Supplementary Information

High seroreactivity against SARS-CoV-2 Spike epitopes in a pre SARS-CoV-2 cohort: implications for antibody testing and vaccine design

Kaia Palm[1*#], Mariliis Jaago[1,2*], Annika Rähni[1,2], Nadežda Pupina[1], Arno Pihlak[1], Helle Sadam[1,2], Annela Avarlaid[2], Anu Planken[3], Margus Planken[3], Liina Haring[4], Eero Vasar[5,6], Miljana Baćević[7], France Lambert[8], Eija Kalso[9], Pirkko Pussinen[10], Pentti J. Tienari[11], Antti Vaheri[12], Dan Lindholm[13,14], Tõnis Timmusk[1,2] and Amir M Ghaemmaghami[15#]

[1]Protobios Llc, Tallinn, Estonia. Electronic address: kaia@protobios.com

[2]Department of Chemistry and Biotechnology, Tallinn University of Technology, Tallinn, Estonia;

[3]North Estonia Medical Centre Foundation, Tallinn, Estonia;

[4]Institute of Clinical Medicine, University of Tartu, Psychiatry Clinic of Tartu University Hospital, Estonia;

[5]Department of Physiology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia;

[6]Center of Excellence for Genomics and Translational Medicine, University of Tartu, Tartu, Estonia;

[7]Dental Biomaterial Research Unit (d-BRU), Faculty of Medicine, University of Liege, Liege, Belgium;

[8]Department of Periodontology and Oral Surgery, Faculty of Medicine, University of Liege, Belgium;

[9]Department of Anaesthesiology, Intensive Care and Pain Medicine, Helsinki Uiversity Hospital and Department of Pharmacology and SleepWell Research Programme, University of Helsinki, Helsinki, Finland;

[10]Oral and Maxillofacial Diseases, University of Helsinki and Helsinki University Hospital, Helsinki, Finland;

[11]Department of Neurology, Neurocenter, Helsinki University Hospital, and Translational Immunology Research Program, University of Helsinki, Helsinki, Finland;

[12]Department of Virology, Medicum, University of Helsinki, Finland;

[13]Department of Biochemistry and Developmental Biology, Faculty of Medicine, University of Helsinki, Helsinki, Finland;

[14]Minerva Foundation Institute for Medical Research, Helsinki, Finland;

[15]Immunology and Immuno-Bioengineering Group, School of Life Science, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham, United Kingdom. Electronic address: Amir.Ghaemmaghami@nottingham.ac.uk

*These authors contributed equally to the work.

Correspondence to Kaia Palm or Amir Ghaemmaghami

Key words: SARS-CoV-2, immunoprfiling, antibody, seroreactivity, COVID-19, antigenic sin

Supplementary Information

Materials and Methods

Peptide antigen clustering

Non-discriminatory clustering (PD-CAD; MI; T2D; BC):

Within the PD-CAD cohort, the 96 subjects are grouped two-ways: 1) by CAD diagnosis; or by 2) PD diagnosis, resulting in 6 partially overlapping disease groups. Most abundant and shared peptide antigens were extracted for each of 6 groups, based on criteria: peptide must be present in ≥10 repeats in one sample; and must be present in ≥10% of samples within disease group (ranging from 2-4, based on group size). Resulting peptide sets were compared with random-generated sets of matching lengths to identify core sequence motifs more enriched in a given group, with the criteria: hypergeometric p-value < 10^{-8} or query/reference ratio >10; present in ≥ 4 distinct peptides; ≥ 4 fixed amino acid positions. As a result, distinct motif sequences were selected: PD-CAD (altogether 8088 distinct; gum-healthy (1668), gingivitis (3776), periodontitis (2155), no-CAD (1885), stable-CAD (906), ACS (1888)), MI (759), HC-MI (3260), and T2D (1169).

