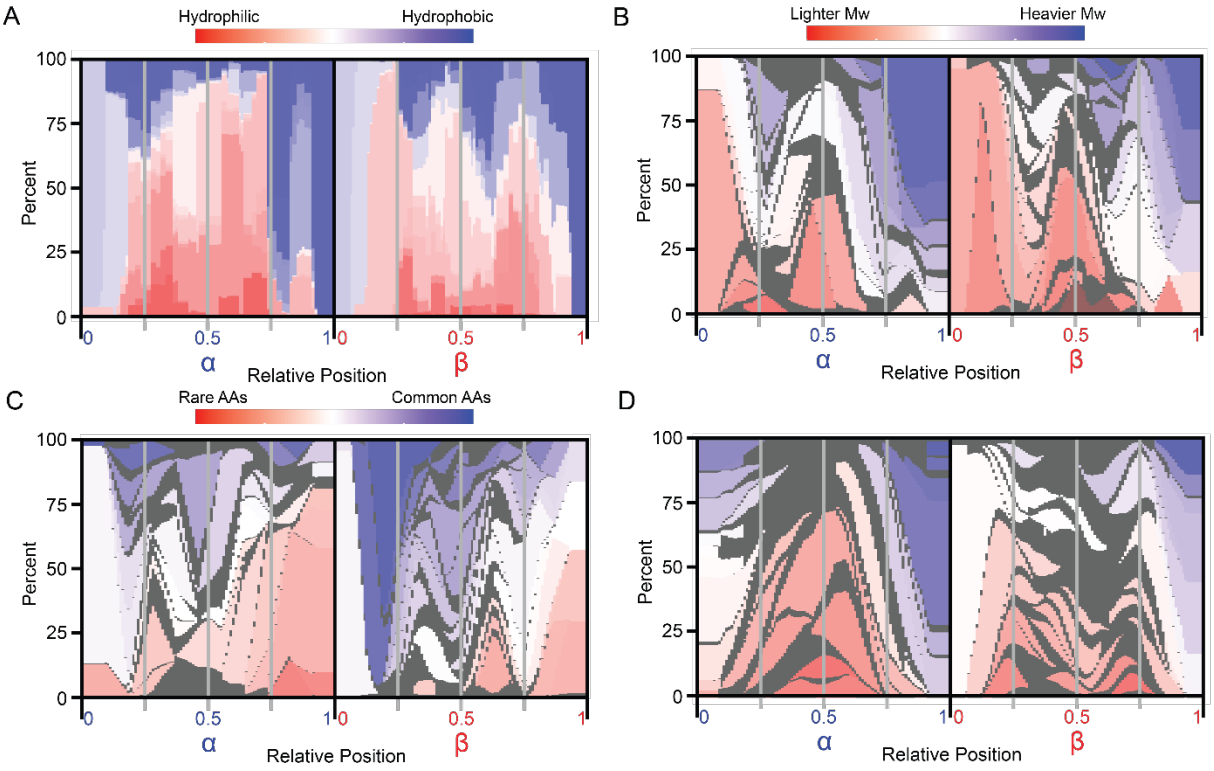
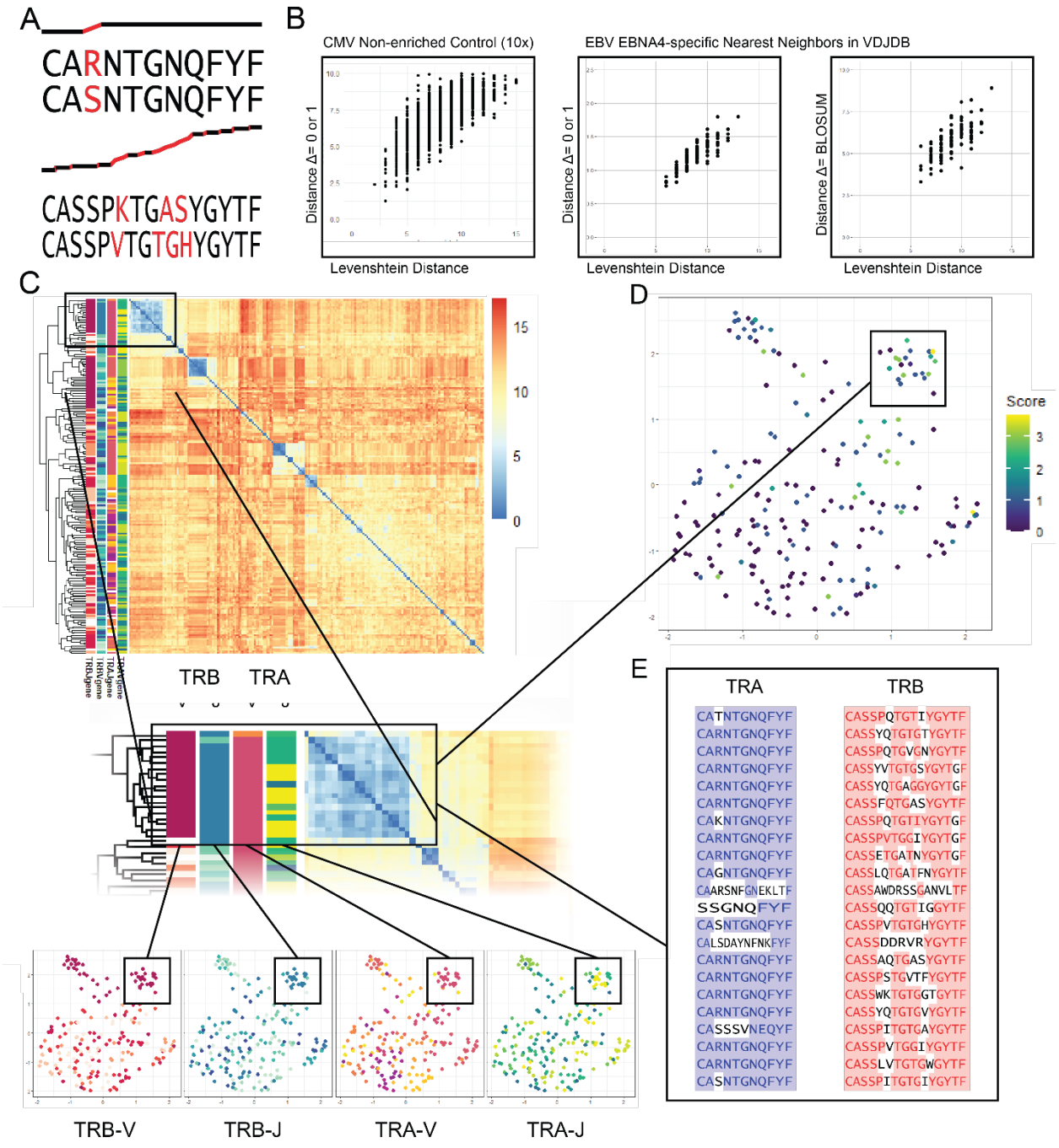
Supplementary Information

