Cell Reports, Volume *31*

Supplemental Information

Landscape of Non-canonical Cysteines in Human V_H Repertoire Revealed by Immunogenetic Analysis

Ponraj Prabakaran and Partha S. Chowdhury

Supplemental Information

Immunogenetic Analysis Reveals the Landscape of Non-Canonical Cysteines in Human V_H repertoire

Ponraj Prabakaran and Partha S. Chowdhury

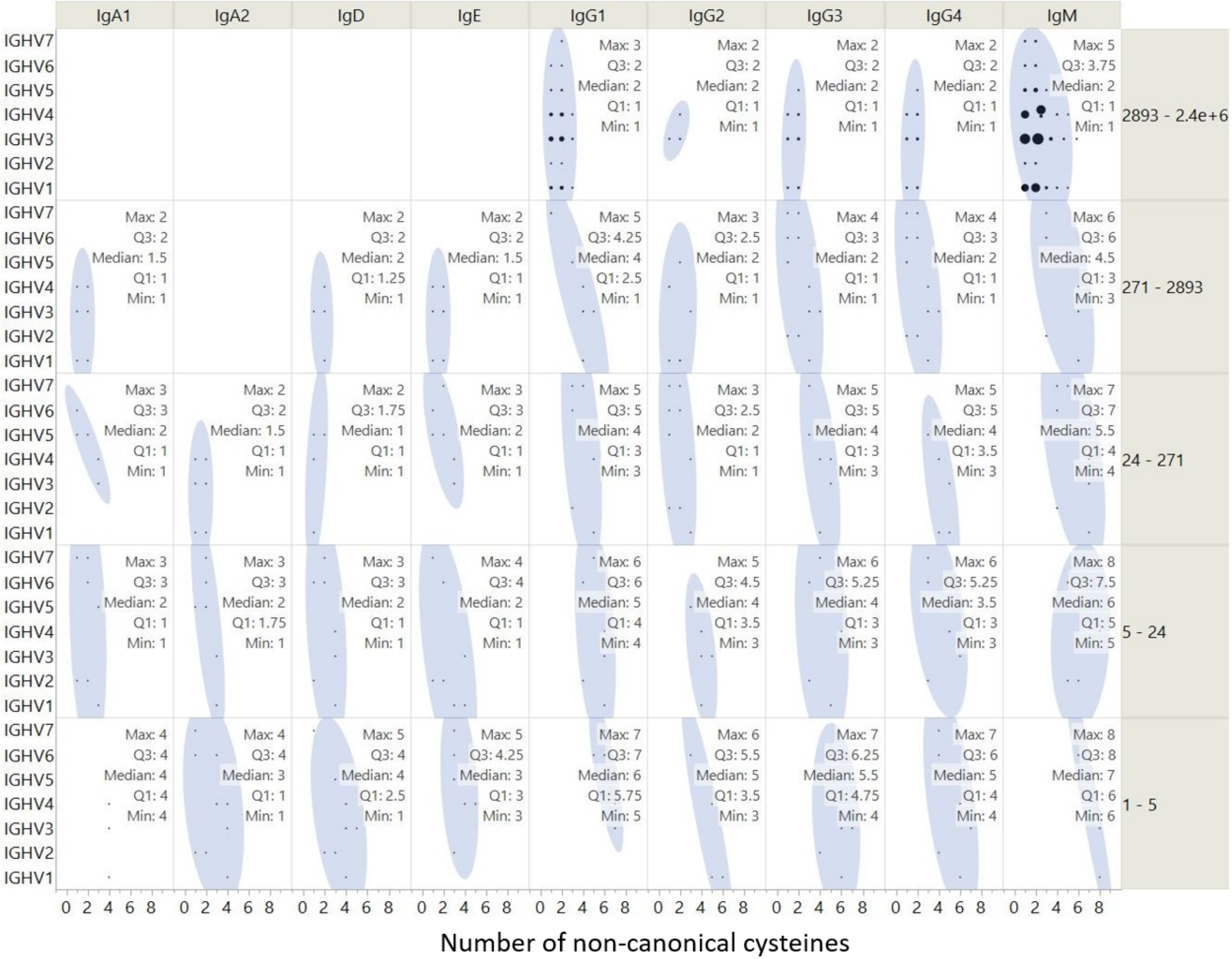


Figure S1. Bivariate normal density plot showing the association of non-canonical cysteine containing human CDR-H3s with different IGHV gene families and isotypes, Related to Figure 1, Table S1 and RESULTS section: Immunogenetic Analysis Reveals High Frequency, Extensive Diversity and Recurring Patterns of Non-Canonical Cysteines

IGHV gene usage and Ig isotype diversity observed in 12,054,263 human VH sequences from dataset A are shown by bivariate normal density plots with a 90% coverage along with statistical summary. Total counts are shown on the right side.

