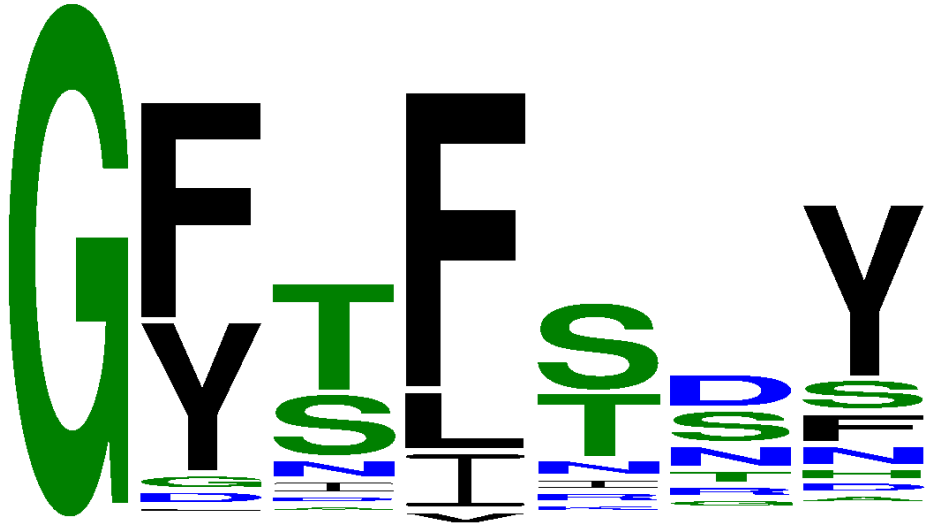
APPENDIX

A1: STRUCTURAL CLUSTERS SEQUENCE LOGOS



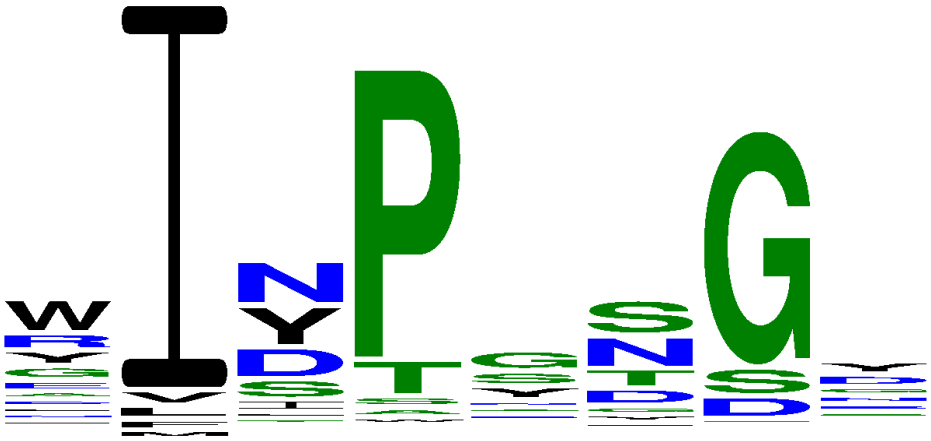
H1-7-A

H1-7-B

H1-8-A

H1-9-A

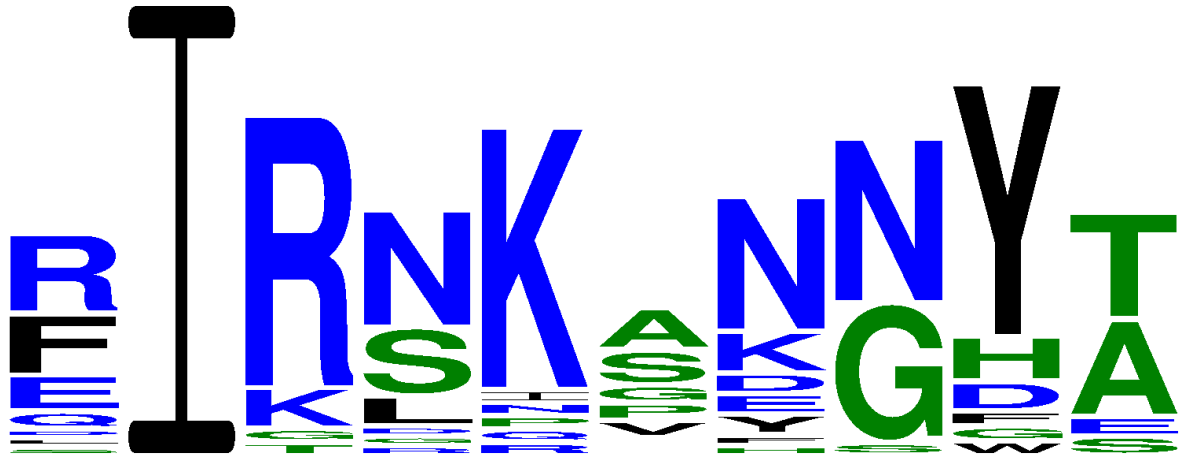
H2-7-A



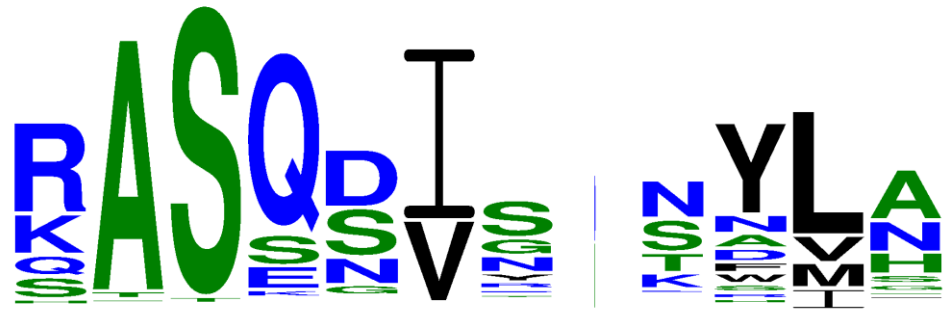
H2-8-A

H2-8-B

H2-8-D



H2-10-A

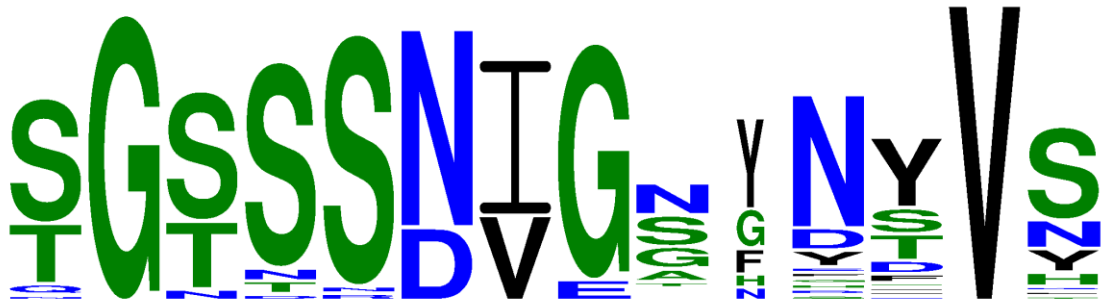


L1-10,11,12-A

L1-11-A

L1-11-B

L1-12-A

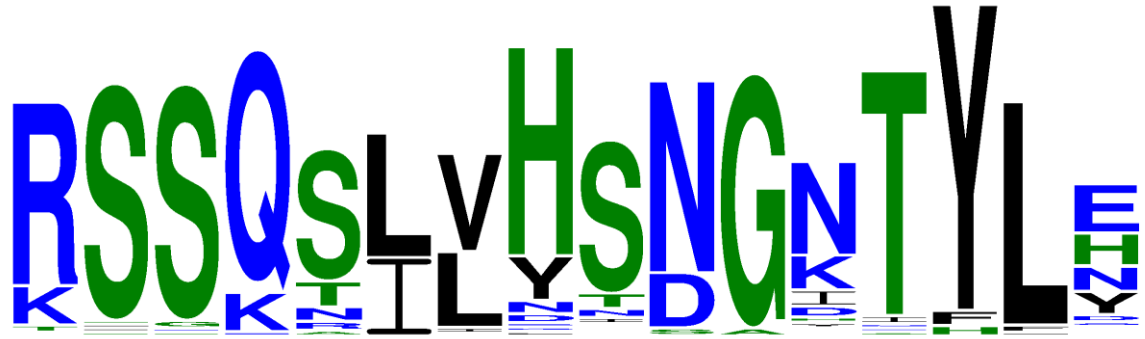


L1-13,14-A

L1-13-A

L1-14-A

L1-15-A



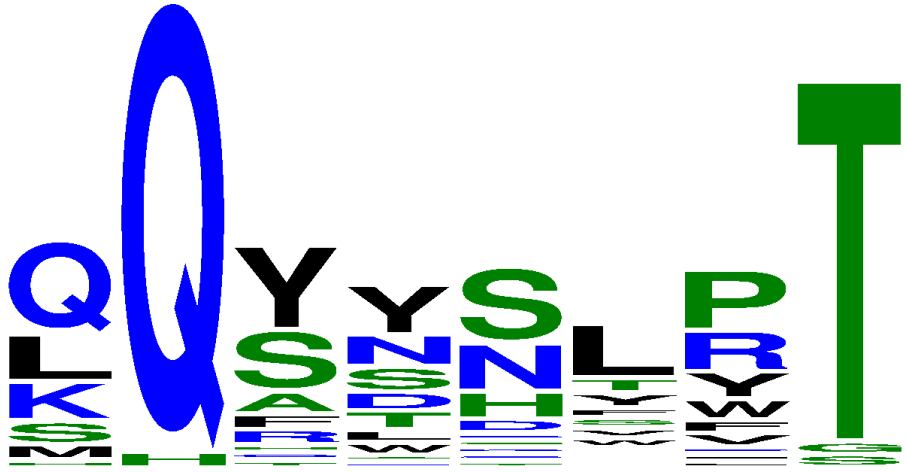
L1-16-A

L1-17-A

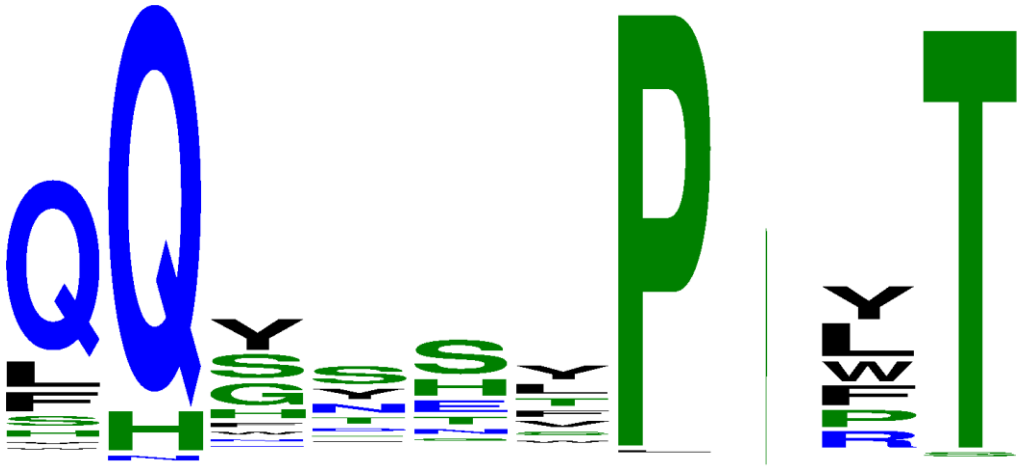
L2-7-A

L2-7-B

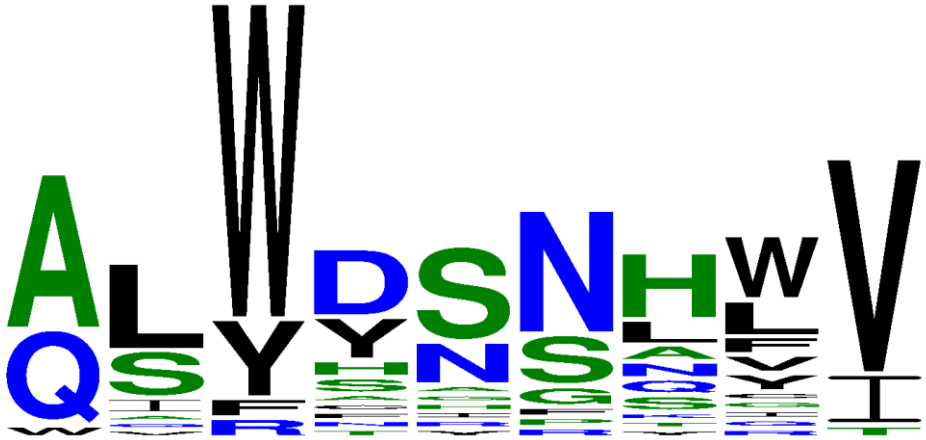
L3-5-A