For myocardial infarction cohort analysis of the 50 subjects (MI), the control group of 50 healthy subjects was selected (MI_HC). For further analysis, the most abundant peptide antigens were identified for MI and MI_HC groups separately. Selected peptides were sequenced >10 times in one patient and presented at least in five MI or MI_HC samples, correspondingly. For revealing recognition patterns, peptides from MI and MI_HC groups were transformed to consensus sequences (motifs) via SPEXS2 software, using the criteria: hypergeometric p-value < 10^{-5}; present in ≥ 4 distinct peptides; ≥ 4 fixed amino acid positions. As a result, 759 and 3260 distinct motifs were identified for MI and MI_HC groups, respectively.

For Type 2 diabetes (T2D) with foot ulcer condition analysis, 25 subjects were selected. Most abundant and shared peptide antigens from MVA immunoprofiles were selected with criteria: peptide must be present in ≥10 repeats in one sample; and must be present in ≥2 samples. The resulting peptide set was compared with random-generated peptide set of same length and enriched motifs were identified using SPEXS2 algorithm, using hypergeometric p-value < 10^{-7} and motif to be present in ≥ 4 distinct peptides. As a result, 1169 motifs with ≥4 fixed amino acid positions were identified.

For patients with breast cancer (BC) and/or neuropathic pain (NP) condition, top 900 most abundant peptides from each sample separately were screened for motif sequences (≥5 fixed amino acid positions). Motifs identified from ≥2 samples were extracted for further analysis and motifs more detected from NP group were selected. Additionally, peptides from a large non-

unique dataset (shared in ≥2 samples) that did not contain any motifs identified by the above-mentioned approach, were examined separately. Peptides present in ≥4 samples (> 10%) in a given subgroup were extracted and common consensus motif sequences were identified with SPEXS2 tool (with hypergeometric p-value < 10-3, ≥5 fixed amino acid positions, motif present in ≥4 unique peptides). Altogether, 1014 distinct motif sequences were selected for further analysis.

Discriminatory clustering (FEP; SZ; MS):

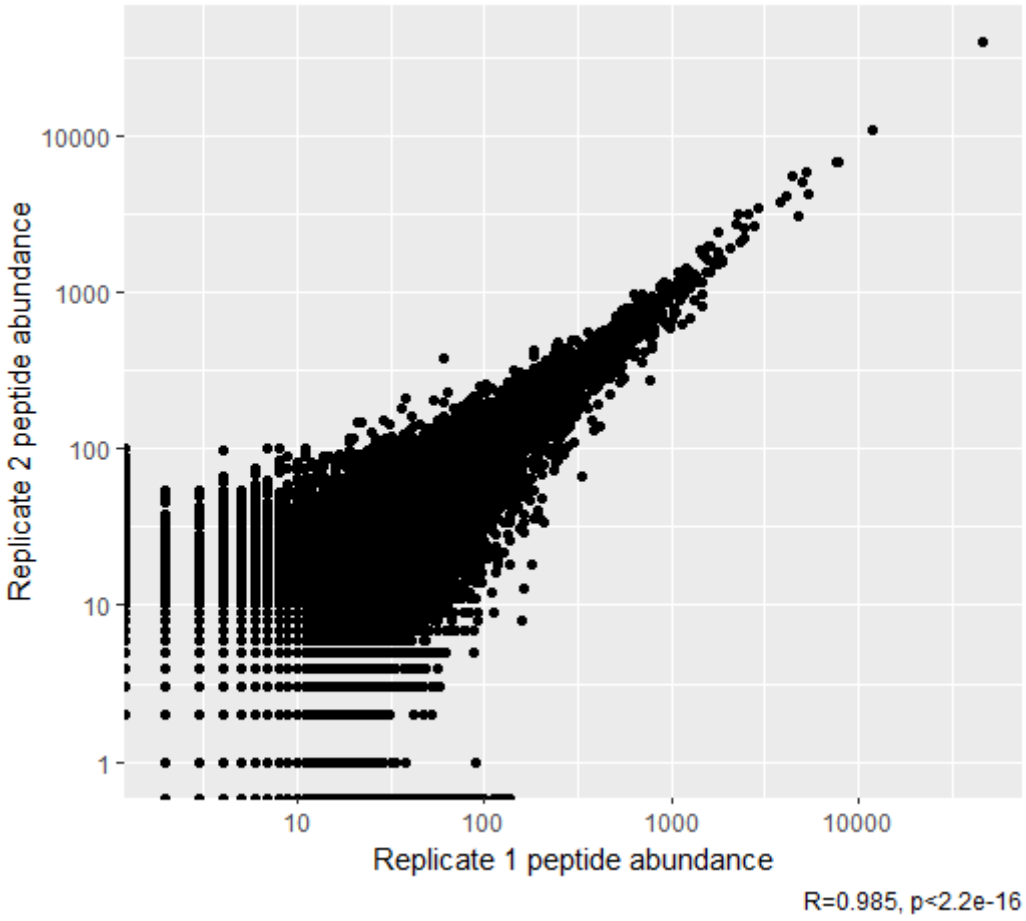
Within the FEP/SZ cohort case and control groups, most abundant and shared group-specific peptide antigens were extracted for each of the groups, based on criteria: the chosen peptide must be present in ≥10 repeats in one sample; and must be present in ≥ 10% of the samples within the given group. Core motif sequences were defined for each group independently as motifs that were more enriched when compared to 3 different reference sets: 1) random-generated reference set of same length, within-peptide scrambling; 2) within amino acid position- and across-peptide scrambling; and 3) the top abundant peptide set from the control group. Distinct motif sequences identified from SPEXS analyses were selected, based on criteria: hypergeometric p-value < 10^{-6} (FEP), < 10^{-5} (FEP_Ctrl), < 10^{-8} (SZ), or < 10^{-6} (SZ_Ctrl); present in ≥ 4 distinct peptides; ≥ 4 fixed amino acid positions. Motifs matching these criteria were selected for FEP (228), SZ (1785), FEP_HC (1935), SZ_HC (760).