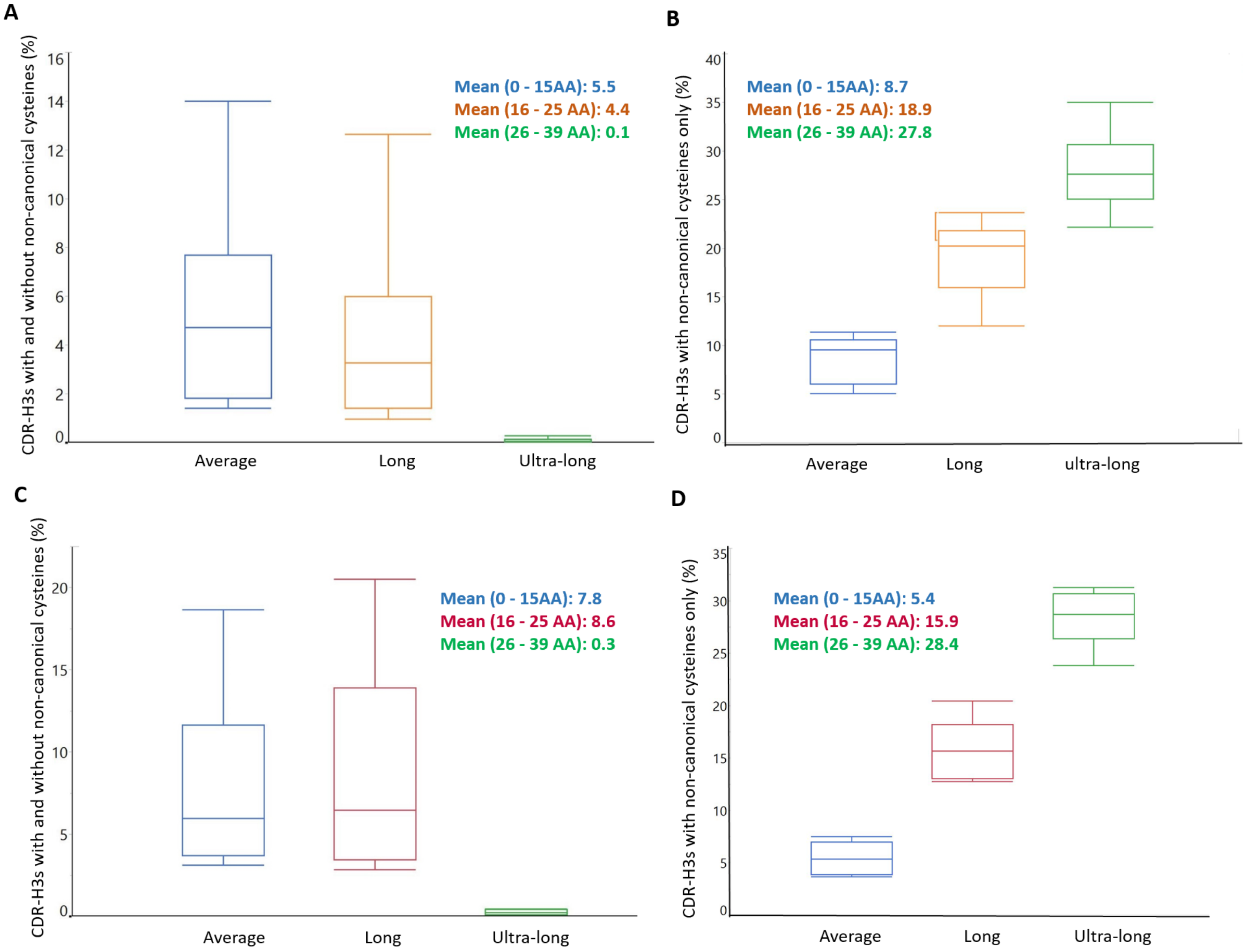


Figure S2. Box plots showing percentages of antibodies that were grouped into three different CDR-H3 length categories, average (up to 15 AA), long (16-25 AA) and ultra-long (26-39 AA), Related to Figure 1B and Tables S1 and S2

(A and B) The percentage of antibodies of different CDR-H3 length categories for dataset A in all sequences, with and without non-canonical cysteines, (A) and in non-canonical cysteine containing CDR-H3s only (B).

(C and D) The percentage of antibodies of different CDR-H3 length categories for dataset B in all sequences, with and without non-canonical cysteines, (C) and in non-canonical cysteine containing CDR-H3s only (D).

