusters, which can provide stabilizing hydrogen bonds between L4 and L1 regardless of L4 structure.

For H4, in addition to the conserved hydrogen bond involving DE residue 4 in most H1/H4 pairs for common H1 lengths and clusters, we note several side-chain/backbone hydrogen bonds that are shared between several H1 clusters and various residues in H4. Most notably, the Arg1 residue in H4 hydrogen bonds with the backbone carbonyl of residue 10 in H1-13-1 using both the NH1 and NH2 atom in the interaction (Figure 6F). For specific hydrogen bonds, the occupancy of the hydrogen bond depends

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

highly on the H4 residue type. DE residue Asn6 uses both its side-chain oxygen atom as well as its side-chain nitrogen atom to form side-chain/backbone hydrogen bonds between residue 2, residue 5 and residue 7 of H1, stabilizing the H1-14-1 conformation with a hydrogen bond network. As described in previous studies (6,48), the Arg1 side-chain forms a hydrogen bond with the backbone carbonyl of residue 3 in H2, which occurs with an occupancy of 62% in conjunction with the H2-10-2 conformation. When this residue is instead Lys1, the occupancy of this hydrogen bond is 57% (11/21 chains with H2-10-2 and H4-8-1 with Lys at position 1).

Analysis of the sequence variability in DE loops arising from somatically mutated and germline sequences.

From a set of ~2.5 million sequences of naïve human antibodies (47–52), we calculated the sequence entropy in four of the most prevalent human germlines for the heavy, κ, and λ genes in the data set (Figure 7A), as well as the entropy of human germline sequences of the same length (Figure 7B). As other studies have noted (53), variability of both framework and CDR residues depends highly upon germline. We did not find any DE loops with insertions in this set, so we did not have to account for insertions in the calculation of the sequence entropy.

For H4 sequence variability, we find that in cases of somatic mutation of any one particular germline, the average sequence entropy of H4 for each of the four germlines exceeds the average sequence entropy for FR1, FR2, and FR3 of the same germline antibodies (Table 5). However, these DE loop residues are less variable than H1 or H2 residues within the same germlines. For κ antibodies, the maximum sequence entropy for the most variable DE loop residues in L4 (residues 2 and 5) are heavily somatically mutated compared to the antibody framework (Figure 7A, Table 5), and in some germlines (e.g. IGKV3-11*01 and IGKV1-39*01) these same residues are as

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

variable as some L1 and L2 residues.

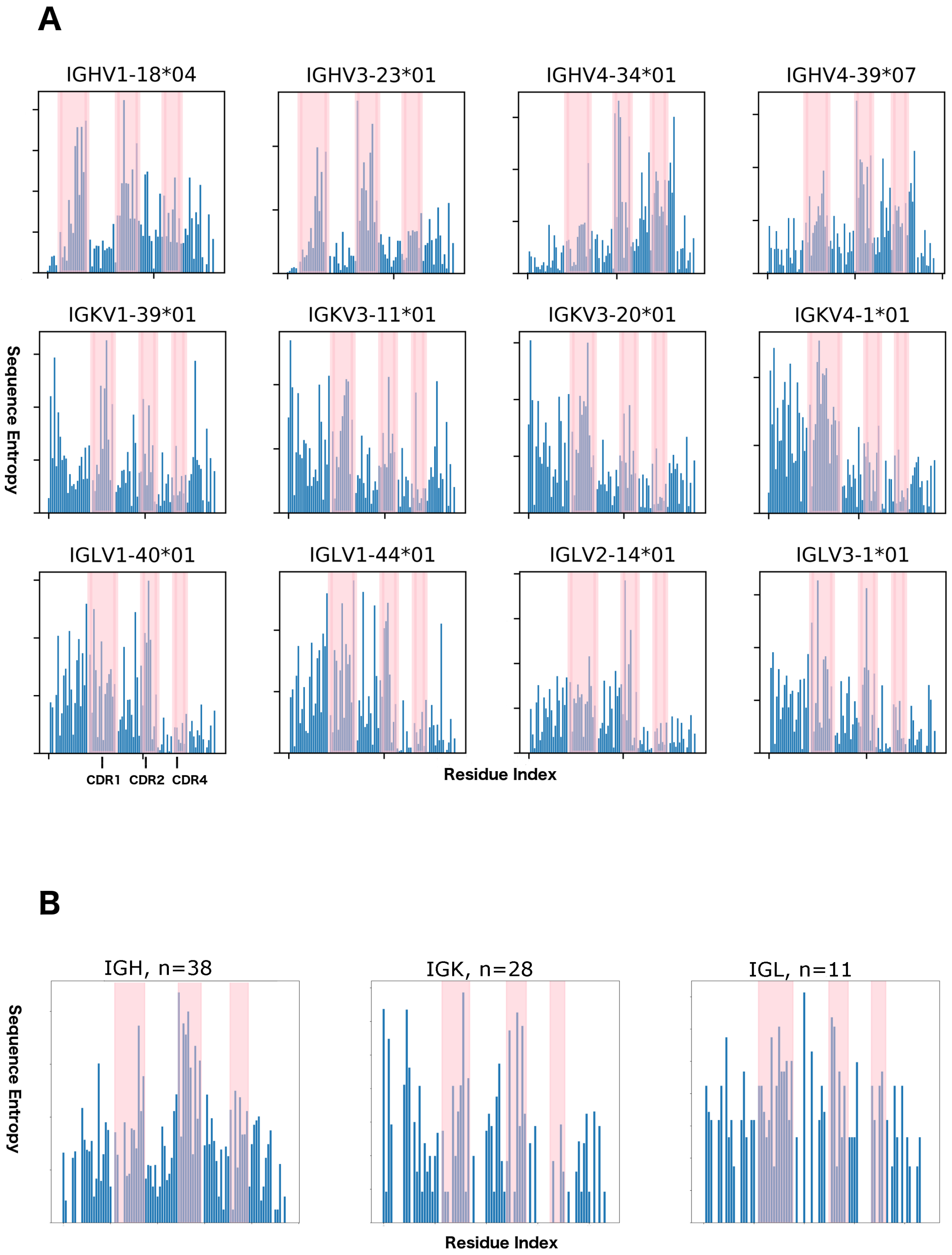

Figure 7: Sequence entropy in naïve human antibodies and human germlines. A. Sequence entropy for 12 common germlines in a naive human antibody sample (>10,000 sequences for each germline). **B.** Sequence entropy for human germlines derived from all IGKV, IGLV, and IGHV sequences from IMGT.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

From left to right, the pink shaded regions indicate CDR1, CDR2, and the DE loop. CDR3 is omitted due to varying lengths and different diversification mechanisms.

Comparing 28 germline sequences for human κ antibodies (Figure 7B), we observe three highly variable residues (DE residue 2, 5, and 6), and 3 completely conserved residues (IMGT residues Gly1, Gly3 and Thr4). In the variable residues of L4, the entropy is comparable to the most variable framework residues in germlines, the average entropy does not compare to the average entropy of L1 or L2, and does not exceed the average entropy of FR1, FR2, or FR3 (Table 5).

Table 5. Average sequence entropies for CDR and framework regions

germline	CDR1	CDR2	FR1	FR2	FR3	CDR4
IGHV1-18*04	0.38	0.44	0.04	0.17	0.21	**0.27**
IGHV3-23*01	0.42	0.74	0.03	0.15	0.21	**0.26**
IGHV4-34*01	0.14	0.38	0.05,	0.14	0.17	**0.29**
IGHV4-39*07	0.21	0.36	0.08	0.13	0.15	**0.20**
IGKV1-39*01	0.29	0.22	0.20	0.12	0.13	0.11
IGKV3-11*01	0.25	0.22	0.21	0.10	0.10	0.11
IGKV3-20*01	0.30	0.21	0.23	0.10	0.10	0.07
IGKV4-1*01	0.30	0.15	0.24	0.09	0.08	0.07
IGLV1-40*01	0.25	0.29	0.22	0.12	0.05	0.07
IGLV1-44*01	0.27	0.26	0.21	0.12	0.06	0.07
IGLV2-14*01	0.24	0.30	0.17	0.13	0.06	0.08
IGLV3-1*01	0.28	0.26	0.20	0.10	0.06	0.13
gene	**CDR1**	**CDR2**	**FR1**	**FR2**	**FR3**	**CDR4**
IGH	0.78	1.50	0.53	0.61	0.53	**0.86**
IGK	0.57	0.71	0.46	0.24	0.20	0.19
IGL	0.80	0.63	0.43	0.33	0.28	**0.52**

Average sequence entropies partitioned by CDR or framework region, excluding CDR3. Bolded values are values where the CDR4 average sequence entropy either compares to CDR1/CDR2, or exceeds the values for FR1, FR2, and FR3

Within each λ germline, L4 sequences are much less somatically mutated than in κ structures (Figure 7A). The amount of sequence variability

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

due to somatic mutation is less than even the most variable framework residues, and does not compare to sequence variability in L1 or L2. However, looking at 11 germline sequences (Figure 7B), sequence variability is comparable to both L1 and L2 at 4 of the 6 L4 residues excluding residue GLY3, and residue ALA6. Average sequence entropy in these λ L4 sequences exceeds that of FR1, FR2, and FR3, but does not compare to L1 or L2. This indicates that sequence variability in λ L4 relates primarily to germline sequence differences, and not somatic mutation. The observations for all antibody germlines show that for H4, the average sequence entropy of the DE loop residues exceeds that of the antibody framework, and also in many cases, is comparable with the variability of H1 and H2 residues. This effect depends highly on antibody germline.