L3-8-A



L3-9,10-A



L3-9-A

L3-10,11-A

A2: Full cluster tables

Cluster name	Length	Number of structures	Middle structure	Number of unique sequences	Species	Germline
L1-10,11,12-A	10	69	4F33_E	28	Mo	IGKV4
	11	707	3SOB_L	174	Mo, Hu, Ra, Rabbit	IGKV
	12	3	3EOO_A	2	Hu	IGKV3
L1-11-A	11	38	4IMK_C	9	Hu	IGLV3
L1-11-B	11	24	3MLS_M	8	Hu	IGLV3
L1-11-C	11	5	4FQC_L	3	Hu	IGLV3
L1-12-A	12	22	1HQ4_A	12	Hu, Mo	Hu_IGKV, Mo_IGKV4
L1-12-B	12	28	4LLV_L	5	Hu	IGKV3-20*01
L1-12-C	12	13	2OTU_E	3	Mo	IGLV3
L1-12-D	12	7	1HZH_M	3	Hu	IGKV3
L1-13,14-A	13	66	4FQJ_L	23	Hu	IGLV1
	14	51	3U2S_L	14	Hu	IGLV2
L1-13-A	13	23	2WOL_C	6	Hu	IGLV6-57*01
L1-14-A	14	92	1YOL_C	7	Hu, Mo	Mo_IGLV1, Hu_IGLV7
L1-14-B	14	2	4KTD_L	2	Hu	Hu_IGLV5
L1-15-A	15	55	3QRG_L	26	Hu, Mo	IGKV
L1-15-B	15	3	3DGG_C	2	Mo	IGKV3-12*01
L1-16-A	16	273	1KFA_M	65	Hu, Mo	IGKV
L1-17-A	17	113	2R1X_A	31	Hu, Mo	Mo_IGKV8, Hu_IGKV4
L1-17-B	17	3	3LDB_B	2	Ra, Mo	IGKV8
Unclustered	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17	104	-	47	-	-

Table S1: Information on canonical forms of CDR-L1. The clusters are ordered first by length, then by number of structures and finally by number of sequences. The germlines are reported in the following way: if over 90% of the structures come from the same allele, that allele is shown in the Table. Otherwise, we report the lowest step in the classification hierarchy that can explain 90% of the data. If two different loci are present in a cluster we report both of them.

Cluster name	Length	Number of structures	Middle structure	Number of unique sequences	Species	Germline
L2-7-A	7	1708	2G5B_A	291	Hu, Mo, Ra,	IGKV, IGLV
L2-7-B	7	21	3I9G_L	6	Hu, Mo	IGKV, IGLV
L2-7-C	7	6	2ABI_A	3	Mo	IGKV10-96*01
L2-11-A	11	13	2GSG_A	2	Mo	IGLV3*01
L2-11-B	11	5	2H3N_C	3	Rhesus monkey, Hu	IGLV, Hu_VPREB1*02
Unclustered	7	9	-	6	-	-

Table S2: Information on canonical forms of CDR-L2. See description below Table S1