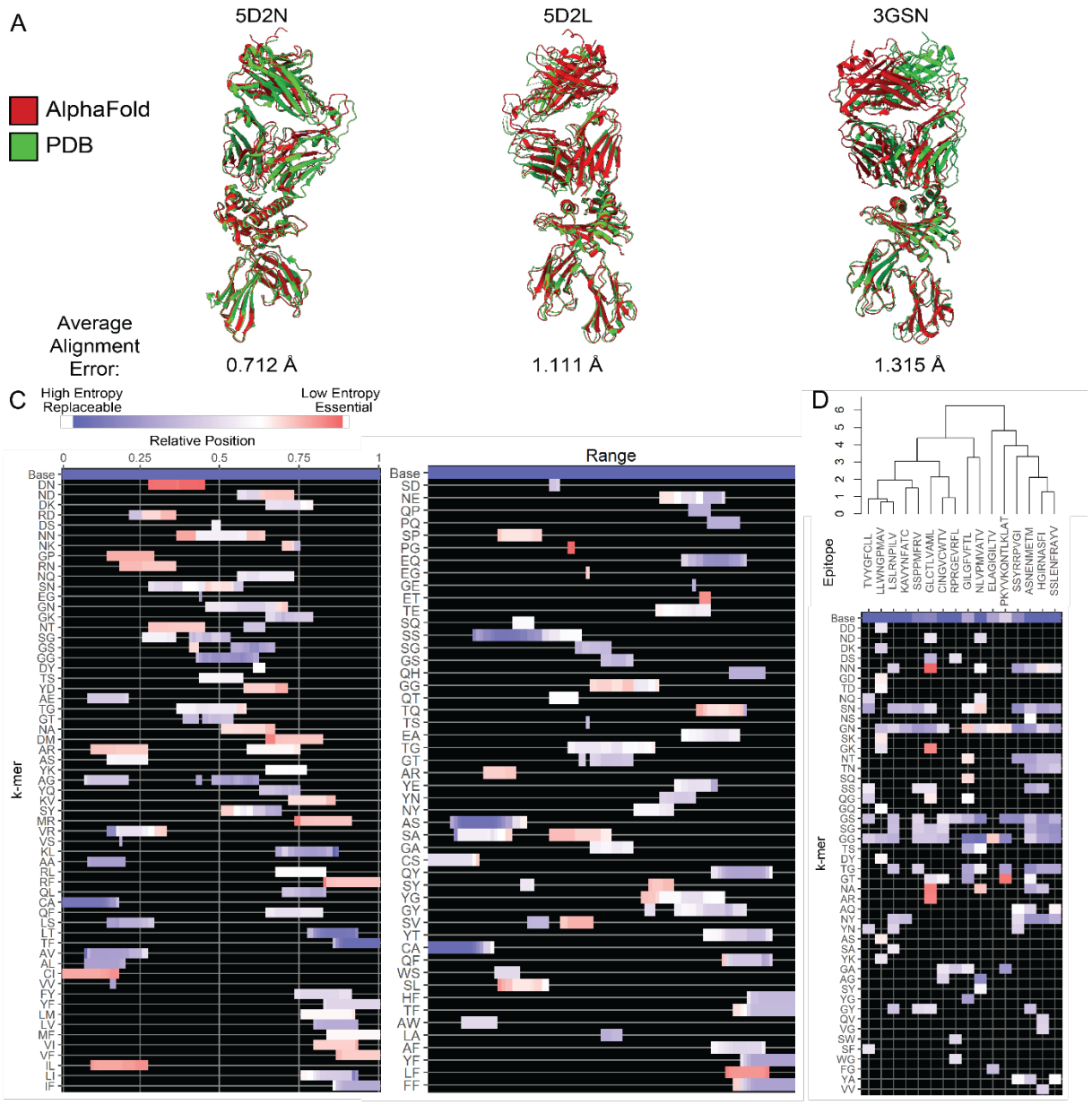


Supplementary Figure 1. Variations of TCR landscape plots for CMV pp65-specific TCRs, related to Figure 2. A. In step increments of 1% relative position, k=1 or single residue usage is shown. B. k=3 k-mers, sorted by molecular weight of the k-mer. C. k=2 k-mers, sorted by frequency of amino acid usage. D. k=3 sorted by hydrophobicity.