For L4, in the case of somatic mutation the average sequence entropy for the whole DE loop is less than L1 and L2, as well as FR1, FR2, and FR3. However, sequence variation across various germline sequences is greater than L1, L2, FR1, FR2, and FR3 in λ L4 sequences, but lower in κ L4 germline sequences.

Non-canonical L4 and H4 length in HIV-1 bnAbs

All known mammalian VH germlines have a DE-loop of 8 residues, except for a small number of rabbit VH genes with a DE-loop of length 6 (rabbit IGHV1S17*01 and IGHV1S69*01 are represented in 42 chains in 18 PDB entries). All known mammalian VL germlines have a DE-loop of either 6 or 8 residues, except for one alpaca germline (IGLV5-12*01) with a DE loop of length 3 (not represented in the PDB). Table 6 lists the various PDB structures that have insertions in either the light or heavy chain DE loop as well as their sequences, germlines, and which bnAb class they belong to. There are 119 chains from 43 entries in the PDB with H4 loops longer than 8 amino acids, ranging from 10 amino acids to 16 amino acids. There are 65

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

chains from 50 entries that have insertions in L4 (all lambda chains), resulting in L4 loops of length 9.

Table 6. HIV-1 bnAbs with insertions in L4 and/or H4.

gene	pdb chains	length	DE loop sequence	bnAb class	germline(s)
Heavy	4toyH, 4tvpD*, 5cezD*, 5fyjD*, 5fykD*, 5fylD*, 5t3sD*, 5u7oD*, 5u7mD*, 5um8D*, 5utfD*, 5utyD*, 5v7jD*, 5w6dD*, 5wduD*, 5wduH*, 5wduM*, 5wduU*, 6ce0D*, 6ch7D, 6ch8D, 6ch9D, 6ck9D*, 6de7D*, 6ieqD, 6mcoD, 6mtjD, 6mdtD, 6mtnD, 6mu6D, 6mu7D, 6mu8D, 6mufD, 6mugD, 6nm6E, 6nnfD, 6nnjD	16	TDTEVPVTSFTSTGAA	35O22	Hu_IGHV1-8*01
Heavy	4jb9H	15	RLFSQDLYYPDRGTA	VRC06	Hu_IGHV1-24*01
Heavy	3se8H, 4cc8F, 4cc8H, 4cc8I, 5jxaH, 6cdeq*, 6cdiQ*, 6cue7, 6cueQ, 6cueq, 6cuf8, 6cufQ, 6cufq, 6e5pI, 6e5pO, 6efpV, 6mpg8, 6mpgQ, 6mpgq, 6mphQ, 6mphf, 6mphg, 6n1vQ, 6n1vf, 6n1vg, 6n1w8, 6n1wQ, 6n1wq, 6nf2C, 6nf2N, 6nf2V	15	RQLSQDPDDPDWGVA	VRC03	Hu_IGHV1-24*01
Heavy	4s1qH, 6nm6U	15	RQLSQDPDDPDWGIA	VRC03	Hu_IGHV1-24*01
Heavy	6nnfU	15	RQLSQDPDDPDWGTA	VRC03	Hu_IGHV1-8*01
Heavy	4xnzB, 4xnzE, 4xnzH	15	RQLSQDPDDPDWGVA	VRC06B	Hu_IGHV1-3*01
Heavy	4p9hH, 4p9mH, 5a7xN, 5a7xP, 5a7xR, 5a8hF, 5a8hL, 5a8hR, 5c7kE, 5cjxA, 5cjxD, 5cjxH, 5js9E, 5jsaE, 5thrP, 5thrR, 5thrT, 5viyK, 5viyM, 5viyI, 5vj6M, 5vj6O, 5vj6Q, 6cm3P, 6cm3R, 6cm3T, 6eduP, 6eduR, 6eduT, 6nqdC, 6nqdG, 6nqdK, 6osy7, 6osyF, 6osyP, 6ot1I, 6ot1S, 6ot1q	12	AVDLTGSSPPIS	8ANC195	Hu_IGHV1-69*01
Heavy	4jpvH, 4lsvH, 5v8lG, 5v8lH, 5v8lI, 5v8mH, 5v8mR, 5v8mS	12	RHASWDFDTYS	3BNC117	Hu_IGHV1-46*01
Heavy	3rpiA, 3rpiH, 4gw4A, 4gw4H	12	RQASWDFDTYSF	3BNC60	Hu_IGHV1-46*01

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Lambda	4jy6A, 4jy6C	9	PDFRPGTTA	PGT123	Hu_IGLV3-21*01
Lambda	4fq2L, 6ccbE, 6ccbL, 6ck9L*, 5ceyA, 5ceyC, 5cezL*, 5t3xL, 5t3zL, 5v7jL*, 5w6dL*, 6mcoL*, 6mdtL*, 6mtjL*, 6mtnL*, 6mu6L*, 6mu7L*, 6mu8L*, 6mufL*, 6mugL*, 6nm6L*, 6nnfL*, 6nnjL*	9	PDINFGTRA	10-1074	Hu_IGLV3-21*01
Lambda	4r26L, 4r2gC, 4r2gI, 4r2gM, 5t3sL*, 5um8L*, 6ce0L*, 6ieqL*,	9	PDINFGTTA	PGT124	Hu_IGLV3-21*01
Lambda	5cexC	9	PDSNFGTTA	32H+109L	Hu_IGLV3-21*01
Lambda	4fq1L, 4fqcL, 4jy4A	9	PDSPFGTTA	PGT121	Hu_IGLV3-21*01
Lambda	4jy5L, 4ncoG, 4ncoK, 4ncoC, 4tvpL*, 5d9qL, 5d9qE, 5d9qM, 5fyjL*, 5fykL*, 5fylL*, 5i8hJ, 5i8hL, 5u7mL*, 5u7oL*, 5utfL*, 5utyL*, 5wduB*, 5wduK*, 5wduS*, 6b0nL, 6cden*, 6cdiN*, 6cue6, 6cuen, 6cuf6, 6cufn, 6de7L*	9	PGSTFGTTA	PGT122	Hu_IGLV3-21*01

The table lists all of the antibody structures in the PDB with insertions in L4 or H4 related to bnABs, along with their sequences, germline, and bnAb lineage. Entries with an asterisk (*) represent structures with insertions in both the light and heavy chains. There is only one other antibody with an inserted DE loop in the PDB: the engineered nanobody towards Higb2 toxin in cholera virus NB6 (PDBID 5mje, DE sequence RDSAEDSAKNTV). It is not listed in the Table.

The L4 and H4 structures longer than germline lengths in the PDB are almost all related to bnAbs isolated from HIV-1 patients that were part of a series of affinity matured bnAbs targeting the HIV-1 envelope gp120-gp41 trimer. In most of these structures, the elongated H4 loops make specific hydrophobic interactions, hydrogen bonds, and salt bridges with the antigen (Figure 8). In these structures, L4 and H4 bind the antigen epitope better than two out of the three heavy chain CDRs as well as any of the light chain CDRs, as demonstrated by the extent of antigen-buried surface area for each CDR including CDR4 (Figure 9). A salt bridge between R1 of H4 and an interfacial Asp of the V1-loop of gp120 also helps stabilize the antibody-antigen interface, and is observed in 7/10 unique elongated H4 sequences. Besides this interaction, much of the buried interface creates hydrophobic

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

contacts across the antibody-antigen interface (Figure 8B). One inserted DE loop structure targets an antigen other than HIV-1 gp120, the engineered nanobody towards Higb2 toxin in cholera virus NB6 (PDBID 5mje, DE sequence RDSAEDSAKNTV).

