Supplementary Information for "A Computationally Optimised Structural Integrity Sequence Enhances Vaccine Stability, Yield and Safety Profile"

Arthur Sarron, Sun B. Sowers, Yaroslav Tsybovsky, Stephen N. Crooke, Guillaume B. E. Stewart-Jones, Ian M. Overton

Table of contents

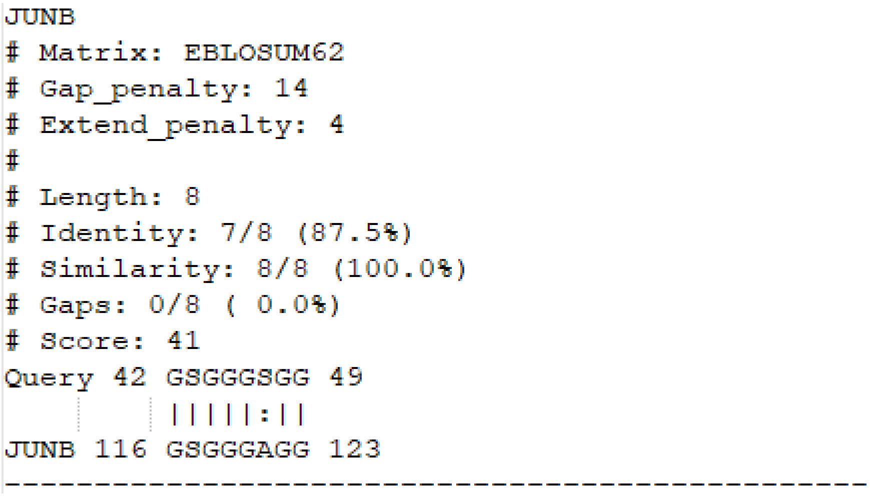
Supplementary Table S1 (SupplementaryTable_S1.xls) as well as the homology models for GCN4_QM (SupplementaryData_S1.pdb), GCN4_init (SupplementaryData_S2.pdb) are available a separate supplementary files.

The Kmer_finder software is available from Github at https://github.com/overton-group/KmerFinder

Please note, reference numbering in the Figure Legends below refers to the main manuscript.

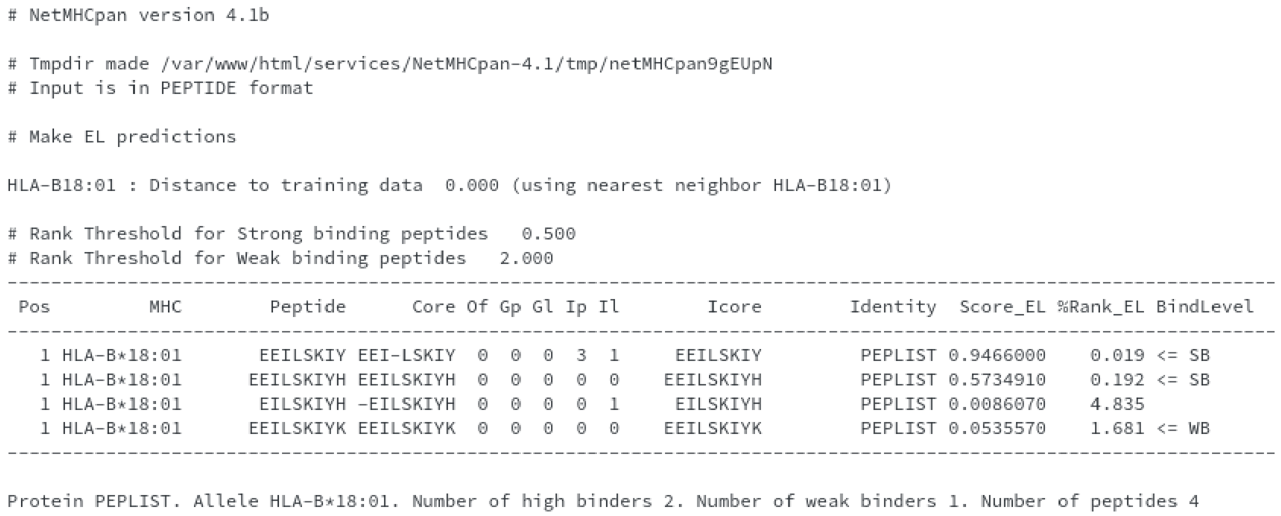
GCN4_init: IEDKIEEILSKIYHIENEIARIKKLIGEAP