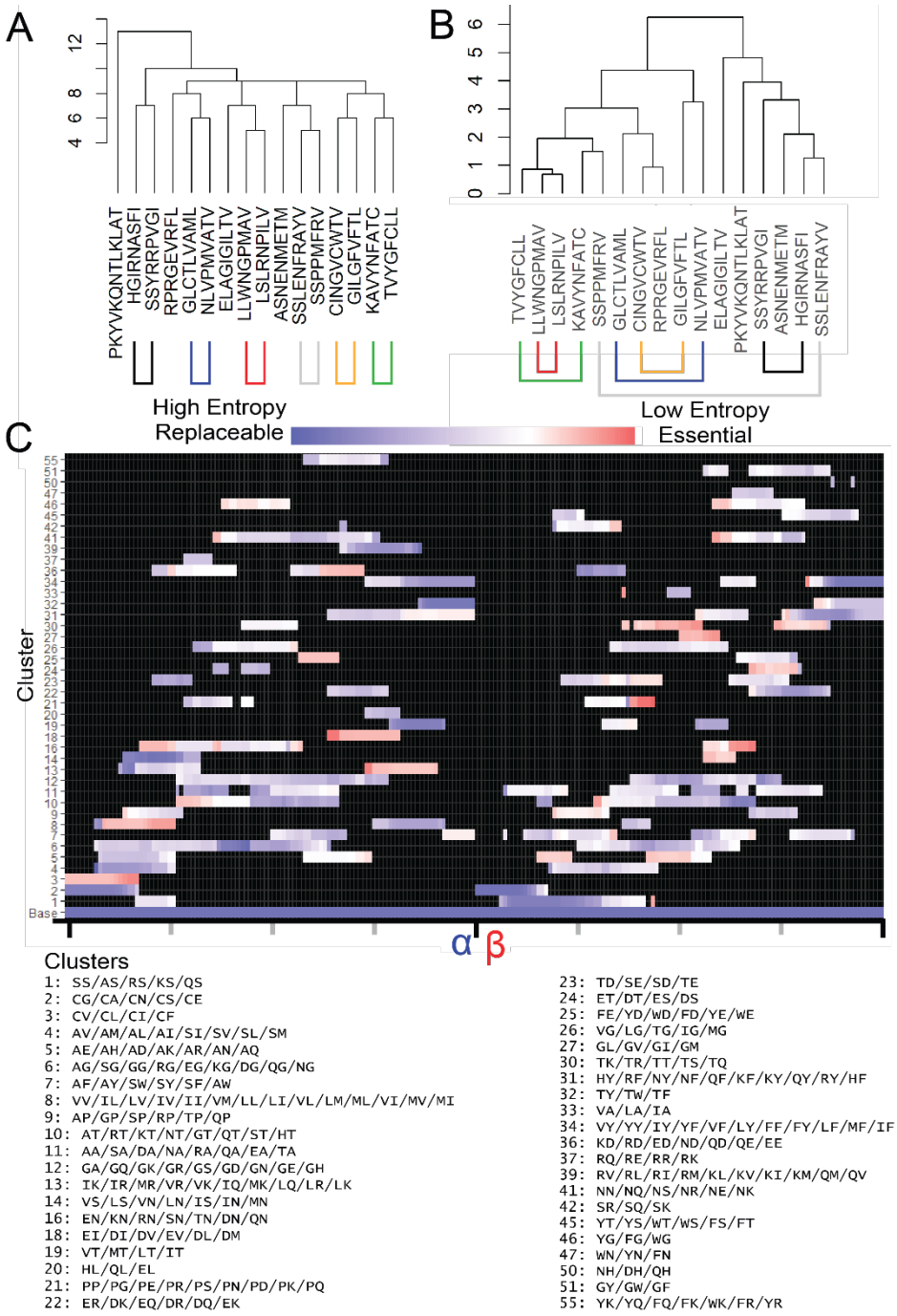
Supplementary Figure 2. Similarity metrics for TCRs, related to Figure 3. A. Schematic illustrating how differences are accrued for real TCR sequences and different length TCRs. B. Distance metrics for generic single cell TCR sequencing data set (10X Genomics) compared to CMV pp65-specific antigens in VDJdb, and distance metrics for putative CMV pp65-specific antigens compared to EBV EBNA4 antigens in VDJdb. No exact or single substitution matches were found in the 10X data set for CMV pp65, and no close matches were found compared to CMV pp65 antigens in VDJdb. C. Heatmap showing SPAN-TCR computed distances between VDJdb CMV pp65-specific TCRs. For the largest identified group of similar CDR3s, VDJ gene usage is conserved except for α J-genes. D. The TCR distance matrix is used to construct

a UMAP embedding of TCRs with the VDJdb score indicated by color. E. The CDR3 sequence similarity is shown among the similar TCRs.