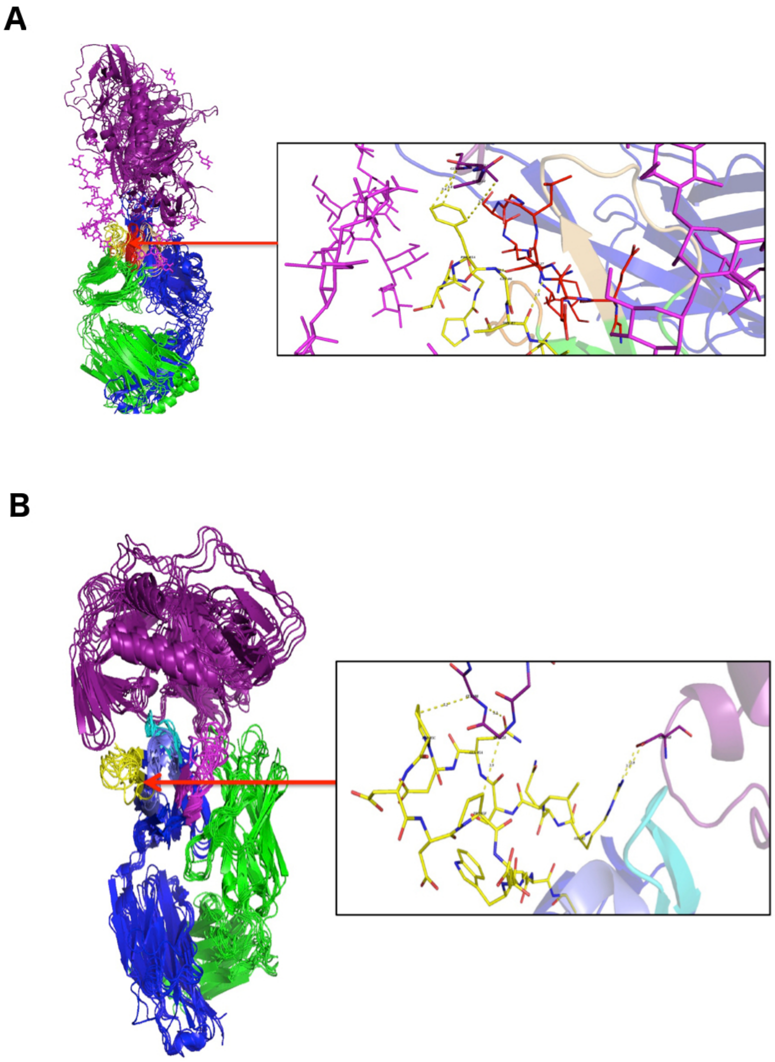

Figure 8. Alignment of a subset of gp120 binding HIV-1 bnAbs representing all unique DE loop sequences. A. Aligned structures of L4-inserted bnAbs binding to HIV-1 gp120 (one representative per unique L4 sequence). The inset shows hydrophobic contacts between the antibody-antigen interface, as well as hydrogen bonds at the antibody-antigen interface and between L1 and L4, stabilizing a unique L1 conformation. **B.** Aligned structures of H4 inserted bnAbs binding to HIV-1 gp120 (one representative per unique H4 sequence). The inset shows hydrophobic contacts between the antibody-antigen interface, as well as a salt bridge between the first Arg residue in H4 and the antigen

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The role of the non-canonical length L4 loops is particularly interesting. These antibodies are related to the Hu_IGLV3_21*01 germline and feature length 9 L4 loops. These loops not only directly bind antigen with hydrophobic interactions at the apex of the loop (Figure 8A), but also stabilize the conformation of L1 through a couple of backbone-backbone and backbone-side-chain hydrogen bonds to a serine in L1, which is sandwiched between L4 and L3. This ‘L1 sandwich’ motif appears to rigidify the binding conformation of the antibody light chain that buries a tremendous amount of binding surface area while binding to gp120 even in the presence of highly glycosylated elements (Figure 8A). The L1-14 conformations associated with these antibodies are exclusive, and no other antibody structures contain these unique conformations of L1. Evolution in L4 may stabilize L1, and enable the formation of new interactions between the antibody and antigen.

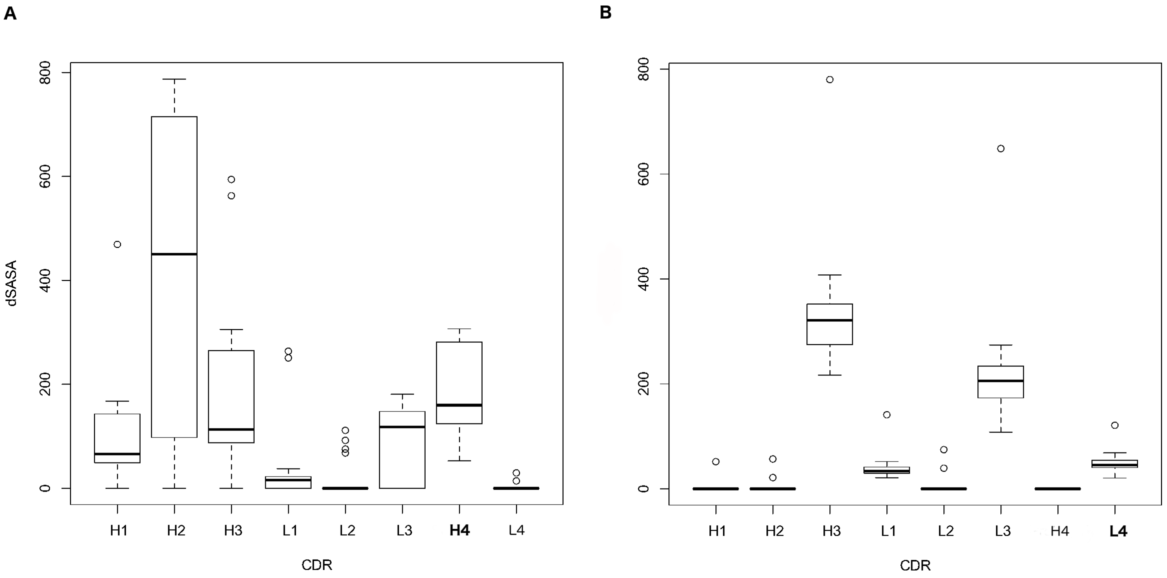

Figure 9. Buried surface area for each CDR at the antibody-antigen interface of HIV-1 bnAbs that bind to gp120 in the PDB. A. Buried surface area plot for 18 PDB structures (non-redundant by chain) with insertions in H4. B. Buried surface area plot for 31 PDB structures (non-redundant by chain) with insertions in L4.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Features of DE loop sequences from HIV-1 infected patients

In searching for HIV-1 bnAb DE loop sequences in high-throughput antibody sequences that are related to the sequences we identified in Table 6, we scanned through ~24 million high-throughput sequences related to 13 studies of HIV-1 bnAbs. None of sequences had DE loop insertion sequences related to those in the PDB listed in Table 6. However, in the high-throughput set, we observed 599 unique (637 total) heavy chain sequences, 640 unique (1354 total) λ sequences, and 2,822 (6,352 total) κ sequences with amino acid insertions in, or around (-10 C-terminal, +10 N-terminal) the DE loop. These sequences are found across 13 HIV-1 high-throughput sequencing datasets related to the affinity maturation of VRC01, CH103, and PTG134-137 lineage antibodies, as well as co-evolution of HIV-1 bnAbs with their founder HIV-1 virus (10,39–44). As previously mentioned, we found no insertions in the DE loop from the naïve human set, possibly indicating a unique response of antibody diversification under HIV-1 infection.

Antibody L4 sequences with κ germlines (Figure 10A) have the most insertions compared to heavy and λ DE loop sequences. Some of the germlines simply have a two amino acid insertion added before the DE loop sequence (e.g. human IGKV4-4*01 inserts GS before GSGTDF). Some germlines have non-frameshift causing insertions, alongside somatic mutation (e.g. human IGKV3D-11*01 has both a GS insertion before the DE loop, as well as eight residue DE loops sequence ASAAGTEF and ASASGTDF). κ L4 sequences part of the human IGKV3D-11*01 germline also have two varieties of frameshift mutations, where a single amino acid is inserted in the FR3 or the DE loop, which results in a highly mutated L4 sequence. In one case, the insertion happens at the beginning of the DE loop, resulting in a hypermutated DE loop, and the surrounding sequence remains untouched (Figure 10A). We verified a sampling of the frameshift causing insertions in IMGT/V-QUEST.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