Supplementary Figure S1. GCN4_init trimerisation region. The sequence above was taken from the PDB record 2B9B, where it was introduced in order to stabilise the parainfluenza virus Fusion (F) protein trimer in the prefusion conformation [21].



```
JUNB
# Matrix: EBLOSUM62
# Gap_penalty: 14
# Extend_penalty: 4
#
# Length: 8
# Identity: 7/8 (87.5%)
# Similarity: 8/8 (100.0%)
# Gaps: 0/8 ( 0.0%)
# Score: 41
Query 42 GSGGGSGG 49
         |||||:||
JUNB 116 GSGGGAGG 123
-----------------------------------------------
```

Supplementary Figure S2. Example PSIBLAST result, aligning to JUNB. The alignment length, score, percentage identity, percentage similarity, query (GCN4_init) and subject (JUNB) sequences are shown.

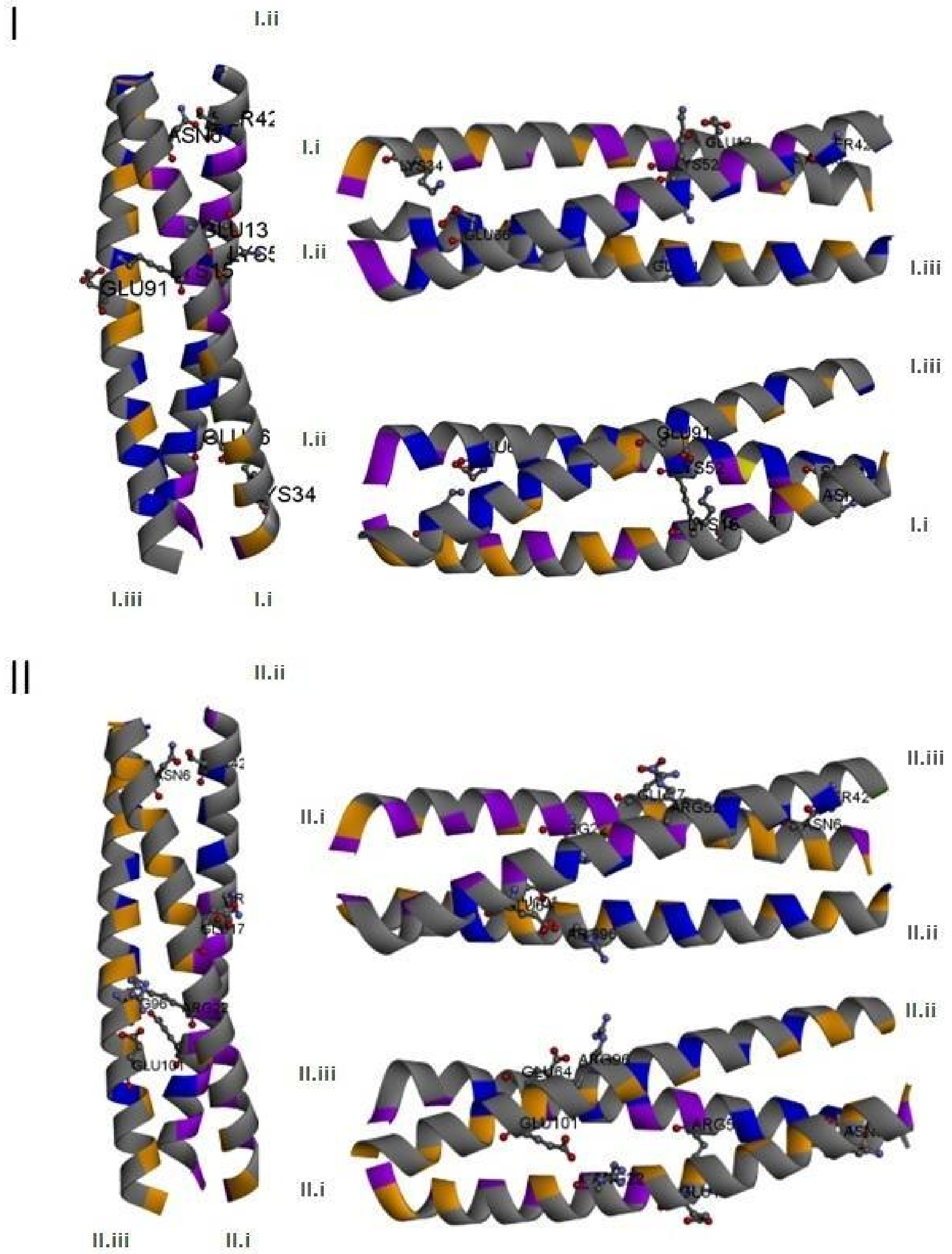


```
# NetMHCpan version 4.1b

# Tmpdir made /var/www/html/services/NetMHCpan-4.1/tmp/netMHCpan9gEUpN
# Input is in PEPTIDE format

# Make EL predictions

HLA-B18:01 : Distance to training data  0.000 (using nearest neighbor HLA-B18:01)

# Rank Threshold for Strong binding peptides   0.500
# Rank Threshold for Weak binding peptides   2.000
-------------------------------------------------------------------------------------------------------------------
 Pos         MHC        Peptide      Core Of Gp Gl Ip Il        Icore        Identity  Score_EL %Rank_EL BindLevel
-------------------------------------------------------------------------------------------------------------------