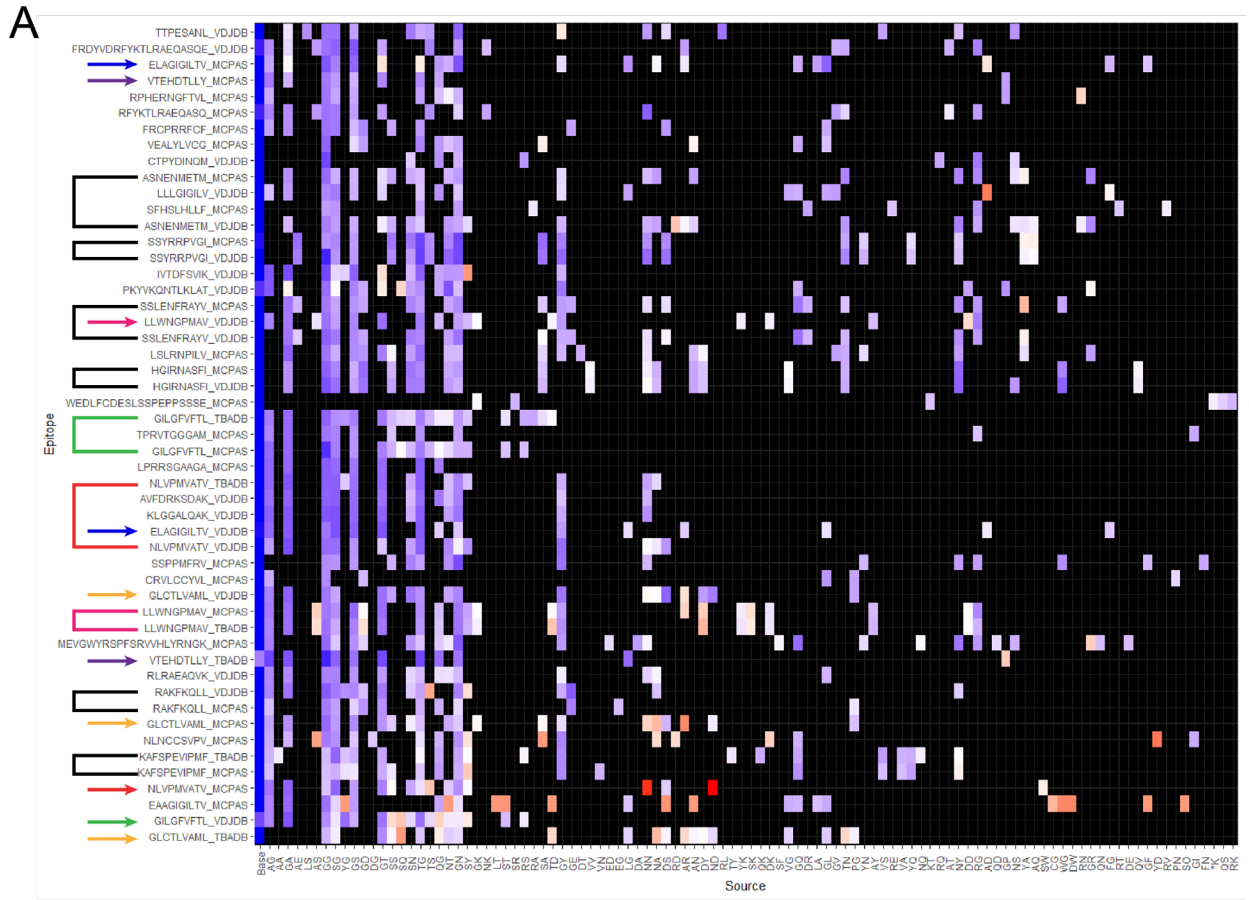


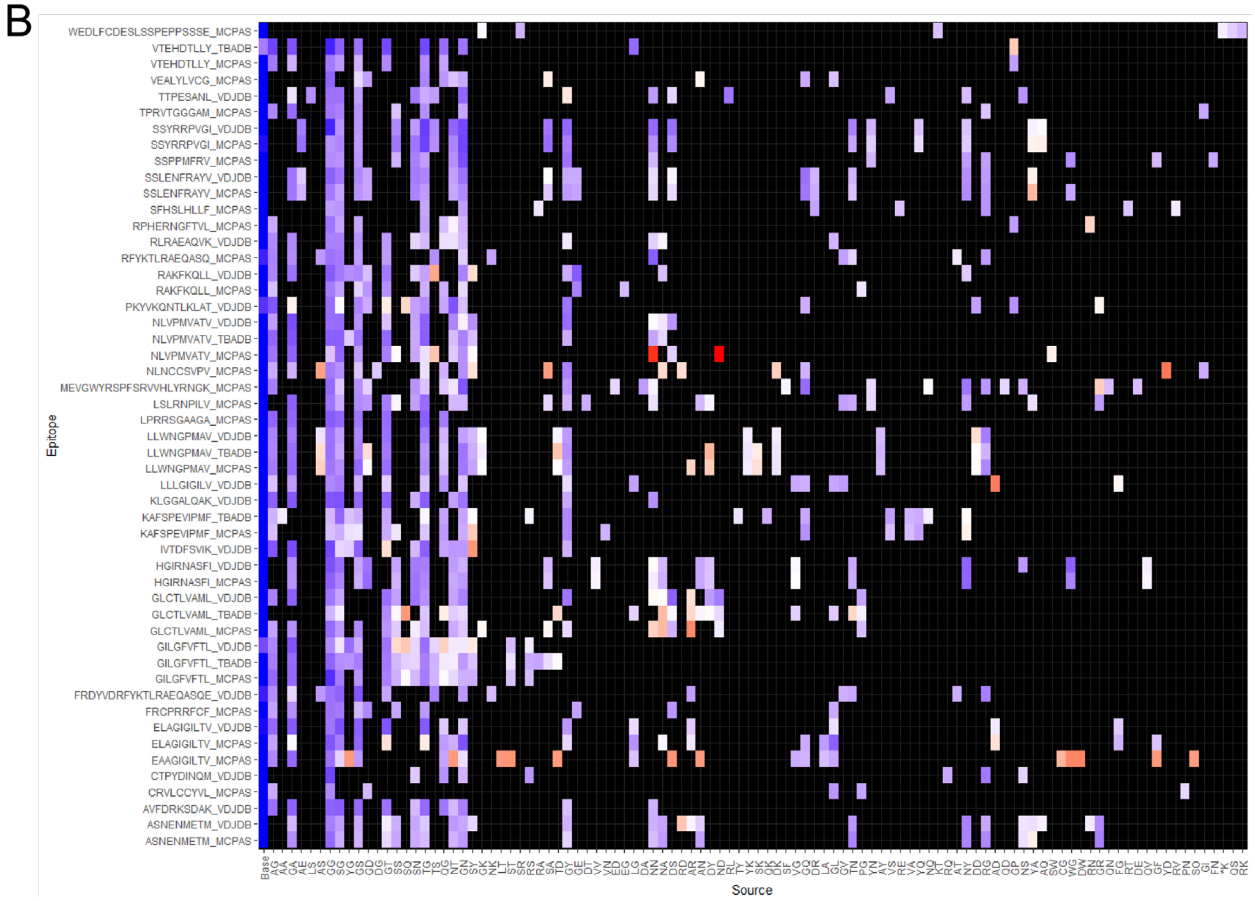
Supplemental Figure 3. Additional entropy scanning by k-mers, related to Figure 4. A. AlphaFold-generated TCR-pMHC structures reproduce crystal structures. B. Detailed table describing hydrogen bonding in AlphaFold-generated TCR-pMHC structures containing the k-mer NN at the center of the α chain CDR3 for NLVPMVATV-specific TCRs. Hydrogen bonds from the α CDR3 sequence to the TCR β, NLVPMVATV epitope, and HLA:A02 molecules are listed. C. Contributions to entropy for k-mers in CMV pp65-specific TCRs at each relative position of the α and β chains identify k-mers essential to binding. The k-mers are ordered by their hydrophobicity (y-axis), and k-mers are only plotted if they are found in at least 5% of the reported TCRs at each relative position. D. For major epitopes reported in VDJdb, essential k-mers at the center of the α chain are plotted. Hierarchical clustering of the epitopes was performed to describe the similarity of the k-mer entropy profile for each epitope. Common k-mers such

as GG or GS have little influence on entropy, while specific epitopes typically have unique essential k-mers.