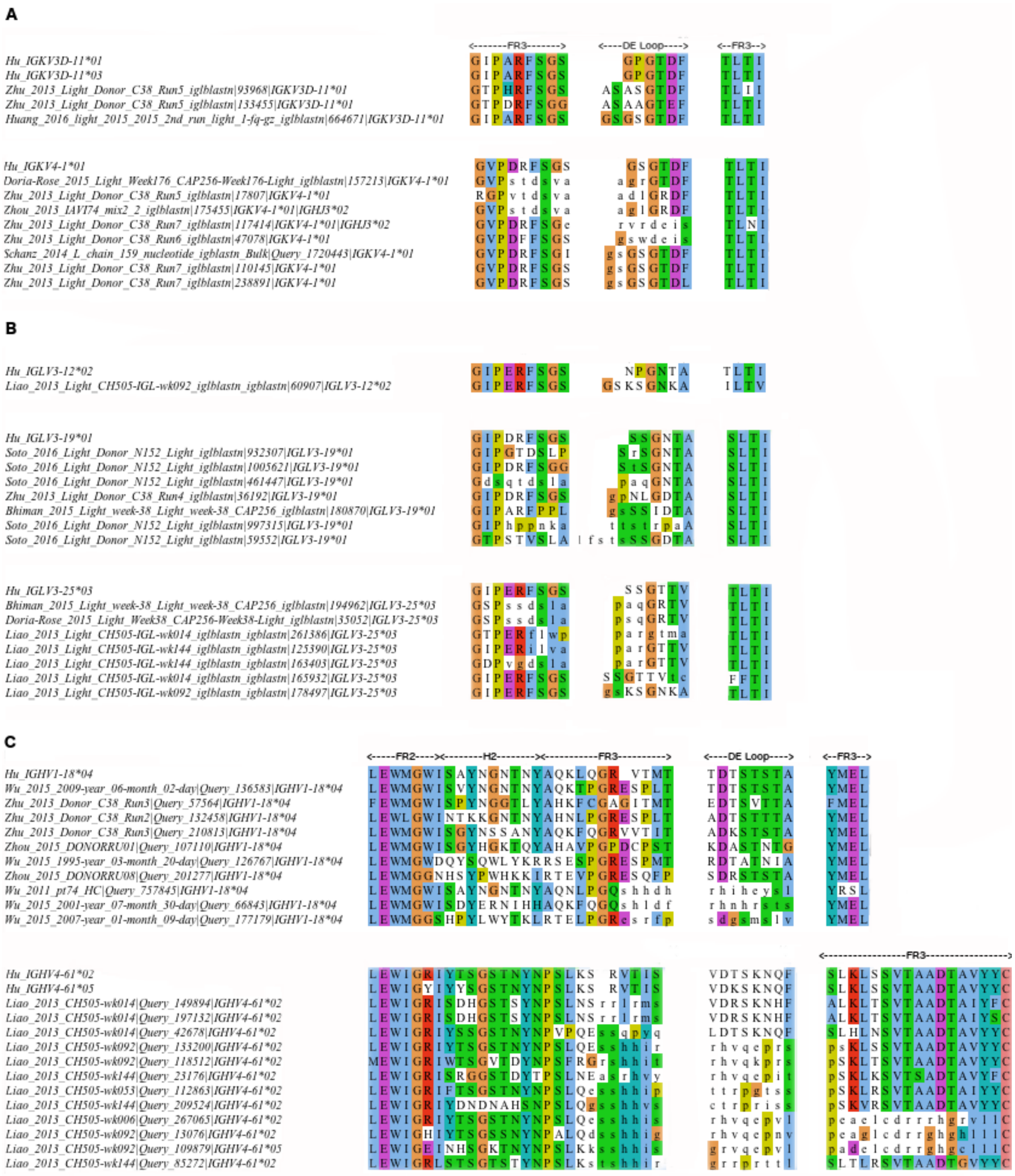
A
<-------FR3------->
<----DE Loop---->
<--FR3-->
B
C
<-----FR2----><--------H2--------><-----------FR3---------->
<----DE Loop---->
<--FR3-->
<-----------------FR3----------------->

Figure 10. DE loop and DE loop adjacent insertions from a large antibody sequencing dataset from HIV-infected individuals. A. Insertions in κ gene antibodies. **B.** Insertions in λ gene antibodies. **C.** Insertions in heavy gene antibodies.

Antibody L4 sequences with λ germlines (Figure 10B) have fewer insertions than κ L4 sequences, but still have features of non-framework shifting insertions, framework shifting insertion, and somatic mutation

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

alongside insertion. Similar to κ germline human IGKV4-4*01, human IGLV3-12*02 inserts a GS before the DE loop, and mutates the germline DE loop sequence from NPGNTA to KSGNKA (whole L4 sequence GSKSGNKA). Human IGLV3-19*01 and IGLV3-25*03 have both frameshift causing mutations, as well a single nucleotide insertions that do no dramatically change the DE loop sequence (e.g. a single amino acid insertion changing the germline sequence from SSGNTA to STSGNTA).

The majority of heavy chain insertions (Figure 10C) are single amino acid insertions arising from nucleotide insertion causing frameshifts in and around the DE loop (identified from IMGT/V-QUEST analysis). These frameshift mutations result in the introduction of charged residues such as Arg, Glu, and His in the DE loop, as well as rare amino acids such as Cys and Pro. Both human IGHV1-18*04 and human IGHV4-61*02 undergo hypersomatic mutation as well as frameshift insertions as shown in the examples (Figure 10C).

Discussion

In this paper, we have analyzed sequence and structural features of the antibody DE loop. We have clustered DE loop conformations of the heavy and light chains of all lengths, identified atomic interactions that are highly associated with various CDR1/CDR2 and DE loop pairs, and have shown features of affinity matured antibodies that utilize DE loops with somatic insertions to directly bind antigen. Our study suggests a new treatment of the antibody binding region for antibody structure analysis and antibody design, by regarding the DE loop as a fourth CDR, denoting it L4 and H4 in the light and heavy chains, along with their respective clusters L4-6-1, L4-6-2, L4-6-3, L4-6-4, L4-8-1, and H4-8-1. With this new structural classification and nomenclature to describe the DE loop of all antibody structures, we encourage all research related to antibody structure to explicitly account for

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the DE loop structure and sequence. We make an argument for the DE loop as a fourth CDR in considering its structure compared to other CDRs, wherein it has a solvent exposed loop flanked by beta strands, its sequence entropy which in the case of the heavy chain and λ germlines sequences exceeds that of the framework residues, and in its ability to insert and mutate in response to antigen, as shown here with the motivating example of HIV-1 bnAbs.

In considering DE loop structure, categorizing the associations and CDR1/CDR2 and DE loops, as well as listing all stabilizing interactions does not describe the entire relationship between the DE loop and various CDR conformations. For example, even if we have a high degree of association between a DE loop and CDR1 pair (e.g. L4-6-1 and L1-17-1), we cannot say that DE loop conformation is a determinant of the CDR1/CDR2 conformation. The association between this CDR1 conformation can be explained by the antibody germline, where specific germlines exist only with L1-17-1 and L4-6-1. Also, in this case there is only one cluster in the L1-17 family of conformations, thus L4-6-1 can only co-occur with a single L1 conformation for this particular set of L1-17-1 germlines. Therefore, when accounting for an impact that the DE loop may have on a neighboring CDR, it is important to note the L4 conformations available for that particular choice of CDR conformation, and also the sequence positions within a singular DE loop cluster that differentially impact CDR1 conformation (e.g. L4-6-2 hydrogen bonds to L1-13-1 with DE residue R1 at a rate of 75%, but bonds to L1-13-1 with DE residue K1 at a rate of >1%).

Furthermore, the motion of CDR1/CDR2 relative to the DE loop may also be an important determinant in finding a stable conformation. While the DE loop conformation for L4-6 DE loops can re-arrange the backbone atoms to readily make hydrogen bonds with various L1 conformations, L1 can also vary in length and conformation to reach towards L4, thus making hydrogen

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bonds to the DE loop more or less accessible. Second, while some hydrogen bonds occur at a very high occupancy (90% or above) in specific DE loops pairings with CDR1/CDR2, others occur at a lower occupancy (70% in the case of L4-6-1 DE residue Asn). Hydrogen bonds of this nature may have an effect on stability of the L1 conformation, but without additional experimental data to test this hypothesis, we cannot determine this.

In considering the DE loop as a fourth CDR, we suggest applications for antibody design and antibody modeling. For example, when designing antibodies using the ‘CDR grafting’ method (56), we suggest that whenever CDR1 is grafted on the light chain, or CDR1 or CDR2 on the heavy chain, L4 or H4 should be ‘co-grafted’ onto the same template structure. This method will preserve contacts between L4/L1, H4/H1, or H4/H2 that are necessary for preserving the structures of CDR1 or CDR2. When considering antibody modeling, a common strategy is to use CDR and framework templates based upon sequence similarity to known structures. We suggest extra attention to the relationships of L4 sequences with their structural clusters. For example, κ antibodies with a somatic mutation at the first position of the L4 from glycine to any other residue should be modeled with representative structures from cluster L4-6-2 instead of the more common L4-6-1 conformation. A similar approach can be considered when selecting template structures for molecular replacement. Taking this information into account is more likely to recapitulate contacts observed in experimental structures. The appropriate cluster, and thus structure, for CDR1 and CDR2 often depends on the sequence and conformation of CDR4, and they should be modeled together in antibody structure prediction methods.

With high-throughput sequencing data in response to HIV-1, we have shown that the DE loop undergoes hyper somatic mutation, alongside nucleotide insertion causing frameshift mutations in several human germline examples. Tracking, and subsequently harnessing useful features from the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DE loop sequences that contribute to antigen binding, and ultimately neutralization of viral infections may prove an important step in identifying functional antibodies from the human repertoire.