Within the MS group, subgroups with or without initial optic neuritis diagnosis were analysed separately. The topmost abundant peptides from both subgroups (n=10) were extracted with criteria: peptide count ≥ 5 in ≥ 1 sample. Using SPEXS2 the peptide sets were compared to age- and sex-matched control top peptide set using hypergeometric p-value < 10^{-7}; motif to be present in ≥ 4 distinct peptides and have ≥4 fixed amino acid positions. As a result, 3500 distinct motif sequences were identified for MS group.

Supplementary Information

Supplementary Figures

Extended Data Figure 1. Pearson correlation of R=0.985 between MVA peptide antigen profiles of two replicates confirms reproducibility of MVA experiment. R programming language and package "ggpubr" was used to calculate correlation and perform hypothesis test on the correlation coefficient ($p < 2.2e-16$), and package "ggplot2" was used to visualize the results (Kassambara 2016; Wickham 2016).



Supplementary Information

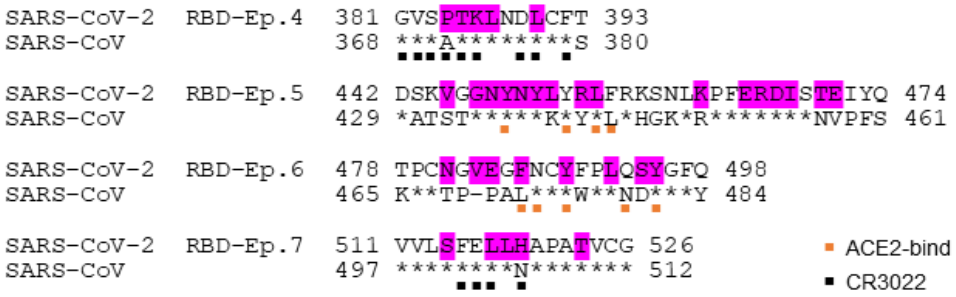
Extended Data Table 1. Predicted epitopes on SARS-CoV-2 S glycoprotein (P0DTC2).

Pattern represents one of motif patterns with exact alignment on the predicted epitope.