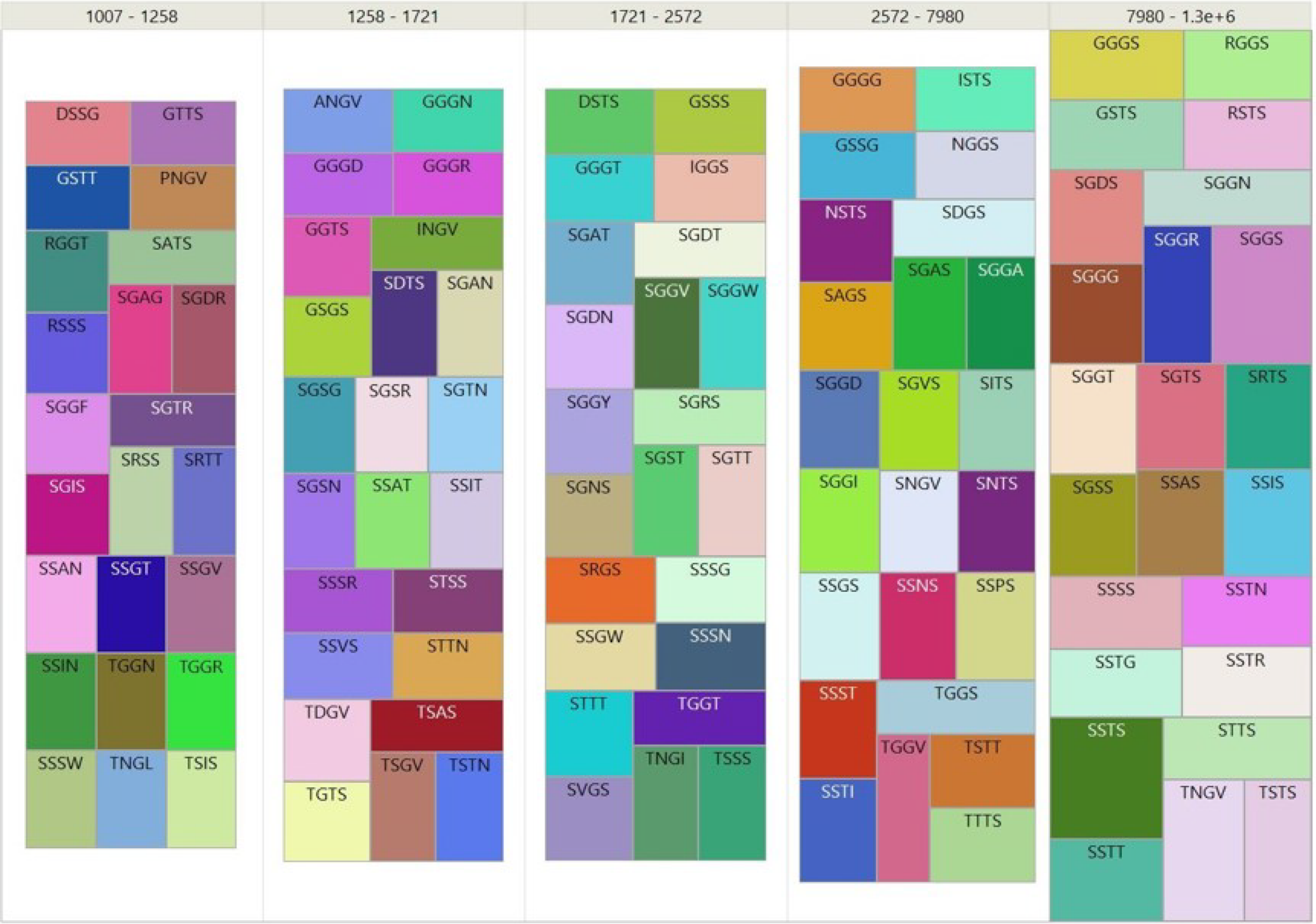


Figure S3, Related to Figure 4
Treemapping of 118 high-frequency tetrapeptides within CX$_4$C motifs of human CDR-H3s appearing more than 1000 times, as observed in dataset A, is shown with frequencies at the top.

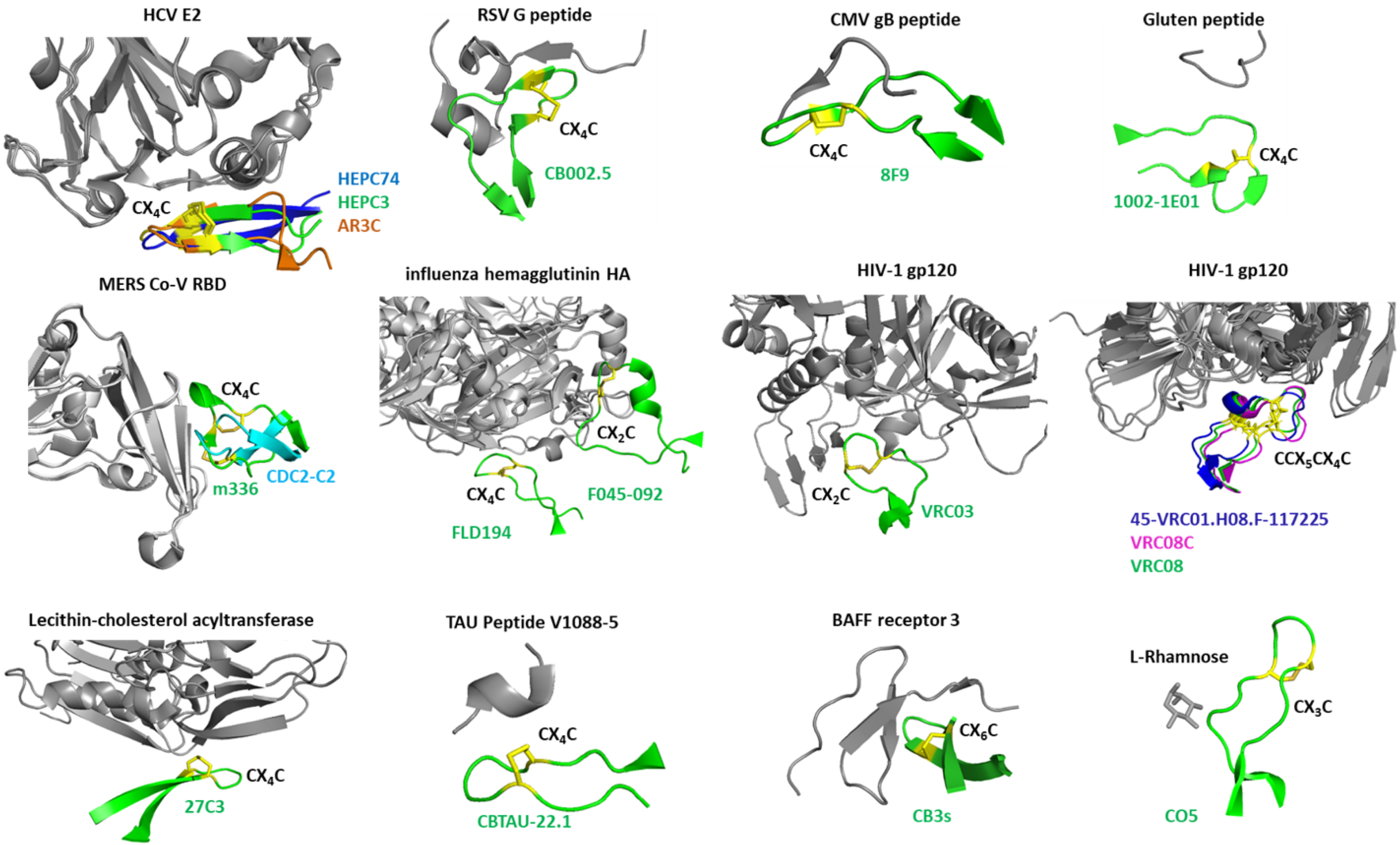
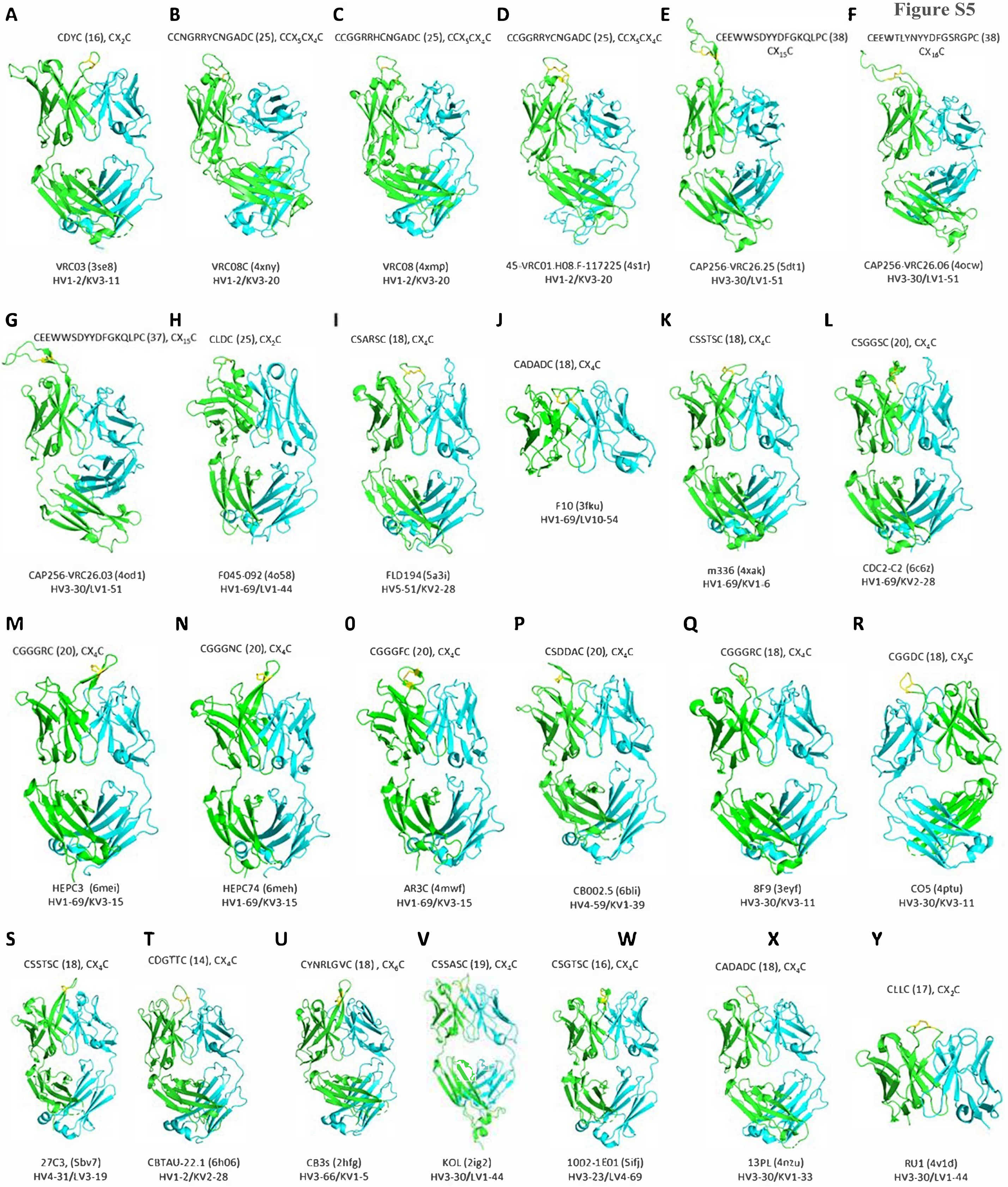