Materials and Methods

Antibody structure and sequence data

We compiled sequence and structure data for all antibodies from the Protein Data Bank (PDB). To collect the list of antibodies in the PDB, we used a lab maintained software, PyIgClassify (57). PyIgClassify compiles all antibody structures from the PDB by applying a set of hidden Markov models (HMMs) for each antibody gene to all sequences in the PDB using HMMER3.0. PyIgClassify also renumbers antibodies according to a modified Honegger-Plückthun CDR scheme and numbering system described in North et al. (48,58) In order to identify CDRs in PyIgClassify, the software uses sequence alignment to the match states of the HMMs.

In order to identify which residues are structurally variable, we plotted ϕ and ψ for all residues in and around the solvent exposed DE loop (3 before the loop, and 3 after the loop, Figure 2). We updated PyIgClassify to recognize L4 and H4 in each antibody sequence, and subsequently assign them to residue numbers within the range of 82-89, adding insert codes appropriately for loops long to exhaust this pre-allocated range of numbers.

We determine germline by comparing each PDB sequence to a curated set of IMGT germline protein sequences with BLAST taking into account the author-provided species designation. However, these are often incorrect. We use a germline from a different species from the author-provided one if the sequence identity of the antibody is at least 8 percentage points higher than the author-provided species. This script also handles cases of ambiguous assignment such as humanized antibodies originating from non-human germlines. The dataset is up-to-date as of August 2019 and includes data for

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

approximately 3,700 antibody structures (available at http://dunbrack2.fccc.edu/PyIgClassify)

Analyzing antibody-antigen complex set

For non-canonical length structures, we calculated the antibody-antigen buried surface area with the Rosetta macromolecular modeling suite (59). We calculated buried surface area as the change in antigen surface area of the CDR from the bound structure to unbound structure:

$$BSA = SA_{bound} - SA_{unbound} \tag{1}$$

Where SA represents the surface area calculated in Rosetta using the Shrake-Rupley algorithm and a standard probe radius of 1.4 Å.

Clustering loop structures

In order to group various conformations of L4 and H4 into structural families, we implemented a density based clustering method for dihedral angles based on the DBSCAN algorithm (46). This unsupervised learning method represents an improvement over previous implementations of internal dihedral metric clustering due to its identification of outliers in the dataset, while simultaneously finding robust clusters by identifying dense regions in the metric space which are separated by low density. DBSCAN's noise detection inherently removes outlier structures due to poor crystal structure determination or other crystallization artifacts. We used the implementation of DBSCAN in the sci-kit learn library in python.

To compare two loops *i* and *j* with identical lengths, we first calculate the dihedral similarity between two angles θ_1 and θ_2 for each pair of corresponding residues, where θ represents any chosen combination of backbone dihedrals angles ϕ, ψ, or ω:

$$d = 2(1 - \cos(\theta_1 - \theta_2)) \qquad \text{for } \theta_i = \{\phi, \psi, \omega\} \tag{2}$$

For our purposes we chose to include ϕ, ψ, and ω, which provides the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

maximum capability to resolve structures with both cis- and trans- peptide bonds. Next, we take as the final clustering distance the maximum value out of the set of calculations of *d for* {ϕ, ψ, ω}, which we call the *L*∞ norm:

$$L_{\infty} = \max\left(d_{\phi 1}, d_{\psi 1}, d_{\omega 1}, \ldots, d_{\phi N}, d_{\psi N}, d_{\omega N}\right) \tag{3}$$

We chose the *L*∞ norm due to its sensitivity in separating loops which are different even at one single dihedral, giving our final clustering single dihedral resolution.

The resulting set of pairwise *L*∞ distances are then clustered in a *N*x*N* pairwise matrix using DBSCAN. This algorithm requires two parameters: ε and *MinPts*. The first parameter, ε, describes a distance from a given data point to search for neighboring data points. The second parameter, *MinPts*, specifies the requirement for the minimum number of neighboring data points within ε of a data point to label the data point under consideration a 'core point'. Data points which are within ε of a core point, but do have *MinPts* data points within ε are called 'border points'; points which do not meet either criterion are labeled as 'noise points.' The final clusters are the connected graphs of all of the core points, together with their border points.

Each selection of a combination of ε and *MinPts* produces a different set of clusters. Two main obstacles exist in identifying all of the interesting clusters from DBSCAN. First, at specific choices of ε and *MinPts*, DBSCAN may merge clusters that ought to be separated, these are identified in the context of Ramachandran data by plotting φ and ψ for each residue in each cluster. Merged clusters are easily identified by their non-convergent Ramachandran φ/ψ conformation at specific residues within a cluster. Second, clusters of varying density arise at different selections of ε and *MinPts*, which may coincide with the choice of ε and *MinPts* that produced a merged cluster. This means that no singular selection of ε and *MinPts* will generate the entire set of interesting clusters. To overcome these two issues, we developed a method to select a set of final clusters after running

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DBSCAN on a grid of ε and *MinPts*, by combining the results of each run of DBSCAN. First, we establish a parameter grid of ε and *MinPts* by selecting a range of both parameters, and run DBSCAN at each parameter selection. We then filter out any merged clusters by removing any clusters in which any two members of the cluster are more than 150° apart. Next, the remaining clusters that pass the merge filtering criterion are treated as nodes on a graph, where the nodes have edges connected to them based on the calculation of Simpson's similarity score:

$$\mathrm{S} = \frac{|A \cap B|}{\min\,(|A|,|B|)} \tag{4}$$

Finally, for each connected subgraph with n nodes, we take the final cluster of that subgraph as the union of all nodes n within the connected subgraph. This produces a final clustering set with clusters of varying density, without including merged clusters.

Following the determination of the final cluster set, we determined cluster representatives using angular statistical analysis. For a given cluster C consisting of N data points, for each structure *i* we calculate the average distance d_i to all other points *j* in the same cluster C:

$$d_i = \frac{1}{N}\sum_{j=1,N} d_{ij} \tag{5}$$

We choose the cluster representative as the structure which has the lowest d_i of all of the structures.

Identifying important hydrogen bonds between CDR4 and CDR1/CDR2

We calculated all hydrogen bonds between CDR4, CDR1, and CDR2 using Rosetta's distance and orientation-dependent hydrogen bond energy accessed through the *report_hbonds_for_plugin.<release>* available in the public release of Rosetta3. We used the resulting contact information to find

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

important contacts that are either frequent or unique over several CDR-lengths and germlines. We analyzed the hydrogen bonds between all CDR1-CDR4 and CDR2-CDR4 pairs for which both CDR1 and CDR4 have defined cluster membership. We then calculated the hydrogen bond occupancy for a particular hydrogen bond as the following:

$$occupancy = \frac{\#\ hbonds\ to\ cdr \mid deresidue, deatom, cdr\ cluster}{\#\ structures\ with\ deresidue, deatom, cdr\ cluster} \quad (6)$$

High-throughput sequence analysis of naïve human antibodies

We accessed high-throughput sequencing data through the antibodymap.org server (www.antibodymap.org). To gain an understanding of how variable L4 and H4 are compared to the other CDRs, we analyzed 12 human germlines (IGHV1-18*04, IGHV3-23*01, IGHV4-34*01, IGHV4-39*07, IGKV1-39*01, IGKV3-11*01, IGKV3-20*01, IGKV4-1*01, IGLV1-40*01, IGLV1-44*01, IGLV2-14*01, IGLV3-1*01) collected from naïve donor deep sequencing samples with thousands of sequences for each germline (download shell script included in supplementary data). Separately, we compared sequence variability between all human germlines for each heavy, λ, and κ gene compiled from IMGT for all germline sequences of the same length. We calculated sequence variability according to the Shannon entropy, denoted H), which represents the most robust method to calculate antibody CDR variability according to Stewart et al. (60).

$$H = -\sum_{i=1,N} p_i \log_2 p_i \quad (7)$$

We calculated *H* only for residues up until the conserved cysteine before CDR3 on both the light and heavy chains.

High-throughput sequence analysis of HIV-1 bnABs

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

In order to search for insertions in L4 or H4 amongst HIV-1 infected patients, we collected all studies referring to HIV-1 from the antibodymap API (download shell script included in supplementary data). We identified CDRs for all of the FASTA files using the HMMER3.0 hmmsearch command, providing the profile HMMs implemented in PyIgClassify for IGHV, IGKV, IGLV, and IGLV6 genes (provided in supplementary data). We searched for sequences that had insertions compared to the profile, and examined these for features related to the long L4 or H4 structures we found in the PDB (sequences are provided in FASTA format in the supplementary data). From this set we used Clustal-omega to align all of the sequences to the germline sequence which matched the IMGT germline assignment (provided in supplementary data). We observed frameshift mutations using the IMGT/V-QUEST tool, which notates nucleotide insertions that result in frameshifts. PDF results of all of our IMGT/V-QUEST queries are supplied in the supplemental data.