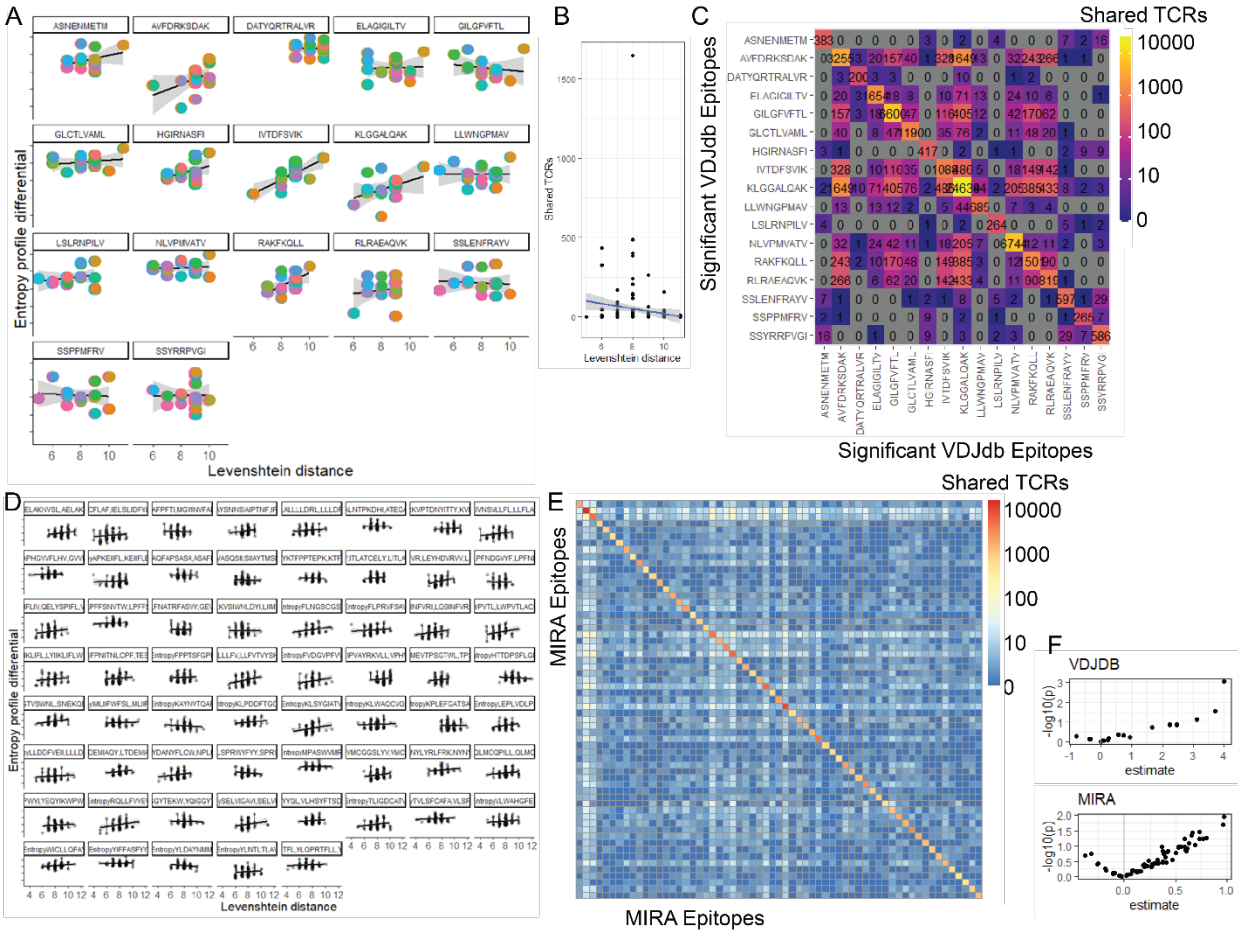


Supplementary Figure 4. Clustering similarities by entropy, related to Figure 4. A. The hierarchical clustering of epitope sequences by edit distance is plotted. For nearest neighbors, some similarities are recreated in B, the clustering of epitopes by their entropy contributions. C. Each k-mer is placed in a cluster based on the modified BLOSUM difference metric of Supplementary Table 3. A height cutoff of 6 is used to generate clusters.