Figure S4. Structurally known disulfide-bonded motifs in CDR-H3s of human antibodies in complex with diverse anti-viral and other antigens, Related to Figure 3A and Results section: CX_nC Motifs Play a Determining Role in the Structure and Function of Antibodies

Complex crystal structures showing the CDR-H3s (in colors) with intra-disulfide bonded motifs (yellow) in human antibodies targeting different antigens (gray) as found in the Protein Data Bank (PDB). CDR-H3s of antibodies with disulfide-bonded motifs recognizing different antigens are only shown. While CX_4C motif is found widespread in CDR-H3s, other motifs of types CX_2C, CX_3C, CCX_5CX_4C and CX_6C were also observed. See Figure S5 for more information on these and other uncomplexed antibodies containing disulfide-bonded cysteine motifs in CDR-H3s.

Figure S5



A
CDYC (16), CX$_2$C
VRC03 (3se8)
HV1-2/KV3-11
B
CCNGRRYCNGADC (25), CCX$_5$CX$_4$C
VRC08C (4xny)
HV1-2/KV3-20
C
CCGGRRHCNGADC (25), CCX$_5$CX$_4$C
VRC08 (4xmp)
HV1-2/KV3-20
D
CCGGRRYCNGADC (25), CCX$_5$CX$_4$C
45-VRC01.H08.F-117225 (4s1r)
HV1-2/KV3-20
E
CEEWWSDYYDFGKQLPC (38)
CX$_{15}$C
CAP256-VRC26.25 (5dt1)
HV3-30/LV1-51
F
CEEWTLYNYYDFGSRGPC (38)
CX$_{16}$C
CAP256-VRC26.06 (4ocw)
HV3-30/LV1-51
G
CEEWWSDYYDFGKQLPC (37), CX$_{15}$C
CAP256-VRC26.03 (4od1)
HV3-30/LV1-51
H
CLDC (25), CX$_2$C
F045-092 (4o58)
HV1-69/LV1-44
I
CSARSC (18), CX$_4$C
FLD194 (5a3i)
HV5-51/KV2-28
J
CADADC (18), CX$_4$C
F10 (3fku)
HV1-69/LV10-54
K
CSSTSC (18), CX$_4$C
m336 (4xak)
HV1-69/KV1-6
L
CSGGSC (20), CX$_4$C
CDC2-C2 (6c6z)
HV1-69/KV2-28
M
CGGGRC (20), CX$_4$C
HEPC3 (6mei)
HV1-69/KV3-15
N
CGGGNC (20), CX$_4$C
HEPC74 (6meh)
HV1-69/KV3-15
0
CGGGFC (20), CX$_4$C
AR3C (4mwf)
HV1-69/KV3-15
P
CSDDAC (20), CX$_4$C
CB002.5 (6bli)
HV4-59/KV1-39
Q
CGGGRC (18), CX$_4$C
8F9 (3eyf)
HV3-30/KV3-11
R
CGGDC (18), CX$_3$C
CO5 (4ptu)
HV3-30/KV3-11
S
CSSTSC (18), CX$_4$C
27C3, (5bv7)
HV4-31/LV3-19
T
CDGTTC (14), CX$_4$C
CBTAU-22.1 (6h06)
HV1-2/KV2-28
U
CYNRLGVC (18) , CX$_6$C
CB3s (2hfg)
HV3-66/KV1-5
V
CSSASC (19), CX$_4$C
KOL (2ig2)
HV3-30/LV1-44
W
CSGTSC (16), CX$_4$C
1002-1E01 (5ifj)
HV3-23/LV4-69
X
CADADC (18), CX$_4$C
13PL (4n2u)
HV3-30/KV1-33
Y
CLLC (17), CX$_2$C
RU1 (4v1d)
HV3-30/LV1-44