Cluster name	Length	Number of Structures	Middle structure	Number of unique structures	Species	Framework germline	Joining germline
L3-5-A	5	17	4JPI_B	6	Hu	IGKV3, IGLV2	IGLJ, IGKJ
L3-7-A	7	2	1DFB_L	2	Hu	IGKV01-5*03	IGKJ3
L3-8-A	8	106	4HGW_A	29	Hu, Mo, Ra	IGKV	IGKJ
L3-8-B	8	9	3VW3_L	4	Mo	IGKV1-117*01	IGKJ
L3-8-C	8	6	2FD6_L	2	Mo	IGKV	IGKJ
L3-8-D	8	3	1TZH_A	2	Mo	IGKV12-44*01	IGKJ
L3-9,10-A	9	1123	3RVV_C	331	Hu, Mo, Ra	IGKV	IGKJ
	10	10	4HHA_A	4	Hu, Mo	Mo_IGKV1-110*01, Hu_IGKV3-11*01	IGKJ5*01
L3-9-A	9	107	1Y0L_C	22	Hu, Mo	IGLV, IGKV4-86*01	IGLJ, IGKL5*01
L3-9-B	9	5	2VXS_O	2	Hu	IGLV6-57*01	IGLJ
L3-9-C	9	2	2HWZ_L	2	Mo	IGKV1-117*01	IGKJ4
L3-10,11-A	10	4	3MLX_L	1	Hu	IGLV1-51*02	IGLJ
	11	49	4NZT_L	22	Hu	IGLV1, IGLV3-21	IGLJ
L3-10-A	10	31	3U2S_L	5	Hu	IGLV2-14*01	IGLJ
L3-10-B	10	9	3U79_B	2	Hu	IGKV1-33*01	IGKJ
L3-10-C	10	4	2DD8_L	3	Hu	IGLV3	IGLJ
L3-10-D	10	3	4JAM_L	2	Hu	IGLV3	IGLJ6
L3-11-A	11	3	3MA9_L	3	Hu	IGLV3, IGKJ1	IGLJ, IGKJ1
L3-12-A	12	4	3QHZ_L	2	Hu	IGLV1-44*01	IGLJ
L3-12-B	12	2	4JY5_L	2	Hu	IGLV3-21	IGLJ
L3-13-A	13	5	2BOS_L	2	Hu	IGLV1-47*01	IGLJ
Unclustered	6, 8, 9, 10, 11, 12 , 13, 19	248	-	80	-	-	-

Table S3: Information on canonical forms of CDR-L3. See description below Table S1

Cluster name	Length	Number of structures	Middle structure	Number of unique sequences	Species	Germline
H1-4-A	4	8	1KXQ_H	3	Camel	IGHV1S45*01
H1-6-A	6	7	2QQQ_B	2	Channel catfish	NITR11
H1-7-A	7	1267	1PLG_H	257	Hu, Mo, Ra, Camel, Llama, Rabbit, Rhesus Monkey, Sheep, Hamster	IGHV
H1-7-B	7	18	4FQQ_F	6	Hu, Mo	Hu_IGHV, Mo_IGHV5S21*01
H1-7-C	7	10	4KPH_H	2	Mo	IGHV3-8*02
H1-7-D	7	8	1BZQ_K	3	Camel, Llama	IGHV1S45
H1-7-E	7	7	4DKA_A	2	Llama	IGV1S3*01
H1-7-F	7	6	4NBZ_D	3	Llama, Mo	Mo_IGHV9-1*01, Llama_IGHV1S3*01
H1-7-G	7	6	3EZJ_B	3	Camel, Llama	IGHV
H1-8-A	8	37	3RVW_D	8	Hu, Mo	Hu_IGHV4-38-2*01, Mo_IGHV3
H1-8-B	8	14	1RVK_H	3	Mo	Mo_IGHV3-2*02
H1-8-C	8	7	1F58_H	2	Hu, Mo	Hu_IGHV4-38-2*01, Mo_IGHV3-1*02
H1-9-A	9	86	3IDN_B	9	Hu, Mo	Mo_IGHV8, Hu_IGHV2
H1-9-B	9	5	3BKJ_H	2	Mo	IGHV8-12*01
Unclustered	3,6,7, 8,9,10, 12,14	248	-	102	-	-

Table S4: Information on canonical forms of CDR-H1. See description below Table S1

Cluster name	Length	Number of structures	Middle structure	Number of unique sequences	Species	Germline
H2-7-A	7	387	3ZKM_H	91	Hu, Mo, Rat, Sheep, Camel, Llama, Rabbit	IGHV
H2-7-B	7	4	4FQC_H	3	Hu	IGHV4
H2-8-A	8	650	1I8M_B	197	Hu, Mo, Rat, Camel, Rhesus Monkey, Llama, Rabbit	IGHV
H2-8-B	8	305	2VXS_K	93	Hu, Mo, Camel, Rat, Llama	IGHV
H2-8-C	8	23	1ZLV_H	2	Hu	IGHV3
H2-8-D	8	19	1YQV_H	9	Hu, Mo	Hu_IGHV1, Mo_IGHV1
H2-8-E	8	8	3OGO_E	3	Hu, Rat, Camel	IGHV
H2-8-F	8	7	2XVM_B	2	Hu, Llama	Hu_IGHV1, Llama_IGHV1S4
H2-8-G	8	6	1ZA6_B	3	Mo	IGHV1
H2-8-H	8	4	3EYV_H	2	Mo, Hu	IGHV
H2-8-I	8	3	3QOS_B	3	Hu, Mo	Mo_IGHV5, Hu_IGHV3
H2-8-J	8	3	1F4X_H	2	Mo	IGHV5
H2-10-A	10	147	3HZV_B	25	Mo, Hu, Ra, Hamster	IGHV
Unclustered	8,9,10, 11,12	213	-	103	-	-