Acknowledgements

This work was funded by NIH grant R35 GM1225167 (to R.L.D.) as well as the NIH T32 Structural Biology and Molecular Biophysics training grant awarded by the department of Biochemistry and Molecular Biophysics (GM008275).

References

1. Kabat EA, Wu TT. Attempts to Locate Complementarity-Determining Residues in the Variable Positions of Light and Heavy Chains *. *Ann N Y Acad Sci* (1971) **190**:382–393. doi:10.1111/j.1749-6632.1971.tb13550.x

2. Bork P, Holm L, Sander C. The Immunoglobulin Fold. *J Mol Biol* (1994) **242**:309–320. doi:10.1006/jmbi.1994.1582

3. Lesk AM, Chothia C. Evolution of proteins formed by β-sheets: II. The core of the immunoglobulin domains. *J Mol Biol* (1982) **160**:325–342. doi:10.1016/0022-2836(82)90179-6

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

4. Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Barbié V, Chaume D. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* (1999) **27**:209–212. doi:10.1093/nar/27.1.209

5. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) **196**:901–917. doi:10.1016/0022-2836(87)90412-8

6. Tramontano A, Chothia C, Lesk AM. Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. *J Mol Biol* (1990) **215**:175–182. doi:10.1016/S0022-2836(05)80102-0

7. Foote J, Winter G. Antibody framework residues affecting the conformation of the hypervariable loops. *J Mol Biol* (1992) **224**:487–499. doi:10.1016/0022-2836(92)91010-M

8. Lehmann A, Wixted JH, Shapovalov MV, Roder H, Dunbrack Jr RL, Robinson MK. Stability engineering of anti-EGFR scFv antibodies by rational design of a lambda-to-kappa swap of the VL framework using a structure-guided approach. in *mAbs* (Taylor & Francis), 1–14. doi:DOI: 10.1080/19420862.2015.1088618

9. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, Chen X, Longo NS, Louder M, McKee K, et al. Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. *Science* (2011) **333**:1593–1602. doi:10.1126/science.1207532

10. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, Georgiev IS, Altae-Tran HR, Chuang G-Y, Joyce MG, et al. Multidonor Analysis Reveals Structural Elements, Genetic Determinants, and Maturation Pathway for HIV-1 Neutralization by VRC01-Class Antibodies. *Immunity* (2013) **39**:245–258. doi:10.1016/j.immuni.2013.04.012

11. Huang J, Kang BH, Ishida E, Zhou T, Griesman T, Sheng Z, Wu F, Doria-Rose NA, Zhang B, McKee K, et al. Identification of a CD4-Binding-Site Antibody to HIV that Evolved Near-Pan Neutralization Breadth. *Immunity* (2016) **45**:1108–1121. doi:10.1016/j.immuni.2016.10.027

12. Rutten L, Lai Y-T, Blokland S, Truan D, Bisschop IJM, Strokappe NM, Koornneef A, Manen D van, Chuang G-Y, Farney SK, et al. A Universal Approach to Optimize the Folding and Stability of Prefusion-Closed HIV-1 Envelope Trimers. *Cell Rep* (2018) **23**:584–595. doi:10.1016/j.celrep.2018.03.061

13. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TYK, Pietzsch J, Fenyo D, Abadir A, Velinzon K, et al. Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. *Science* (2011) **333**:1633–1637.

14. Mouquet H, Scharf L, Euler Z, Liu Y, Eden C, Scheid JF, Halper-Stromberg A, Gnanapragasam PNP, Spencer DIR, Seaman MS, et al. Complex-type N-glycan

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

recognition by potent broadly neutralizing HIV antibodies. *Proc Natl Acad Sci* (2012) **109**:E3268–E3277. doi:10.1073/pnas.1217207109

15. Klein F, Diskin R, Scheid JF, Gaebler C, Mouquet H, Georgiev IS, Pancera M, Zhou T, Incesu R-B, Fu BZ, et al. Somatic Mutations of the Immunoglobulin Framework Are Generally Required for Broad and Potent HIV-1 Neutralization. *Cell* (2013) **153**:126–138. doi:10.1016/j.cell.2013.03.018

16. Julien JP, Sok D, Khayat R, Lee JH, Doores KJ, Walker LM, Ramos A, Diwanji DC, Pejchal R, Cupo A, et al. Broadly neutralizing antibody PGT121 allosterically modulates CD4 binding via recognition of the HIV-1 gp120 V3 base and multiple surrounding glycans. *Plos Pathog* (2013) **9**:e1003342–e1003342. doi:10.2210/pdb4jy4/pdb

17. Julien J-P, Cupo A, Sok D, Stanfield RL, Lyumkis D, Deller MC, Klasse P-J, Burton DR, Sanders RW, Moore JP, et al. Crystal Structure of a Soluble Cleaved HIV-1 Envelope Trimer. *Science* (2013) **342**:1477–1483. doi:10.1126/science.1245625

18. Scharf L, Scheid JF, Lee JH, West AP, Chen C, Gao H, Gnanapragasam PNP, Mares R, Seaman MS, Ward AB, et al. Antibody 8ANC195 Reveals a Site of Broad Vulnerability on the HIV-1 Envelope Spike. *Cell Rep* (2014) **7**:785–795. doi:10.1016/j.celrep.2014.04.001

19. Garces F, Sok D, Kong L, McBride R, Kim HJ, Saye-Francisco KF, Julien J-P, Hua Y, Cupo A, Moore JP, et al. Structural Evolution of Glycan Recognition by a Family of Potent HIV Antibodies. *Cell* (2014) **159**:69–79. doi:10.1016/j.cell.2014.09.009

20. Pancera M, Zhou T, Druz A, Georgiev IS, Soto C, Gorman J, Huang J, Acharya P, Chuang G-Y, Ofek G, et al. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* (2014) **514**:455–461. doi:10.1038/nature13808

21. Wu X, Zhang Z, Schramm CA, Joyce MG, Do Kwon Y, Zhou T, Sheng Z, Zhang B, O'Dell S, McKee K, et al. Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. *Cell* (2015) **161**:470–485. doi:10.1016/j.cell.2015.03.004

22. Scharf L, Wang H, Gao H, Chen S, McDowall AW, Bjorkman PJ. Broadly Neutralizing Antibody 8ANC195 Recognizes Closed and Open States of HIV-1 Env. *Cell* (2015) **162**:1379–1390. doi:10.1016/j.cell.2015.08.035

23. Kong L, Torrents de la Peña A, Deller MC, Garces F, Sliepen K, Hua Y, Stanfield RL, Sanders RW, Wilson IA. Complete epitopes for vaccine design derived from a crystal structure of the broadly neutralizing antibodies PGT128 and 8ANC195 in complex with an HIV-1 Env trimer. *Acta Crystallogr D Biol Crystallogr* (2015) **71**:2099–2108. doi:10.1107/S1399004715013917

24. Kong R, Xu K, Zhou T, Acharya P, Lemmin T, Liu K, Ozorowski G, Soto C, Taft JD, Bailer RT, et al. Fusion peptide of HIV-1 as a site of vulnerability to neutralizing antibody. *Science* (2016) **352**:828–833. doi:10.1126/science.aae0474

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

25. Davenport TM, Gorman J, Joyce MG, Zhou T, Soto C, Guttman M, Moquin S, Yang Y, Zhang B, Doria-Rose NA, et al. Somatic Hypermutation-Induced Changes in the Structure and Dynamics of HIV-1 Broadly Neutralizing Antibodies. *Structure* (2016) **24**:1346–1357. doi:10.1016/j.str.2016.06.012

26. Steichen JM, Kulp DW, Tokatlian T, Escolano A, Dosenovic P, Stanfield RL, McCoy LE, Ozorowski G, Hu X, Kalyuzhniy O, et al. HIV Vaccine Design to Target Germline Precursors of Glycan-Dependent Broadly Neutralizing Antibodies. *Immunity* (2016) **45**:483–496. doi:10.1016/j.immuni.2016.08.016

27. Wang H, Cohen AA, Galimidi RP, Gristick HB, Jensen GJ, Bjorkman PJ. Cryo-EM structure of a CD4-bound open HIV-1 envelope trimer reveals structural rearrangements of the gp120 V1V2 loop. *Proc Natl Acad Sci* (2016) **113**:E7151–E7158. doi:10.1073/pnas.1615939113

28. Pancera M, Lai Y-T, Bylund T, Druz A, Narpala S, O'Dell S, Schön A, Bailer RT, Chuang G-Y, Geng H, et al. Crystal structures of trimeric HIV envelope with entry inhibitors BMS-378806 and BMS-626529. *Nat Chem Biol* (2017) **13**:1115–1122. doi:10.1038/nchembio.2460

29. Guenaga J, Garces F, Val N de, Stanfield RL, Dubrovskaya V, Higgins B, Carrette B, Ward AB, Wilson IA, Wyatt RT. Glycine Substitution at Helix-to-Coil Transitions Facilitates the Structural Determination of a Stabilized Subtype C HIV Envelope Glycoprotein. *Immunity* (2017) **46**:792–803.e3. doi:10.1016/j.immuni.2017.04.014