Figure S5, Related to Figures 3A and S4, and Results section: CXnC Motifs Play a Determining Role in the Structure and Function of Antibodies

Twenty-five human antibodies bearing a variety of disulfide-bonded CDR-H3s with distinctive IGHV/IGLV germline pairings, as analyzed from crystal structures available in the Protein Data Bank (PDB), are shown. These antibodies have longer CDR-H3s with lengths ranging from 16 to 38 AAs by IMGT numbering scheme. Heavy chains are in green, light chains in cyan and disulfide bonds in yellow. The sequence and type of cysteine motif along with CDR-H3 length are given at the top of each structure. Antibody name, PDB code and IGHV/IGLV germline information are given at the bottom of each structure. These antibodies target a wide range of antigens; (A-G) HIV, (H-J) Influenza, (K and L) MERS CoV, (M-O) HCV, (P) RSV, (Q) HCMV, (R) L-rhamnose of Streptococcus pneumoniae, (S) Lecithin cholesterol acyltransferase (LCAT), (T) Tau peptide, (U) BLyS receptor 3 (BR3), (V) Unknown, (W) Celiac disease-specific gluten peptide, (X) Protein M, (Y) Cn2 toxin from scorpion.

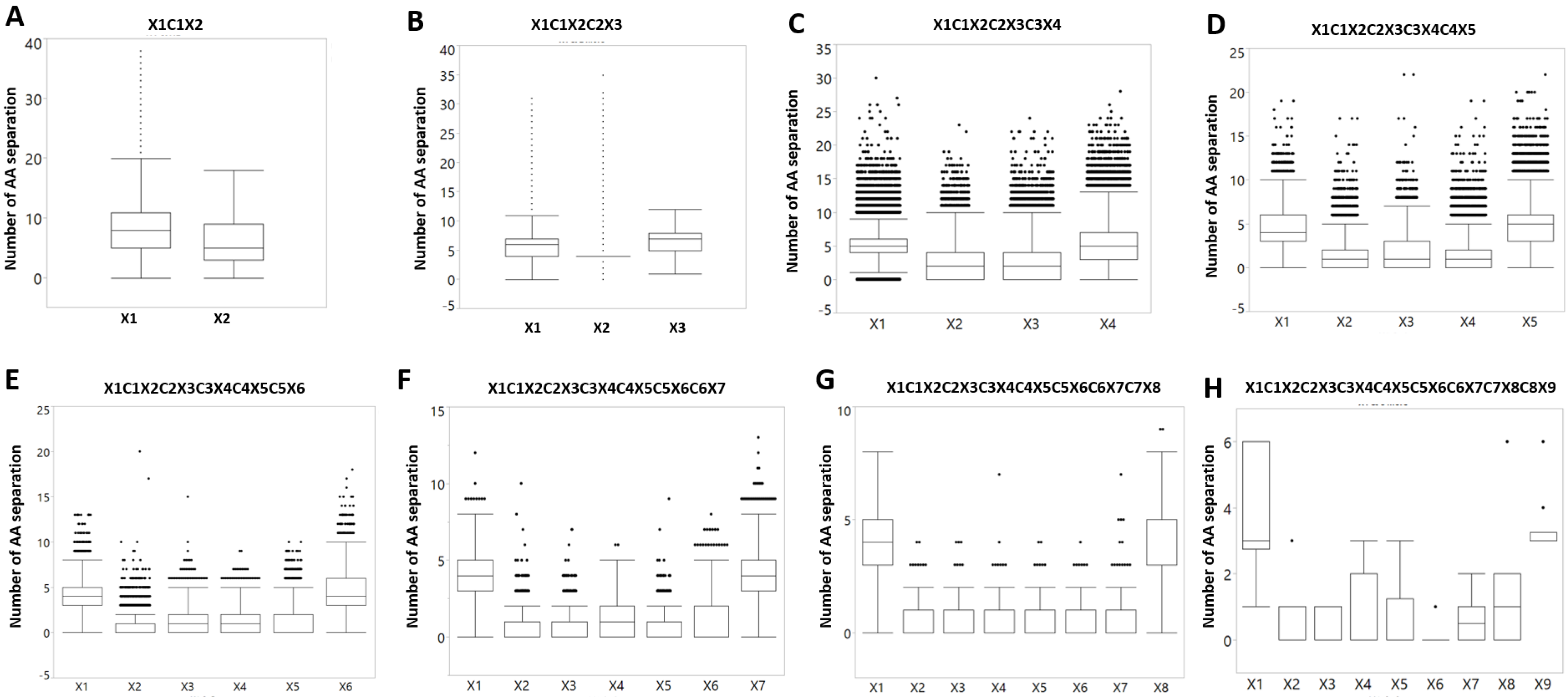


Figure S6. Boxplots showing the variations in number of AAs separating and/or flanking the cysteines in human CDR-H3s, Related to Figure 1C

(A-H) Distributions of number of AA separating and/or flanking the non-canonical cysteines observed in 8,792,995 unique CDR-H3s from dataset A. The cysteine motifs consisting of one to eight cysteines, as designated with C1 through C8 for cysteines and X1 through X9 for number of other AAs, are shown.