Table S5: Information on canonical forms of CDR-H2. See description below Table S1

A3: CLUSTERING COMPARISON TABLES

This file shows the detailed comparison between our length-independent clustering and the recent clustering of CDR structures by North & Dunbrack *et al.*

Our cluster	North & Dunbrack *et al.*	Fraction of CDRs in North & Dunbrack *et al.* cluster found in our cluster
L1-10,11,12-A	**L1-10-1**	**95%**
	L1-11-1	**100%**
	L1-11-1	**100%**
L1-11-A	**L1-11-3**	**40%**
L1-11-B	**L1-11-3**	**20%**
L1-11-C	-	-
L1-12-A	**L1-12-1**	**100%**
L1-12-B	L1-12-2	80%
L1-12-C	L1-12-3	100%
L1-12-D	-	-
L1-13,14-A	**L1-13-1**	**83%**
	L1-14-2	**75%**
L1-13-A	**L1-13-2**	**100%**
L1-14-A	**L1-14-1**	**100%**
L1-14-B	-	-
L1-15-A	**L1-15-1**	**100%**
L1-15-B	L1-15-2	100%
L1-16-A	**L1-16-1**	**90%**
L1-17-A	**L1-17-1**	**94%**
L1-17-B	-	-

Table S6: The comparison between North & Dunbrack *et al.* work and our clustering for CDR-L1. The first column shows the cluster label in our set while the second contains the labels of corresponding North & Dunbrack *et al.* clusters. The third column shows the fraction of CDRs in North & Dunbrack *et al.* cluster that is contained within our cluster. Two clusters were considered equivalent if our cluster contained at least 50% of CDRs present in the corresponding North & Dunbrack *et al.* cluster. North & Dunbrack *et al.* cluster was considered to be split between two of our clusters if the two clusters together contained at least 50% of CDRs from North & Dunbrack *et al.* cluster (for example the CDRs from North & Dunbrack *et al.* cluster L1-11-3 are split between our clusters L1-11-A and L1-11-B). The clusters in our work which contain at least six unique sequences are shown in bold

Our cluster	North & Dunbrack *et al.*	Fraction of CDRs in North & Dunbrack *et al.* cluster found in our cluster
L2-7-A	**L2-8-1**	**100%**
	L2-8-2	**100%**
	L2-8-4	**100%**
	L2-8-5	**100%**
L2-7-B	**L2-8-3**	**100%**
L2-7-C	-	-
L2-11-A	L2-12-2	100%
L2-11-B	L2-12-1	100%

Table S7: The comparison between North & Dunbrack *et al.* work and our clustering for CDR-L2. See description under Table S6

Our cluster	North & Dunbrack *et al.*	Fraction of CDRs in North & Dunbrack *et al.* cluster found in our cluster
L3-5-A	**-**	**-**
L3-7-A	L3-7-1	100%
L3-8-A	**L3-8-1**	**93%**
L3-8-B	L3-8-2	50%
L3-8-C	L3-8-cis6-1	66%
L3-8-D	L3-8-cis6-1	33%
L3-9,10-A	**L3-9-2**	**100%**
	L3-9-cis7-1	**98%**
	L3-9-cis7-2	**88%**
	L3-9-cis7-3	**100%**
	L3-10-cis7and8-1	**100%**
L3-9-A	**L3-9-1**	**68%**
	L3-9-cis6-1	**100%**
L3-9-B	-	-
L3-9-C	-	-
L3-10,11-A	**L3-11-1**	**63%**
L3-10-A	-	-
L3-10-B	-	-
L3-10-C	-	-
L3-10-D	-	-
L3-11-A	-	-
L3-12-A	-	-
L3-12-B	-	-
L3-13-A	-	-