30. Chuang G-Y, Geng H, Pancera M, Xu K, Cheng C, Acharya P, Chambers M, Druz A, Tsybovsky Y, Wanninger TG, et al. Structure-Based Design of a Soluble Prefusion-Closed HIV-1 Env Trimer with Reduced CD4 Affinity and Improved Immunogenicity. *J Virol* (2017) **91**:e02268–16. doi:10.1128/JVI.02268-16

31. Zhou T, Doria-Rose NA, Cheng C, Stewart-Jones GBE, Chuang G-Y, Chambers M, Druz A, Geng H, McKee K, Kwon YD, et al. Quantification of the Impact of the HIV-1-Glycan Shield on Antibody Elicitation. *Cell Rep* (2017) **19**:719–732. doi:10.1016/j.celrep.2017.04.013

32. Wang H, Gristick HB, Scharf L, West AP Jr, Galimidi RP, Seaman MS, Freund NT, Nussenzweig MC, Bjorkman PJ. Asymmetric recognition of HIV-1 Envelope trimer by V1V2 loop-targeting antibodies. *eLife* (2017) **6**:e27389. doi:10.7554/eLife.27389

33. Medina-Ramírez M, Garces F, Escolano A, Skog P, Taeye SW de, Moral-Sanchez ID, McGuire AT, Yasmeen A, Behrens A-J, Ozorowski G, et al. Design and crystal structure of a native-like HIV-1 envelope trimer that engages multiple broadly neutralizing antibody precursors in vivo. *J Exp Med* (2017) **214**:2573–2590. doi:10.1084/jem.20161160

34. Sarkar A, Bale S, Behrens A-J, Kumar S, Sharma SK, Val N de, Pallesen J, Irimia A, Diwanji DC, Stanfield RL, et al. Structure of a cleavage-independent HIV Env recapitulates the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

glycoprotein architecture of the native cleaved trimer. *Nat Commun* (2018) **9**:1–14. doi:10.1038/s41467-018-04272-y

35. Moyo T, Ereño-Orbea J, Jacob RA, Pavillet CE, Kariuki SM, Tangie EN, Julien J-P, Dorfman JR. Molecular Basis of Unusually High Neutralization Resistance in Tier 3 HIV-1 Strain 253-11. *J Virol* (2018) **92**:e02261–17. doi:10.1128/JVI.02261-17

36. Xu K, Acharya P, Kong R, Cheng C, Chuang G-Y, Liu K, Louder MK, O'Dell S, Rawi R, Sastry M, et al. Epitope-based vaccine design yields fusion peptide-directed antibodies that neutralize diverse strains of HIV-1. *Nat Med* (2018) **24**:857–867. doi:10.1038/s41591-018-0042-6

37. Barnes CO, Gristick HB, Freund NT, Escolano A, Lyubimov AY, Hartweger H, West AP, Cohen AE, Nussenzweig MC, Bjorkman PJ. Structural characterization of a highly-potent V3-glycan broadly neutralizing antibody bound to natively-glycosylated HIV-1 envelope. *Nat Commun* (2018) **9**:1–12. doi:10.1038/s41467-018-03632-y

38. Zhang P, Gorman J, Geng H, Liu Q, Lin Y, Tsybovsky Y, Go EP, Dey B, Andine T, Kwon A, et al. Interdomain Stabilization Impairs CD4 Binding and Improves Immunogenicity of the HIV-1 Envelope Trimer. *Cell Host Microbe* (2018) **23**:832–844.e6. doi:10.1016/j.chom.2018.05.002

39. Bhiman JN, Anthony C, Doria-Rose NA, Karimanzira O, Schramm CA, Khoza T, Kitchin D, Botha G, Gorman J, Garrett NJ, et al. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. *Nat Med* (2015) **21**:1332–1336. doi:10.1038/nm.3963

40. Zhou T, Lynch RM, Chen L, Acharya P, Wu X, Doria-Rose NA, Joyce MG, Lingwood D, Soto C, Bailer RT, et al. Structural Repertoire of HIV-1-Neutralizing Antibodies Targeting the CD4 Supersite in 14 Donors. *Cell* (2015) **161**:1280–1292. doi:10.1016/j.cell.2015.05.007

41. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, Fire AZ, Roskin KM, Schramm CA, Zhang Z, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* (2013) **496**:469–476. doi:10.1038/nature12053

42. Schanz M, Liechti T, Zagordi O, Miho E, Reddy ST, Günthard HF, Trkola A, Huber M. High-Throughput Sequencing of Human Immunoglobulin Variable Regions with Subtype Identification. *PLOS ONE* (2014) **9**:e111726. doi:10.1371/journal.pone.0111726

43. Soto C, Ofek G, Joyce MG, Zhang B, McKee K, Longo NS, Yang Y, Huang J, Parks R, Eudailey J, et al. Developmental Pathway of the MPER-Directed HIV-1-Neutralizing Antibody 10E8. *PLOS ONE* (2016) **11**:e0157409. doi:10.1371/journal.pone.0157409

44. Zhu J, Wu X, Zhang B, McKee K, O'Dell S, Soto C, Zhou T, Casazza JP, Mullikin JC, Kwong PD, et al. De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

generation sequencing of B-cell transcripts. *Proc Natl Acad Sci U S A* (2013) **110**:E4088–E4097. doi:10.1073/pnas.1306262110

45. Kepler TB, Liao H-X, Alam SM, Bhaskarabhatla R, Zhang R, Yandava C, Stewart S, Anasti K, Kelsoe G, Parks R, et al. Immunoglobulin Gene Insertions and Deletions in the Affinity Maturation of HIV-1 Broadly Reactive Neutralizing Antibodies. *Cell Host Microbe* (2014) **16**:304–313. doi:10.1016/j.chom.2014.08.006

46. Ester M, Kriegel H-P, Sander J, Xu X. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* KDD'96. (Portland, Oregon: AAAI Press), 226–231. Available at: http://dl.acm.org/citation.cfm?id=3001460.3001507 [Accessed February 2, 2017]

47. Wall JS, Gupta V, Wilkerson M, Schell M, Loris R, Adams P, Solomon A, Stevens F, Dealwis C. Structural basis of light chain amyloidogenicity: comparison of the thermodynamic properties, fibrillogenic potential and tertiary structural features of four Vλ6 proteins. *J Mol Recognit* (2004) **17**:323–331. doi:10.1002/jmr.681

48. North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. *J Mol Biol* (2011) **406**:228–256.

49. Galson JD, Clutterbuck EA, Trück J, Ramasamy MN, Münz M, Fowler A, Cerundolo V, Pollard AJ, Lunter G, Kelly DF. BCR repertoire sequencing: different patterns of B cell activation after two Meningococcal vaccines. *Immunol Cell Biol* (2015) **93**:885–895. doi:10.1038/icb.2015.57

50. Ellebedy AH, Jackson KJL, Kissick HT, Nakaya HI, Davis CW, Roskin KM, McElroy AK, Oshansky CM, Elbein R, Thomas S, et al. Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat Immunol* (2016) **17**:1226–1234. doi:10.1038/ni.3533

51. Johnson EL, Doria-Rose NA, Gorman J, Bhiman JN, Schramm CA, Vu AQ, Law WH, Zhang B, Bekker V, Abdool Karim SS, et al. Sequencing HIV-neutralizing antibody exons and introns reveals detailed aspects of lineage maturation. *Nat Commun* (2018) **9**:1–13. doi:10.1038/s41467-018-06424-6

52. Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, M. Euskirchen G, Mamedov MR, Swan GE, Dekker CL, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun* (2016) **7**:1–12. doi:10.1038/ncomms11112

53. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, Kirgizova VI, Merzlyak EM, Staroverov DB, Bolotin DA, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) **11**:1599–1616. doi:10.1038/nprot.2016.093

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

54. Vander Heiden JA, Stathopoulos P, Zhou JQ, Chen L, Gilbert TJ, Bolen CR, Barohn RJ, Dimachkie MM, Ciafaloni E, Broering TJ, et al. Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. *J Immunol Baltim Md 1950* (2017) **198**:1460–1473. doi:10.4049/jimmunol.1601415

55. Kirik U, Persson H, Levander F, Greiff L, Ohlin M. Antibody Heavy Chain Variable Domains of Different Germline Gene Origins Diversify through Different Paths. *Front Immunol* (2017) **8**: doi:10.3389/fimmu.2017.01433

56. Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, Weitzner BD, Hu X, Adachi Y, Schief WR, Dunbrack RL. Rosetta Antibody Design (RAbD): A General Framework for Computational Antibody Design. *bioRxiv* (2018)183350. doi:10.1101/183350

57. Adolf-Bryfogle J, Xu Q, North B, Lehmann A, Dunbrack RL. PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Res* (2015) **43**:D432–D438.