(A)

	IGHD6-13*01	IGHD2-21*01/*02	IGHD2-2*01/*02/*03	IGHD2-15*01	IGHD2-8*01	IGHD2-8*02
D-D fusion	CC	CGGDC	CSSTSC	CSGGSC	CTNGVC	CTGGVC
IGHD6-13*01 CC	15815	X	X	X	X	X
IGHD2-21*01/*02 CGGDC	349	7	X	X	X	X
IGHD2-2*01/*02/*03 CSSTSC	556	218	3	X	X	X
IGHD2-15*01 CSGGSC	545	149	584	13	X	X
IGHD2-8*01 CTNGVC	64	14	127	0	1	X
IGHD2-8*02 CTGGVC	3	1	2	0	0	0

(B)

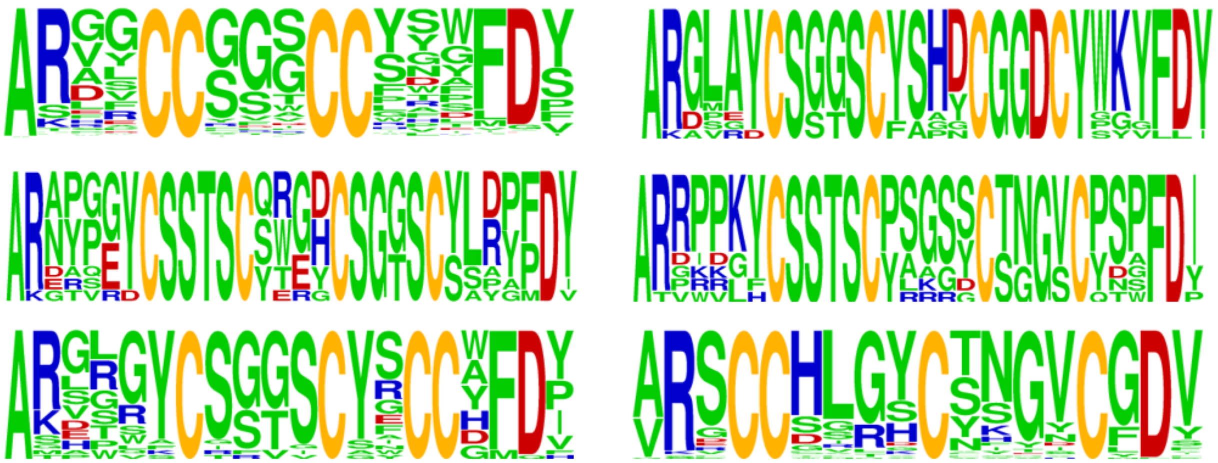
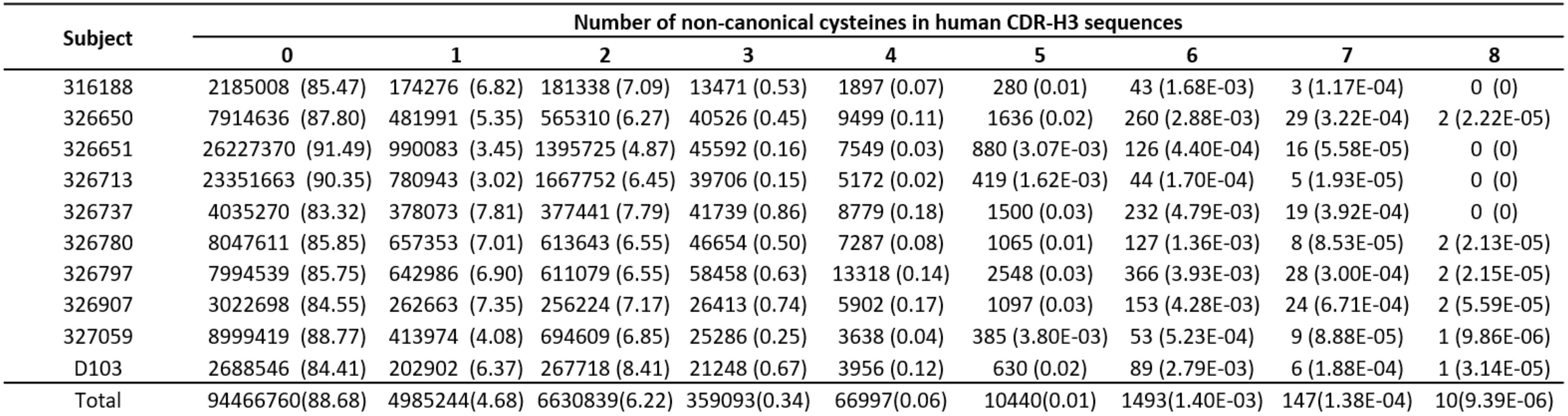


Figure S7. The D-D fusions occurring in four cysteine motifs of human CDR-H3s, Related to Figure 7 and Results Sections: Multiple Non-Canonical Cysteine Motifs Exist and Reveal Immunogenetic Mechanisms

(A) IGHD germline segments encoding two-cysteine motifs (CC, CX_3C, CX_4C) that could potentially undergo the V(DD)J recombination to create the four-cysteine motifs. Possible frequencies for the D-D fusions which might have occurred in the CDR-H3s were shown using the dataset A. Note that the total number of actual frequencies for such V(DD)J recombination with other IGHD germlines, with possible cysteines generated through SHM, could tremendously increase the four-cysteines motif landscape.