Table S8: The comparison between North & Dunbrack *et al.* work and our clustering for CDR-L3. See description under Table S6

Our cluster	North & Dunbrack *et al.*	Fraction of CDRs in North & Dunbrack *et al.* cluster found in our cluster
H1-4-A	H1-10-1	100%
H1-6-A	-	-
H1-7-A	**H1-13-1**	**98%**
	H1-13-2	**57%**
H1-7-B	**-**	**-**
H1-7-C	H1-13-7	67%
H1-7-D	-	-
H1-7-E	H1-13-5	33%
H1-7-F	-	-
H1-7-G	H1-13-5	33%
H1-8-A	**H1-14-1**	**91%**
H1-8-B	-	-
H1-8-C	-	-
H1-9-A	**H1-15-1**	**78%**
H1-9-B	-	-

Table S9: The comparison between North & Dunbrack *et al.* work and our clustering for CDR-H1. See description under Table S6

Our cluster	North & Dunbrack *et al.*	Fraction of CDRs in North & Dunbrack *et al.* cluster found in our cluster
H2-7-A	**H2-9-1**	**96%**
H2-7-B	-	-
H2-8-A	**H2-10-1**	**95%**
H2-8-B	**H2-10-2**	**90%**
H2-8-C	-	-
H2-8-D	**H2-10-3**	**66%**
H2-8-E	-	-
H2-8-F	-	-
H2-8-G	H2-10-3	22%
H2-8-H	H2-10-6	50%
H2-8-I	-	-
H2-8-J	-	-
H2-10-A	**H2-12-1**	**100%**

Table S10: The comparison between North & Dunbrack *et al.* work and our clustering for CDR-H2. See description under Table S6

A4: Datasets of loop structures

Table S1: The general protein loop targets (dataset 1). The table list the loop structure prediction targets from our general protein dataset. The first column shows the PDB id of the protein's crystal structure, the second column shows the chain identifier, the third column shows the first residue of the loop, the fourth column shows the last residue of the loop and the last column shows the length of the loop.

PDB id	Chain	Start residue	End residue	Length
1smr	A	113	118	6
4j3v	A	313	318	6
3u1l	A	78	83	6
2jc5	A	150	155	6
4l6d	A	270	275	6
1trb	A	289	294	6
1q35	A	237	242	6
1vhe	A	267	272	6
1s9r	A	348	353	6
1ofd	A	218	223	6
4x00	A	175	182	8
3pvk	A	176	183	8
1tke	A	47	54	8
3gve	A	59	66	8
1f46	A	118	125	8
2car	A	103	110	8
2yna	A	241	248	8
2bw4	A	221	228	8
2z9w	A	115	122	8
1n08	A	85	92	8
3khy	A	165	174	10
2cfc	A	208	217	10
4ci7	A	480	489	10
2oem	A	271	280	10
2dri	A	89	98	10
3kwe	A	130	139	10
1pg4	A	450	459	10
1nhs	A	362	371	10
4gek	A	116	125	10
3dci	A	117	126	10
2ddx	A	155	166	12
2nuw	A	7	18	12
1eok	A	170	181	12
4i3g	A	198	209	12
4g29	A	294	305	12
1vhq	A	148	159	12
4gwi	A	81	92	12
3amn	A	121	132	12

4ag1	A	56	67	12
4mjk	A	145	156	12
3w8k	A	200	213	14
4r3f	A	66	79	14
1gp0	A	1593	1606	14
2gdm	A	44	57	14
4p6b	A	277	290	14
2q4w	A	422	435	14
1kv9	A	288	301	14
4wks	C	715	728	14
3f4l	A	153	166	14
2axq	A	321	334	14
2r16	A	857	872	16
4fe9	A	326	341	16
1ei5	A	139	154	16
2ddb	A	18	33	16
1rgz	A	112	127	16
4j94	A	133	148	16
1qcx	A	12	27	16
3r6a	A	26	41	16
1eok	A	127	142	16
4lih	A	453	468	16
4x84	A	86	103	18
2y8t	A	127	144	18
1rp0	A	71	88	18
4p3f	A	96	113	18
4cc2	A	1521	1538	18
2ocg	A	78	95	18
2oct	A	32	49	18
3t8w	A	133	150	18
4jic	A	268	285	18
1v4a	A	212	229	18

Table S2: The antibody CDR-H3 targets (datasets 2 and 3). The table list the loop structure prediction targets from our antibody datasets. The first column shows the PDB id of the antibody's crystal structure, the second column shows the chain identifier, the third column shows the first residue of the CDR-H3 (in Chothia numbering), the fourth column shows the last residue of the CDR-H3 (in Chothia numbering) and the last column shows the length of the CDR-H3.