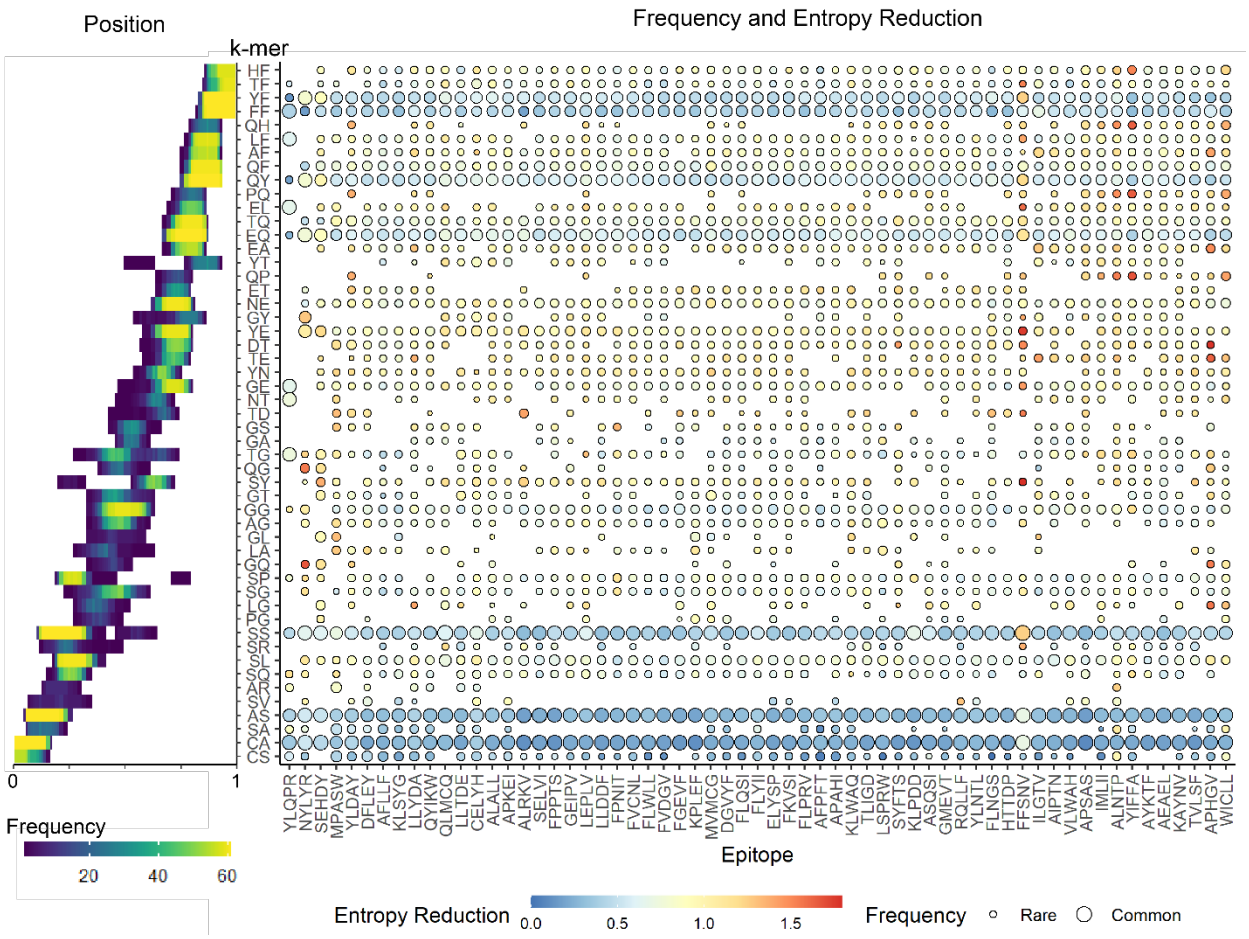

A
TTPESANL_VDJDB
FRDYVDRFYKTLRAEQASQE_VDJDB
ELAGIGILTV_MCPAS
VTEHDTLLY_MCPAS
RPHERNGFTVL_MCPAS
RFYKTLRAEQASQ_MCPAS
FRCPRRFCF_MCPAS
VEALYLVCG_MCPAS
CTPYDINQM_VDJDB
ASNENMETM_MCPAS
LLLGIGILV_VDJDB
SFHSLHLLF_MCPAS
ASNENMETM_VDJDB
SSYRRPVGI_MCPAS
SSYRRPVGI_VDJDB
IVTDFSVIK_VDJDB
PKYVKQNTLKLAT_VDJDB
SSLENFRAYV_MCPAS
LLWNGPMAV_VDJDB
SSLENFRAYV_VDJDB
LSLRNPILV_MCPAS
HGIRNASFI_MCPAS
HGIRNASFI_VDJDB
WEDLFCDESLSSPEPPSSSE_MCPAS
GILGFVFTL_TBADB
TPRVTGGGAM_MCPAS
GILGFVFTL_MCPAS
LPRRSGAAGA_MCPAS
NLVPMVATV_TBADB
AVFDRKSDAK_VDJDB
KLGGALQAK_VDJDB
ELAGIGILTV_VDJDB
NLVPMVATV_VDJDB
SSPPMFRV_MCPAS
CRVLCCYVL_MCPAS
GLCTLVAML_VDJDB
LLWNGPMAV_MCPAS
LLWNGPMAV_TBADB
MEVGWYRSPFSRVVHLYRNGK_MCPAS
VTEHDTLLY_TBADB
RLRAEAQVK_VDJDB
RAKFKQLL_VDJDB
RAKFKQLL_MCPAS
GLCTLVAML_MCPAS
NLNCCSVPV_MCPAS
KAFSPEVIPMF_TBADB
KAFSPEVIPMF_MCPAS
NLVPMVATV_MCPAS
EAAGIGILTV_MCPAS
GILGFVFTL_VDJDB
GLCTLVAML_TBADB
Epitope
Base
Source


B
WEDLFCDESLSSPEPPSSSE_MCPAS
VTEHDTLLY_TBADB
VTEHDTLLY_MCPAS
VEALYLVCG_MCPAS
TTPESANL_VDJDB
TPRVTGGGAM_MCPAS
SSYRRPVGI_VDJDB
SSYRRPVGI_MCPAS
SSPPMFRV_MCPAS
SSLENFRAYV_VDJDB
SSLENFRAYV_MCPAS
SFHSLHLLF_MCPAS
RPHERNGFTVL_MCPAS
RLRAEAQVK_VDJDB
RFYKTLRAEQASQ_MCPAS
RAKFKQLL_VDJDB
RAKFKQLL_MCPAS
PKYVKQNTLKLAT_VDJDB
NLVPMVATV_VDJDB
NLVPMVATV_TBADB
NLVPMVATV_MCPAS
NLNCCSVPV_MCPAS
MEVGWYRSPFSRVVHLYRNGK_MCPAS
LSLRNPILV_MCPAS
LPRRSGAAGA_MCPAS
LLWNGPMAV_VDJDB
LLWNGPMAV_TBADB
LLWNGPMAV_MCPAS
LLLGIGILV_VDJDB
KLGGALQAK_VDJDB
KAFSPEVIPMF_TBADB
KAFSPEVIPMF_MCPAS
IVTDFSVIK_VDJDB
HGIRNASFI_VDJDB
HGIRNASFI_MCPAS
GLCTLVAML_VDJDB
GLCTLVAML_TBADB
GLCTLVAML_MCPAS
GILGFVFTL_VDJDB
GILGFVFTL_TBADB
GILGFVFTL_MCPAS
FRDYVDRFYKTLRAEQASQE_VDJDB
FRCPRRFCF_MCPAS
ELAGIGILTV_VDJDB
ELAGIGILTV_MCPAS
EAAGIGILTV_MCPAS
CTPYDINQM_VDJDB
CRVLCCYVL_MCPAS
AVFDRKSDAK_VDJDB
ASNENMETM_VDJDB
ASNENMETM_MCPAS
Epitope
Base
Source

Supplementary Figure 5. Entropy contributions at the center of the α chain for 3 databases of unpaired TCRs, related to Figure 4. A. Entropy contributions for all 3 databases are plotted and hierarchically clustered. Epitopes with many entries (>200 for VDJdb, >100 for MCPAS or PIRB) that appear in multiple databases are noted on the left. Most shared epitopes also cluster together, with some exceptions, showing that databases are somewhat consistent. The clustering is also subject to the choice of weight function, as described in Supplementary Methods. B. Entropy contributions are plotted by alphabetical order of epitopes.



Supplementary Figure 6. Epitope similarity vs Entropy Reduction Profiles, related to Figure 6. A. For each epitope in VDJDB (>200 paired TCRs), a correlation of Levenshtein distance to other epitopes and entropy profile similarity between epitope-specific TCRs was established. B. There was a statistically insignificant negative correlation between Levenshtein epitope distance and the number of shared TCRs between epitope groups. C. Significant TCR sharing was observed in VDJdb. An average of 44.85 TCRs were shared and 18/136 pairs of epitopes had >50 shared TCRs. TCR sharing did not contribute to the positive correlation between epitope Levenshtein distance and entropy profile similarity. D. The correlation between epitope Levenshtein distance and entropy profile similarity was established for epitopes in MIRA. E. There was minimal TCR sharing in MIRA. An average of 3.12 TCRs were shared and 8/1830 pairs of epitopes had >50 shared TCRs. F. In both VDJdb and MIRA, considering entire epitope groups, the correlation of epitope Levenshtein distance to entropy reduction profile similarity was statistically greater than zero (VDJdb: N=17, p=0.007, MIRA: N=61, p=5.6e-10).