   1 HLA-B*18:01       EEILSKIY EEI-LSKIY  0  0  0  3  1     EEILSKIY         PEPLIST 0.9466000    0.019 <= SB
   1 HLA-B*18:01      EEILSKIYH EEILSKIYH  0  0  0  0  0    EEILSKIYH         PEPLIST 0.5734910    0.192 <= SB
   1 HLA-B*18:01       EILSKIYH -EILSKIYH  0  0  0  0  1     EILSKIYH         PEPLIST 0.0086070    4.835
   1 HLA-B*18:01      EEILSKIYK EEILSKIYK  0  0  0  0  0    EEILSKIYK         PEPLIST 0.0535570    1.681 <= WB
-------------------------------------------------------------------------------------------------------------------

Protein PEPLIST. Allele HLA-B*18:01. Number of high binders 2. Number of weak binders 1. Number of peptides 4
```

Supplementary Figure S3. GCN4_init peptides that overlap with ANK3 and have predicted binding to HLA-B18:01. Results are shown for the MHC allelle HLA-B*18:01. The ‘BindLevel’ column shows NetMHCpan4.1 predicted binding strength as follows: ‘SB’ (strong binder), ‘WB’ (weak binder). The %Rank_EL defines predicted strong binders (<0.5%) and weak binders (<2%). The SB sequence EEILSKIY would involve an amino acid insertion at position 3, the SB sequence EEILSKIYH does not require an insertion but involves two amino acid changes in ANK3 relative to the reference genome. The WB sequence EEILSKIYK could be produced by a single nucleotide change relative to the reference ANK3 sequence.