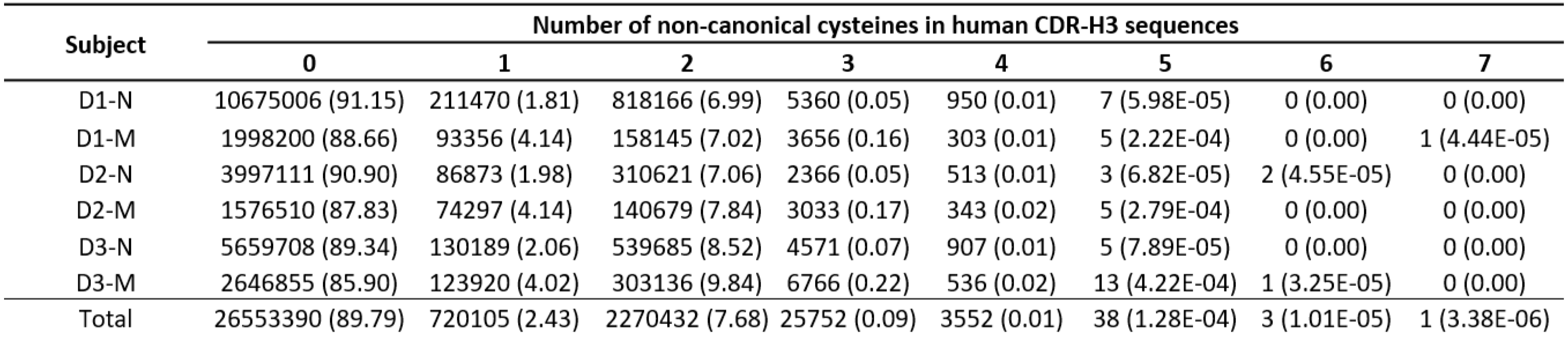
(B) WebLogos depicting the selected D-D fusions in several human CDR-H3s containing the four-cysteine motifs (dataset A).



Subject	Number of non-canonical cysteines in human CDR-H3 sequences								
	0	1	2	3	4	5	6	7	8
316188	2185008 (85.47)	174276 (6.82)	181338 (7.09)	13471 (0.53)	1897 (0.07)	280 (0.01)	43 (1.68E-03)	3 (1.17E-04)	0 (0)
326650	7914636 (87.80)	481991 (5.35)	565310 (6.27)	40526 (0.45)	9499 (0.11)	1636 (0.02)	260 (2.88E-03)	29 (3.22E-04)	2 (2.22E-05)
326651	26227370 (91.49)	990083 (3.45)	1395725 (4.87)	45592 (0.16)	7549 (0.03)	880 (3.07E-03)	126 (4.40E-04)	16 (5.58E-05)	0 (0)
326713	23351663 (90.35)	780943 (3.02)	1667752 (6.45)	39706 (0.15)	5172 (0.02)	419 (1.62E-03)	44 (1.70E-04)	5 (1.93E-05)	0 (0)
326737	4035270 (83.32)	378073 (7.81)	377441 (7.79)	41739 (0.86)	8779 (0.18)	1500 (0.03)	232 (4.79E-03)	19 (3.92E-04)	0 (0)
326780	8047611 (85.85)	657353 (7.01)	613643 (6.55)	46654 (0.50)	7287 (0.08)	1065 (0.01)	127 (1.36E-03)	8 (8.53E-05)	2 (2.13E-05)
326797	7994539 (85.75)	642986 (6.90)	611079 (6.55)	58458 (0.63)	13318 (0.14)	2548 (0.03)	366 (3.93E-03)	28 (3.00E-04)	2 (2.15E-05)
326907	3022698 (84.55)	262663 (7.35)	256224 (7.17)	26413 (0.74)	5902 (0.17)	1097 (0.03)	153 (4.28E-03)	24 (6.71E-04)	2 (5.59E-05)
327059	8999419 (88.77)	413974 (4.08)	694609 (6.85)	25286 (0.25)	3638 (0.04)	385 (3.80E-03)	53 (5.23E-04)	9 (8.88E-05)	1 (9.86E-06)
D103	2688546 (84.41)	202902 (6.37)	267718 (8.41)	21248 (0.67)	3956 (0.12)	630 (0.02)	89 (2.79E-03)	6 (1.88E-04)	1 (3.14E-05)
Total	94466760(88.68)	4985244(4.68)	6630839(6.22)	359093(0.34)	66997(0.06)	10440(0.01)	1493(1.40E-03)	147(1.38E-04)	10(9.39E-06)

Table S1, Related to Figure 1 and RESULTS section: Immunogenetic Analysis Reveals High Frequency, Extensive Diversity and Recurring Patterns of Non-Canonical Cysteines

Analysis of non-canonical cysteines in human CDR-H3 repertoire. A total of 106,521,023 CDR-H3 sequences of NGS data sets from ten subjects (dataset A) were retrieved and analyzed. The data sets were binned by number of non-canonical cysteines, 0 to 8, as observed in the CDR-H3s for each subject. The numbers under each column represents the number of unique sequences and the numbers in parenthesis represent % of total.



Subject	Number of non-canonical cysteines in human CDR-H3 sequences							
	0	1	2	3	4	5	6	7
D1-N	10675006 (91.15)	211470 (1.81)	818166 (6.99)	5360 (0.05)	950 (0.01)	7 (5.98E-05)	0 (0.00)	0 (0.00)
D1-M	1998200 (88.66)	93356 (4.14)	158145 (7.02)	3656 (0.16)	303 (0.01)	5 (2.22E-04)	0 (0.00)	1 (4.44E-05)
D2-N	3997111 (90.90)	86873 (1.98)	310621 (7.06)	2366 (0.05)	513 (0.01)	3 (6.82E-05)	2 (4.55E-05)	0 (0.00)
D2-M	1576510 (87.83)	74297 (4.14)	140679 (7.84)	3033 (0.17)	343 (0.02)	5 (2.79E-04)	0 (0.00)	0 (0.00)
D3-N	5659708 (89.34)	130189 (2.06)	539685 (8.52)	4571 (0.07)	907 (0.01)	5 (7.89E-05)	0 (0.00)	0 (0.00)
D3-M	2646855 (85.90)	123920 (4.02)	303136 (9.84)	6766 (0.22)	536 (0.02)	13 (4.22E-04)	1 (3.25E-05)	0 (0.00)
Total	26553390 (89.79)	720105 (2.43)	2270432 (7.68)	25752 (0.09)	3552 (0.01)	38 (1.28E-04)	3 (1.01E-05)	1 (3.38E-06)