PDB id	Chain	Start residue	End residue	Length
2ddq	H	95	102	4
1dqq	D	95	102	5
1z3g	H	95	102	6
1tet	H	95	102	7
2c1p	H	95	102	7
1bql	H	95	102	7
1mlb	B	95	102	7
1cgs	H	95	102	7
1jpt	H	95	102	8
1fgn	H	95	102	8
1a6t	D	95	102	8
1vfa	B	95	102	8
1iqd	B	95	102	8
1kem	H	95	102	8
2bdn	H	95	102	8
1qbl	H	95	102	8
2fd6	H	95	102	9
2jel	H	95	102	9
2adf	H	95	102	9
2fbj	H	95	102	9
2aep	H	95	102	9
1k4c	A	95	102	9
1igt	B	95	102	9
1jhl	H	95	102	9
1ynt	D	95	102	9
1kb5	H	95	102	10
2aju	H	95	102	10
1dba	H	95	102	10
2cju	H	95	102	10
1ztx	H	95	102	10
1for	H	95	102	10
1clz	H	95	102	10
2b2x	H	95	102	10
2fjg	H	95	102	11
2adg	B	95	102	11
1fpt	H	95	102	11
1mcp	H	95	102	11
2h1p	H	95	102	11
1nca	H	95	102	11

1igm	H	95	102	12
2fjh	H	95	102	12
2h2p	C	95	102	12
2g5b	H	95	102	12
1fbi	H	95	102	13
1wc7	H	95	102	14
2aj3	B	95	102	14
1zan	H	95	102	14
1bj1	H	95	102	14
2dqu	H	95	102	14
1f58	H	95	102	17
1hzh	H	95	102	18
1g9m	H	95	102	19
2b4c	H	95	102	22

A5: INTERFACE RESIDUES

Table S1: The list of interface residues. The table shows the VH-VL interface residues, identified by the program NACCESS, for the purpose of Fv interface orientation prediction.

Chain	Residue IMGT number	Insertion code
VH	1	
VH	3	
VH	4	
VH	38	
VH	40	
VH	42	
VH	44	
VH	46	
VH	47	
VH	48	
VH	49	
VH	50	
VH	51	
VH	52	
VH	53	
VH	54	
VH	55	
VH	57	
VH	63	
VH	64	
VH	65	
VH	66	
VH	67	
VH	68	
VH	69	
VH	70	
VH	72	
VH	101	
VH	103	
VH	105	
VH	106	
VH	107	
VH	108	
VH	109	
VH	110	
VH	111	
VH	111	A
VH	111	B
VH	111	C
VH	111	D
VH	111	E

VH	112	E
VH	112	D
VH	112	C
VH	112	B
VH	112	A
VH	112	
VH	113	
VH	114	
VH	115	
VH	116	
VH	117	
VH	118	
VH	119	
VH	120	
VH	121	
VH	122	
VH	123	
VL	1	
VL	2	
VL	3	
VL	4	
VL	5	
VL	28	
VL	29	
VL	31	
VL	32	
VL	33	
VL	34	
VL	36	
VL	37	
VL	38	
VL	39	
VL	40	
VL	41	
VL	42	
VL	44	
VL	45	
VL	46	
VL	47	
VL	48	
VL	49	
VL	50	
VL	51	
VL	52	

VL	53	
VL	54	
VL	55	
VL	56	
VL	57	
VL	58	
VL	63	
VL	64	
VL	65	
VL	66	
VL	67	
VL	68	
VL	69	
VL	70	
VL	71	
VL	81	
VL	101	
VL	103	
VL	105	
VL	106	
VL	107	
VL	108	
VL	109	
VL	110	
VL	112	
VL	113	
VL	114	
VL	115	
VL	116	
VL	117	
VL	118	
VL	119	
VL	120	
VL	121	
VL	123	