Allele	Population	% of individuals that have the allele	Allele Frequency (in_decimals)	Sample Size
B*18:01	Austria	9.0	0.0500	200
B*18:01	Azores Terceira Island		0.0390	130
B*18:01	Belgium	10.2	0.0510	99
B*18:01	Croatia		0.1130	150
B*18:01	Croatia pop 4		0.0816	4,000
B*18:01	Czech Republic		0.0750	106
B*18:01	Czech Republic NMDR		0.0643	5,099
B*18:01	England North West	9.1	0.0450	298
B*18:01	Finland		0.0780	91
B*18:01	France French Bone Marrow Donor Registry		0.0580	42,623
B*18:01	France Southeast	8.5	0.0430	130
B*18:01	Germany DKMS - Austria minority		0.0459	1,698
B*18:01	Germany DKMS - Bosnia and Herzegovina minority		0.0749	1,028
B*18:01	Germany DKMS - China minority		0.0047	1,282
B*18:01	Germany DKMS - Croatia minority		0.0848	2,057
B*18:01	Germany DKMS - France minority		0.0551	1,406
B*18:01	Germany DKMS - German donors		0.0477	3,456,066
B*18:01	Germany DKMS - Greece minority		0.0940	1,894
B*18:01	Germany DKMS - Italy minority		0.1066	1,159
B*18:01	Germany DKMS - Netherlands minority		0.0415	1,374
B*18:01	Germany DKMS - Portugal minority		0.0492	1,176
B*18:01	Germany DKMS - Romania minority		0.0875	1,234
B*18:01	Germany DKMS - Spain minority		0.0678	1,107
B*18:01	Germany DKMS - Turkey minority		0.0587	4,856
B*18:01	Germany DKMS - United Kingdom minority		0.0358	1,043
B*18:01	Germany pop 6		0.0473	8,862
B*18:01	Germany pop 8		0.0505	39,689
B*18:01	Greece pop 6		0.1580	242
B*18:01	Ireland Northern	5.7	0.0290	1,000
B*18:01	Ireland South	8.4	0.0420	250
B*18:01	Italy North pop 3	33.3	0.1670	97
B*18:01	Italy pop 5		0.0580	975
B*18:01	Kosovo	22.6	0.1331	124
B*18:01	Netherlands Leiden		0.0330	1,305
B*18:01	Poland		0.0700	200
B*18:01	Poland DKMS		0.0727	20,653
B*18:01	Portugal Azores Terceira Island	7.9	0.0395	130
B*18:01	Romania	21.3	0.1130	348
B*18:01	Serbia pop 2	15.7	0.0780	102
B*18:01	Spain (Catalunya, Navarra, Extremadura, Aaragón, Cantabria,	15.9	0.0835	4,335

Supplementary Figure S4. Population frequencies of HLA-B*18:01 in European populations. The frequency of HLA-B*18:01 is shown, covering up to 33.3% of individuals in some populations. Allele frequency values are expected to be approximately half of the proportion of individuals with the allele due to heterozygosity. Data is from the Allele Frequency Net Database (AFND) [48].



Supplementary Figure S5. Structural models for GCN4_init and GCN4_QM, highlighting inter-helix interactions. Homology models of the trimeric GCN4_init (I, top) and GCN4_QM (II, bottom) are shown. The template was the yeast GCN4 structure (PDB identifier: 1GCM). Helix colouring indicates regions of hydrophobic interactions for the pairs i-ii (purple), i-iii(orange) and ii-iii (blue); grey shows residues not involved in hydrophobic packing. Van der Waals radii are shown for residues participating in electrostatic and hydrogen bond interactions (CPK space-filling), which are also labelled in black text. Atoms are shown in standard colours (oxygen in red, nitrogen in blue, carbon in grey). Helix identifiers (I.i, I.ii etc.) are equivalent to those shown in main Figure 3. Notably, helices I.ii and I.iii (GCN4_init) only have hydrophobic packing interactions, while helices II.ii and II.iii (GCN4_QM) have an electrostatic interaction between Glu64 (II.ii) and Arg96 (II.iii).