Epitope number	Start position	End position	Sequence	Pattern
1	26	34	PAYTNSFTR	NSF.R
2	47	58	VLHSTQDLFLPF	V..S..D…P
3	170	185	YVSQPFLMDLEGKQGN	L..K.GN
4	384	390	PTKLNDL	PTKL..L
5	445	471	VGGNYNYLYRLFRKSNLKPFERDISTE	K….DI.T
6	481	495	NGVEGFNCYFPLQSY	N.VE.F
7	514	523	SFELLHAPAT	S…LH…T
8	570	582	ADTTDAVRDPQTL	PQTL
9	599	612	TPGTNTSNQVAVLY	GTN.S
10	650	660	LIGAEHVNNSY	L..A…..SY
11	757	768	GSFCTQLNRALT	T.LNR
12	804	815	QILPDPSKPSKR	I.P…KP
13	858	869	LTVLPPLLTDEM	V.P.L…E
14	937	944	SLSSTASA	SL.S…A
15	1151	1161	ELDKYFKNHTS	EL….K...S

Supplementary Information

Extended Data Figure 2. Sequence alignment of SARS-CoV-2 (P0DTC2) with SARS-CoV S glycoprotein (P59594) reveals amino acid mismatches in MVA-predicted epitopes. The conservation by amino acid sequence between SARS-CoV and SARS-CoV-2 in our selected Ep.5 (spanning 27 amino acids) is ~ 59.3 %. For Ep.6, out of 5 amino acids important for ACE2 binding only 3 (60%) are conserved between SARS-CoV and SARS-CoV-2. Amino acids on SARS-CoV-2 S RBD with exact alignments with MVA motifs are coloured in magenta. Amino acids on SARS-CoV S with asterisks (*) are identical with SARS-CoV-2 S, hyphen (-) denotes absent amino acid. Orange squares denote amino acids important for binding to human ACE2 protein (Yuan et al., 2020; Chowdhury and Maranas, 2020), black squares designate amino acids important for binding to anti-Spike neutralizing antibody CR3022 (Yuan et al., 2020).