Table S2, Related to Figure S2C and D, and RESULTS section: Immunogenetic Analysis Reveals High Frequency, Extensive Diversity and Recurring Patterns of Non-Canonical Cysteines

Non-canonical cysteines in human CDR-H3s of naïve (N) and memory (M) repertoires. The NGS data sets containing a total of 29,573,273 sequences from three subjects (dataset B) were analyzed. The data sets were binned by number of non-canonical cysteine residues, 0 to 7, as observed in the CDR-H3 sequences for each individual. The numbers under each column represents the number of unique sequences and the numbers in parenthesis represent % of total.

(A)

J00232	IGHD2-2*01	R I L * * Y Q L L **C** G Y **C** S S T S **C** Y A AGGATATTGTAGTAGTACCAGCTGCTATGCC
X97051	IGHD2-2*02	G Y **C** **S** **S** **T** **S** **C** Y T AGGATATTGTAGTAGTACCAGCTGCTATACC
M35648	IGHD2-2*03	W I L * * Y Q L L **C** G Y **C** **S** **S** **T** **S** **C** Y A TGGATATTGTAGTAGTACCAGCTGCTATGCC
X13972	IGHD2-8*01	R I L Y * W **C** M L Y G Y **C** **T** **N** **G** **V** **C** Y T AGGATATTGTACTAATGGTGTATGCTATACC
J00233	IGHD2-8*02	R I L Y W W **C** M L Y G Y **C** **T** **G** **G** **V** **C** Y T AGGATATTGTACTGGTGGTGTATGCTATACC
J00234	IGHD2-15*01	G Y **C** **S** **G** **G** **S** **C** Y S AGGATATTGTAGTGGTGGTAGCTGCTACTCC
J00235	IGHD2-21*01	A Y **C** G G D **C** Y S AGCATATTGTGGTGGTGATTGCTATTCC
X97051	IGHD2-21*02	A Y **C** G G D **C** Y S AGCATATTGTGGTGGTGACTGCTATTCC
X93615	IGHD3-10*02	V L L **C** S G S Y Y N GTATTACTATGTTCGGGGAGTTATTATAAC
X93614	IGHD3-16*01	V L * L R L G E L **C** L Y GTATTATGATTACGTTTGGGGGAGTTATGCTTATACC
X97051	IGHD1-1*01	R S S **C** T GTCGTTCCAGTTGTACC
J00235	IGHD2-21*01	G I A I T T T I **C** GGAATAGCAATCACCACCACAATATGCT
X97051	IGHD2-21*02	G I A V T T T I **C** GGAATAGCAGTCACCACCACAATATGCT
X93618	IGHD3-3*02	V * * P L Q K **C** * Y GGTATAATAACCACTCCAAAAATGCTAATAC
X13972	IGHD4-4*01	S Y **C** S GTAGTTACTGTAGTCA
X13972	IGHD4-11*01	S Y **C** S GTAGTTACTGTAGTCA
X13972	IGHD5-5*01	N H S **C** I H GTAACCATAGCTGTATCCAC
X97051	IGHD5-18*01	N H S **C** I H GTAACCATAGCTGTATCCAC
X97051	IGHD5-24*01	N **C** S H L Y GTAATTGTAGCCATCTCTAC
X13972	IGHD6-6*01	T S **C** Y T GGACGAGCTGCTATACTC
X13972	IGHD6-13*01	T S **C** **C** Y T GTACCAGCTGCTGCTATACCC
X97051	IGHD6-19*01	T S H **C** Y T GTACCAGCCACTGCTATACCC
X97051	IGHD6-25*01	S R **C** Y T GTAGCCGCTGCTATACCC

(B)

IGHD	Number of all CDR-H3s	(%)	Number of Cys containing CDR-H3s	(%)
IGHD1-1	1093660	1.03	47435	0.04
IGHD1-20	257626	0.24	9892	0.01
IGHD1-26	6014104	5.65	239936	0.23
IGHD1-7	1275693	1.20	46841	0.04
IGHD2-15	5876364	5.52	2367365	2.22
IGHD2-2	7810904	7.33	3995203	3.75
IGHD2-21	4076340	3.83	1047839	0.98
IGHD2-8	2628608	2.47	482875	0.45
IGHD3-10	13973853	13.12	687987	0.65
IGHD3-16	5001391	4.70	274312	0.26
IGHD3-22	11588386	10.88	743634	0.70
IGHD3-3	6999534	6.57	440238	0.41
IGHD3-9	3926989	3.69	202020	0.19
IGHD4-17	4777260	4.48	182354	0.17
IGHD4-4	1019379	0.96	43275	0.04
IGHD5-12	4327874	4.06	195328	0.18
IGHD5-5	4570933	4.29	220258	0.21
IGHD6-13	8481459	7.96	385882	0.36
IGHD6-19	8328442	7.82	467849	0.44
IGHD6-25	547368	0.51	26502	0.02
IGHD6-6	2854008	2.68	115072	0.11
IGHD7-27	1002322	0.94	27249	0.03

Table S3. IGHD gene segments in human V$_H$ repertoires, Related to Figure1 and RESULTS section: Immunogenetic Analysis Reveals High Frequency, Extensive Diversity and Recurring Patterns of Non-Canonical Cysteines

(A) IGHD germline segments that encode cysteines with their corresponding accession numbers from the IMGT database. * indicates a stop codon in the sequence.

(B) Total numbers of CDR-H3s as well as those that contain non-canonical cysteines involving different IGHD segments are given. IGHD2 segments encoding two cysteines (highlighted in gray) observed in human V$_H$ repertoire as seen in the dataset A.