Supplementary Figure 7. Meta-analysis of common 2-mers, their significant positions, and entropy reductions, related to Figure 6. The 2-mer order (top-to-bottom) is determined by their average position in the CDR3 sequence (left, color = relative frequency of appearance). For each 2-mer and epitope pair, the frequency of observation of the 2-mer (dot size) and average reduction in entropy (color) are plotted. The most common k-mers are found at the beginning and end of TCRs, especially CA/AS/SS and QF/FF/QY/YF. These k-mers have minimal effects on entropy reduction, with the exception of epitope group FFSNV (Sup. Table). Throughout the epitope groups, epitope/essential 2-mer pairs can be identified with large entropy reductions (SEHDY/YE), (NYLDR/GY/QG), etc. Non-essential 2-mers are identified by minimal entropy reduction across epitopes, typically GG/AG/GE/TG, with notable epitope exceptions (YIFFA/GG, MPASW/AG, FFSNV/GE).

Supplementary Table 2. Common k-mers (k=4) for CMV pp65 antigen-specific TCRs in VDJdb and their GLIPH2 group membership, related to Figure 2.

	kmer	matches	Location	GLIPH Matches
1	CASN		0.005	0
2	CASR		0.025	0
3	CSAR		0.005	0
4	CASS		0.005	0
5	CASG	SL%ASGSSYNE	0.005	1
6	CSVE		0.005	0
7	CATS		0.005	0
8	CAST		0.005	0
9	CASM		0.005	0
10	CAWS		0.005	0
11	CAIS		0.005	0
12	ASSP		0.185	0
13	ASSE		0.185	0
14	ASSQ		0.205	0
15	ASSS		0.245	0
16	ASSH		0.175	0
17	ASSY	SG%SSYE	0.195	1
18	ASSL		0.185	0
19	ASSI		0.185	0
20	ASSF		0.205	0
21	SARD		0.105	0
22	ASRG		0.125	0
23	ASSV		0.185	0
24	ASMG		0.075	0
25	ASGL		0.075	0
26	SARG		0.125	0
27	SSHW		0.195	0
28	SSPQ		0.185	0
29	SSPV		0.255	0
30	SSLD		0.185	0
31	SSYQ		0.265	0
32	SSYS	SS%SYNE	0.265	1
33	SSLA		0.255	0
34	SSLV		0.255	0
35	SSQT	S%QTGASYG	0.285	1
36	SSFQ		0.305	0
37	SSSA		0.335	0
38	SHWD	S%WDRSSGANV//S%WDRGSGANV	0.195	2
39	SSSV		0.365	0
40	SPVT	SPV%GAPHE//SP%TGTGHYG//SPVTGQG%YG	0.315	3
41	SLAP	SLAPG%TNEK//SLAPGATN%K//%LAPGATNEK//SLAPGAT%EK//SLAPGA%NEK//SL%PGATNEK//SLAP%ATNEK//SLA%GTDSYE//SL%PGE	0.315	9
42	SLVT	SLV%GPGTDT//SL%TGAGAYG//SL%TDT	0.225	3
43	SYQT	SYQT//SYQTGA%YG//SYQTG%IYG//SYQTGTG%YG//S%QTGASYG	0.335	5
44	SYST		0.375	0
45	SQTQ	S%TQLWET//SQTQLWE%//SQTQ%WET//SQ%QLWET//SQTQLW%T//%QTQLWET//SQTQ	0.355	7
46	SSAN	SSA%YG//SS%NE//%SANYG//SS%NG	0.415	4
47	PVTG	SPV%GAPHE//SP%TGTGHYG//SPVTGQG%YG	0.375	3
48	LAPG	SLAPG%TNEK//SLAPGATN%K//%LAPGATNEK//SLAPGAT%EK//SLAPGA%NEK//SL%PGATNEK//SLAP%ATNEK//SLA%GTDSYE//SL%PGE	0.375	9
49	SSAY	SSA%YG//SSAYY%	0.415	2
50	SSVN	SSV%E//SSVN%//S%VNE//SS%NE//SS%NG//SSV%G	0.455	6
51	YQTG	YQTG//SFY%TGAPYNE//SYQTGA%YG//SYQTG%IYG//SYQTGTG%YG//S%QTGASYG	0.395	6
52	QTQL	QTQL//S%TQLWET//SQTQLWE%//SQTQ%WET//SQ%QLWET//SQTQLW%T//%QTQLWET	0.425	7
53	APGT	SLAPG%TNEK//SLA%GTDSYE	0.435	2
54	VTGG		0.435	0
55	VTGT	SP%TGTGHYG//NQVT%TGAYG//NQVTGTGAY%//NQVTGT%AYG	0.435	4
56	QTGT	SYQTG%IYG//SYQTGTG%YG//SP%TGTGHYG//SPQTG%IYG	0.445	4
57	STGT	SP%TGTGHYG//S%GTSYNE//S%GTGYE	0.405	3
58	QTGA	SFY%TGAPYNE//SYQTGA%YG//SYQTG%IYG//NQVT%TGAYG//SPQTG%IYG//S%QTGASYG//SL%TGAGAYG	0.465	7
59	SANY	SSA%YG//%SANYG	0.495	2
60	SAYY	SSA%YG//SSAYY%//SAYY	0.495	3
61	SVNE	SSV%E//SSVN%//S%VNE//SS%NE//S%NE	0.545	5
62	TQLW	TQLW//S%TQLWET//SQTQLWE%//SQTQ%WET//SQ%QLWET//SQTQLW%T//%QTQLWET//SL%QLWET	0.495	8
63	YSTG	SFY%TGAPYNE	0.375	1
64	PGTT	SLAPG%TNEK	0.495	1
65	TGTG	SYQTGTG%YG//SP%TGTGHYG//NQVT%TGAYG//TTG%GGYE//SALGG%GTGE//NQVTGTGAY%//SPITGTG%YG//NQVTGT%AYG//SAT%TGE//S%GTGYE	0.495	10
66	TGGI	SYQTG%IYG//SPTGG%ADT//SPQTG%IYG//SIGGTGG%P	0.505	4
67	TGAS	SYQTGA%YG//S%QTGASYG	0.535	2
68	TGAA	SYQTGA%YG	0.535	1
69	ANYG	SSA%YG//SYQTGA%YG//%SANYG	0.585	3
70	AYYG	SSA%YG//SSAYY%//SYQTGA%YG	0.585	3
71	QLWE	QLWE//S%TQLWET//SQTQLWE%//SQTQ%WET//SQ%QLWET//SQTQLW%T//%QTQLWET//SL%QLWET	0.575	8
72	GTTN	SLAPG%TNEK	0.565	1
73	GTGG	SYQTGTG%YG//SYG%GGE//S%TLGTGGVE//TTG%GGYE//SPITGTG%YG//SIGGTGG%P	0.565	6
74	GGIY	SYQTG%IYG//SPQTG%IYG	0.565	2
75	VNEQ		0.635	0
76	GASY	SYQTGA%YG//S%QTGASYG//SGG%SYNE	0.605	3
77	GAAY	SYQTGA%YG	0.605	1
78	TTNE	SLAPG%TNEK	0.625	1
79	NYGY		0.685	0
80	GIYG	SYQTG%IYG//SYQTGTG%YG//SPQTG%IYG//SPVTGQG%YG//SPITGTG%YG//SQVG%YG	0.625	6
81	LWET	S%TQLWET//SQTQLWE%//SQTQ%WET//SQ%QLWET//SQTQLW%T//%QTQLWET//LWET//SL%QLWET	0.645	8
82	YYGY		0.665	0
83	ASYG	SSA%YG//SYQTGA%YG//S%QTGASYG	0.665	3
84	AAYG	SSA%YG//SYQTGA%YG//NQVTGT%AYG	0.665	3
85	NEQF		0.805	0
86	TNEK	SLAPG%TNEK//SLAPGATN%K//%LAPGATNEK//SLAPGAT%EK//SLAPGA%NEK//SL%PGATNEK//SLAP%ATNEK	0.685	7
87	IYGY		0.685	0
88	WETQ		0.715	0
89	YGYT		0.805	0