58. Honegger A, Ckthun AP. Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis. *J Mol Biol* (2001)657–670.

59. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* (2011) **487**:545–574. doi:10.1016/B978-0-12-381270-4.00019-6

60. Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M, Litwin S. A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol Immunol* (1997) **34**:1067–1082. doi:10.1016/S0161-5890(97)00130-2

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table S1. Hydrogen bonds between DE loop residues and CDRs

Name	CDR Cluster	DE res num	DE res type	DE atom	CDR res num	CDR res type	CDR atom	Occ.	N in cluster	Hbond type	N cases
1) H4.6O/H1.2N	H1-13-1	6	bb	O	2	bb	N	0.952	2821	bb-bb	2687
	H1-13-2	6	bb	O	2	bb	N	0.652	23	bb-bb	15
	H1-13-3	6	bb	O	2	bb	N	0.961	77	bb-bb	74
	H1-13-4	6	bb	O	2	bb	N	0.978	134	bb-bb	131
	H1-13-5	6	bb	O	2	bb	N	0.981	54	bb-bb	53
	H1-13-7	6	bb	O	2	bb	N	0.947	38	bb-bb	36
	H1-13-10	6	bb	O	2	bb	N	1.000	12	bb-bb	12
	H1-14-1	6	bb	O	2	bb	N	0.972	71	bb-bb	69
	H1-15-1	6	bb	O	2	bb	N	0.967	123	bb-bb	119
2) L4.3O/L1.6N	L1-10-1	3	bb	O	6	bb	N	0.951	122	bb-bb	116
	L1-10-2	3	bb	O	6	bb	N	0.944	72	bb-bb	68
	L1-11-1	3	bb	O	6	bb	N	0.875	1169	bb-bb	1023
	L1-11-2	3	bb	O	6	bb	N	0.842	335	bb-bb	282
	L1-11-3	3	bb	O	6	bb	N	0.889	108	bb-bb	96
	L1-11-5	3	bb	O	6	bb	N	0.302	53	bb-bb	16
	L1-12-1	3	bb	O	6	bb	N	0.935	138	bb-bb	129
	L1-12-2	3	bb	O	6	bb	N	0.987	79	bb-bb	78
	L1-15-1	3	bb	O	6	bb	N	0.978	223	bb-bb	218
	L1-16-1	3	bb	O	6	bb	N	0.799	447	bb-bb	357
	L1-17-1	3	bb	O	6	bb	N	0.969	255	bb-bb	247
3) L4.3N/L1-11-1.7O	L1-11-1	3	bb	N	7	bb	O	0.820	1169	bb-bb	958
4) L4.4O/L1.2N	L1-10-1	4	bb	O	2	bb	N	0.967	120	bb-bb	116
	L1-10-2	4	bb	O	2	bb	N	0.931	72	bb-bb	67
	L1-11-1	4	bb	O	2	bb	N	0.983	1168	bb-bb	1148
	L1-11-2	4	bb	O	2	bb	N	0.991	335	bb-bb	332
	L1-11-3	4	bb	O	2	bb	N	0.860	107	bb-bb	92
	L1-11-4	4	bb	O	2	bb	N	0.961	76	bb-bb	73
	L1-11-5	4	bb	O	2	bb	N	0.750	52	bb-bb	39
	L1-12-1	4	bb	O	2	bb	N	0.949	138	bb-bb	131
	L1-12-2	4	bb	O	2	bb	N	1.000	79	bb-bb	79
	L1-12-3	4	bb	O	2	bb	N	1.000	30	bb-bb	30
	L1-13-1	4	bb	O	2	bb	N	0.975	157	bb-bb	153
	L1-14-2	4	bb	O	2	bb	N	0.985	131	bb-bb	129
	L1-15-1	4	bb	O	2	bb	N	0.951	223	bb-bb	212
	L1-16-1	4	bb	O	2	bb	N	0.982	447	bb-bb	439
	L1-17-1	4	bb	O	2	bb	N	0.988	255	bb-bb	252
5) L4.4O/L1-14-1.7N	L1-14-1	4	bb	O	7	bb	N	0.954	108	bb-bb	103
6) L4.3N/L1-15-1.7OD+OG	L1-15-1	3	bb	N	7	ASP	OD1	0.761	184	bb-sc	140
	L1-15-1	3	bb	N	7	SER	OG	0.857	35	bb-sc	30
7) L4.3N/L1.N(-4)OD	L1-16-1	3	bb	N	-4	ASN	OD1	0.767	30	bb-sc	23
	L1-17-1	3	bb	N	-4	ASN	OD1	0.704	240	bb-sc	169
8) H4.R1NH/H1-13.10O	H1-13-1	1	ARG	NH	10	bb	O	0.774	1417	sc-bb	1097
	H1-13-2	1	ARG	NH	10	bb	O	0.733	15	sc-bb	11
	H1-13-3	1	ARG	NH	10	bb	O	0.611	36	sc-bb	22
	H1-13-4	1	ARG	NH	10	bb	O	0.849	86	sc-bb	73
9) H4.N6ND2/H1.2O	H1-13-1	6	ASN	ND2	2	bb	O	0.214	1534	sc-bb	328
	H1-13-3	6	ASN	ND2	2	bb	O	0.262	42	sc-bb	11
	H1-13-4	6	ASN	ND2	2	bb	O	0.298	57	sc-bb	17
	H1-13-7	6	ASN	ND2	2	bb	O	0.576	33	sc-bb	19
	H1-14-1	6	ASN	ND2	2	bb	O	0.746	67	sc-bb	50
	H1-15-1	6	ASN	ND2	2	bb	O	0.108	102	sc-bb	11
10) H4.K6NZ/H1-13-3.4O	H1-13-3	6	LYS	NZ	4	bb	O	0.733	15	sc-bb	11
11) H4.N6ND2/H1.5O	H1-13-1	6	ASN	ND2	5	bb	O	0.250	1534	sc-bb	384
	H1-13-4	6	ASN	ND2	5	bb	O	0.386	57	sc-bb	22
	H1-13-7	6	ASN	ND2	5	bb	O	0.909	33	sc-bb	30
	H1-14-1	6	ASN	ND2	5	bb	O	0.806	67	sc-bb	54
	H1-15-1	6	ASN	ND2	5	bb	O	0.118	102	sc-bb	12
12) H4.ND6OD1/H1.7N	H1-13-1	6	ASN	OD1	7	bb	N	0.267	1534	sc-bb	410
	H1-13-1	6	ASP	OD1	7	bb	N	0.344	93	sc-bb	32
	H1-13-4	6	ASN	OD1	7	bb	N	0.509	57	sc-bb	29
	H1-13-7	6	ASN	OD1	7	bb	N	0.848	33	sc-bb	28
	H1-14-1	6	ASN	OD1	7	bb	N	0.821	67	sc-bb	55
	H1-15-1	6	ASN	OD1	7	bb	N	0.118	102	sc-bb	12
13) H4.RK1Nsc/H2.3O	H2-10-1	1	ARG	NE,NH	3	bb	O	0.647	184	sc-bb	119
	H2-10-1	1	LYS	NZ	3	bb	O	0.524	21	sc-bb	11
	H2-10-6	1	ARG	NE,NH	3	bb	O	0.389	36	sc-bb	14
	H2-9-1	1	ARG	NE	3	bb	O	0.481	391	sc-bb	188
	H2-9-1	1	LYS	NZ	3	bb	O	0.420	274	sc-bb	115
14) H4.RN3Nsc/H2.3O	H2-10-1	3	ARG	NE,NH	3	bb	O	0.283	106	sc-bb	30
	H2-10-1	3	ASN	ND2	3	bb	O	0.333	63	sc-bb	21
	H2-10-2	3	ASN	ND2	3	bb	O	0.776	817	sc-bb	634
	H2-9-1	3	ASN	ND2	3	bb	O	0.112	321	sc-bb	36

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.12.946350; this version posted February 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

15) L4.KR1Nsc/L1-13,14.7O	L1-13-1	1	ARG	NH	7	bb	O	0.875	16	sc-bb	14
	L1-13-1	1	LYS	NZ	7	bb	O	0.191	141	sc-bb	27
	L1-14-2	1	LYS	NZ	7	bb	O	0.359	131	sc-bb	47
16) L4.Q4NE2/L1-11-2.2O	L1-11-2	4	GLN	NE2	2	bb	O	0.810	21	sc-bb	17
17) L4.N4OD1/L1-11-3.4N	L1-11-3	4	ASN	OD1	4	bb	N	0.677	96	sc-bb	65
18) L4.R6NH/L1-12-3.7O	L1-12-3	6	ARG	NH	7	bb	O	1.000	30	sc-bb	30
19) L4.R6NH/L1-12-3.9O	L1-12-3	6	ARG	NH	9	bb	O	1.000	30	sc-bb	30
20) L4.Y6OH/L1-11-2.8N	L1-11-2	6	TYR	OH	8	bb	N	0.863	291	sc-bb	251