90	SGNT	SF%GNTE//G%GNTE	0.705	2
91	SYNE	SS%NE//SL%ASGSSYNE//S%EASGSSYNE//S%GTSYNE//SGG%SYNE//S%NE//SS%SYNE//S%YNE	0.665	8
92	GNYG	SYQTGTG%YG//SPVTGQG%YG//SPITGTG%YG//SQVG%YG	0.685	4
93	YNEQ		0.745	0
94	SYGY	SYG%GGE//SS%GYE	0.735	2
95	AYGY		0.735	0
96	SYEQ		0.715	0
97	YEQY		0.785	0
98	TEAF		0.785	0
99	NEKL		0.745	0
100	TGEL		0.735	0
101	ETQY		0.785	0
102	EQFF		0.975	0
103	TDTQ		0.735	0
104	GNTI		0.765	0
105	DTQY		0.815	0
106	GELF		0.805	0
107	GYTF		0.975	0
108	EAFF		0.975	0
109	GYTG	NPSGGY%GE	0.815	1
110	EQYF		0.975	0
111	EKLF		0.815	0
112	NTIY		0.825	0
113	TQYF		0.975	0
114	ELFF		0.975	0
115	QPQH		0.805	0
116	PQHF		0.975	0
117	YTGF	S%TGFGNQP	0.995	1
118	KLFF		0.985	0
119	TIYF		0.985	0
120	LEQY		0.805	0
121	NTEA		0.795	0
122	GVEQ		0.795	0
123	VEQY		0.815	0
124	GKYG	SYQTGTG%YG//SPVTGQG%YG//SPITGTG%YG//SQVG%YG	0.805	4
125	GGVE	SSGG%E//S%TLGTGGVE//SFGG%E//SLGG%E	0.805	4
126	HLEQ		0.805	0
127	KYGY		0.805	0
128	YTEA		0.835	0
129	DEQY		0.875	0
130	GGYT	NPSGGY%GE//SLGG%TE	0.875	2
131	NVLT		0.875	0
132	PLHF		0.975	0
133	VLTF		0.975	0
134	SPLH		0.885	0
135	IQYF		0.975	0

Supplementary Table 3. BLOSUM62 derived distance matrix for amino acid similarity, related to Figure 3.

Amino Acid	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y
A	0	4	6	5	6	4	6	5	5	5	5	6	5	5	5	3	4	4	7	5	6
C	4	0	7	8	6	7	7	5	7	5	5	7	7	7	7	5	5	5	6	5	6
D	6	7	0	2	7	5	5	7	5	8	7	3	5	4	6	4	5	7	8	5	7
E	5	8	2	0	7	6	4	7	3	7	6	4	5	2	4	4	5	6	7	5	6
F	6	6	7	7	0	7	5	4	7	4	4	7	8	7	7	6	6	5	3	5	1
G	4	7	5	6	7	0	6	8	6	8	7	4	6	6	6	4	6	7	6	5	7
H	6	7	5	4	5	6	0	7	5	7	6	3	6	4	4	5	6	7	6	5	2
I	5	5	7	7	4	8	7	0	7	2	3	7	7	7	7	6	5	1	7	5	5
K	5	7	5	3	7	6	5	7	0	6	5	4	5	3	2	4	5	6	7	5	6
L	5	5	8	7	4	8	7	2	6	0	2	7	7	6	6	6	5	3	6	5	5
M	5	5	7	6	4	7	6	3	5	2	0	6	6	4	5	5	5	3	5	5	5
N	6	7	3	4	7	4	3	7	4	7	6	0	6	4	4	3	4	7	8	5	6
P	5	7	5	5	8	6	6	7	5	7	6	6	0	5	6	5	5	6	8	5	7
Q	5	7	4	2	7	6	4	7	3	6	4	4	5	0	3	4	5	6	6	5	5
R	5	7	6	4	7	6	4	7	2	6	5	4	6	3	0	5	5	7	7	5	6
S	3	5	4	4	6	4	5	6	4	6	5	3	5	4	5	0	3	6	7	5	6
T	4	5	5	5	6	6	6	5	5	5	5	4	5	5	5	3	0	4	6	5	6
V	4	5	7	6	5	7	7	1	6	3	3	7	6	6	7	6	4	0	7	5	5
W	7	6	8	7	3	6	6	7	7	6	5	8	8	6	7	7	6	7	0	5	2
X	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Y	6	6	7	6	1	7	2	5	6	5	5	6	7	5	6	6	6	5	2	5	0