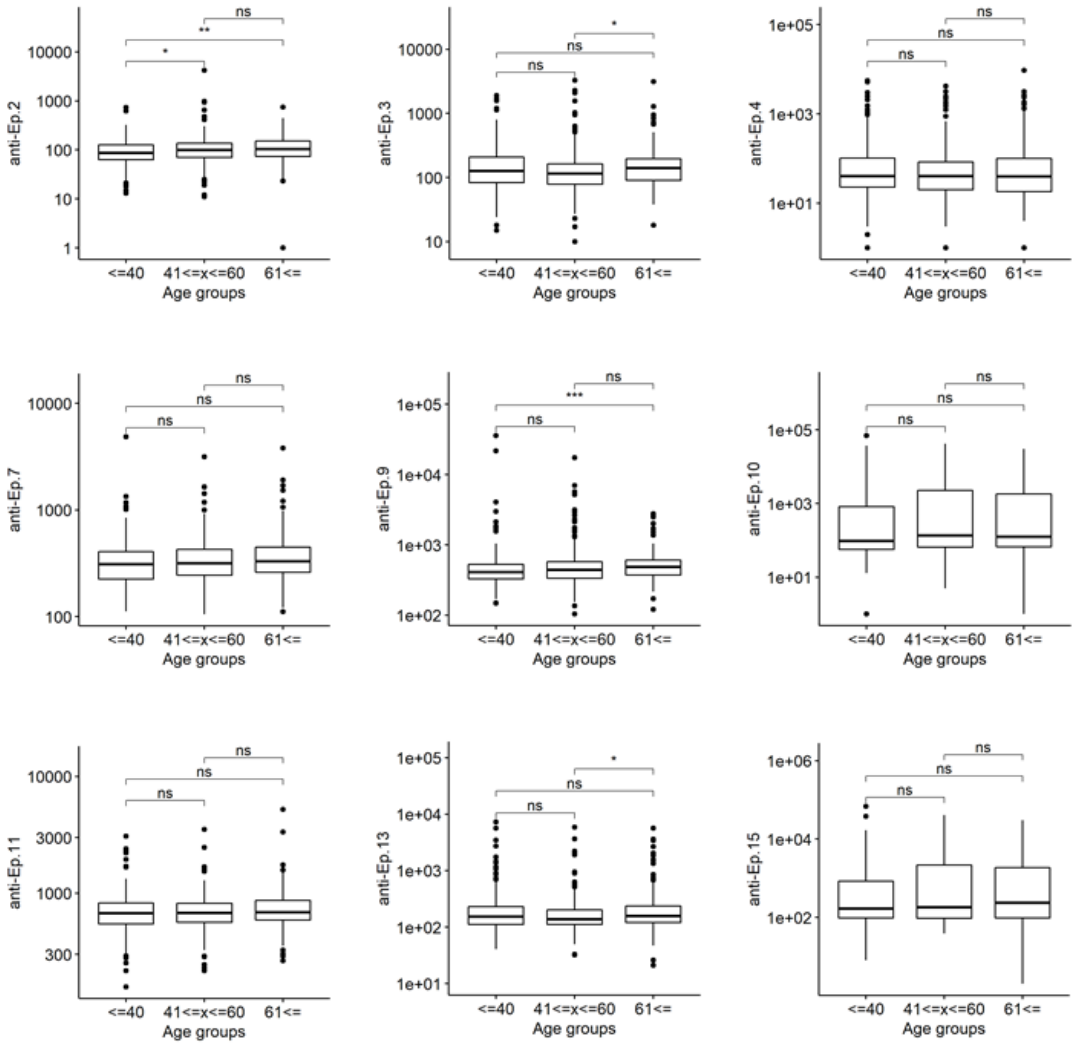
Supplementary Information

Extended Data Table 2. Multiple linear regression analysis for exploratory data analysis. R programming language and package "core" was used to perform multiple linear regression analysis. Multicollinearity of data was assessed using package "car" (Fox and Weisberg, 2019) and the variables used in the models showed minimal collinearity (variance inflation factor (VIF) < 5 or GVIF ^(1/2*Df) < 5 (generalized VIF)). The regression coefficient estimates for each considered categorical variable (1st column) are given for each epitope (columns 2-16) and significant predictors are marked accordingly: (.) p < 0.1, * p < 0.05; ** p < 0.01; *** p < 0.001. HT – hypertension, CVD – cardiovascular disease, HC – healthy control

Epitope	**Ep.1**	**Ep.2**	**Ep.3**	**Ep.4**	**Ep.5**	**Ep.6**	**Ep.7**	**Ep.8**	**Ep.9**	**Ep.10**	**Ep.11**	**Ep.12**	**Ep.13**	**Ep.14**	**Ep.15**
Model 1 (Age group + gender, n=524)															
Age group: 41<=x<=60	111.1(.)	41.4(.)	-12.6	-32.3	37.4	-76.7(.)	15.0	33.7	-61.9	117.0	-3.8	-154.9*	-66.7	137.9	105.0
Age group: 61<=	96.1	17.8	6.5	33.9	143.5***	-178.8***	45.4	126.7***	-228.0	112.7	58.6	2.9	35.5	-157.8	109.5
Gender: Male	-29.1	20.3	-35.2	202.2	22.1	-23.1	-41.8	15.6	-56.8	-1115.6*	-35.9	-82.4	-94.2	-50.4	-1120.9*
Model 2 (Age group + gender + HT + CVD, n=193)															
HT: noHT	141.2	-67.7***	58.4	118.0	-180.7**	40.4	-65.3	-91.2*	-124.5(.)	-574.9	28.58	-110.6	44.4	134.2	-570.6
CVD: HC	42.6	26.6	-2.5	20.8	57.9	53.8	-67.1	-99.9*	33.5	906.4	-71.62	-185.6	-86.9	-464.7(.)	905.6
Gender: Male	-136.9	-8.4	-35.3	170.6	107.1(.)	22.8	-26.3	72.4*	86.5	-265.2	-14.36	-202.8(.)	-10.5	202.8	-270.4
Age group: 41<=x<=60	300.8	57.9*	53.9	-62.8	57.4	20.2	23.2	135.2**	186.6(.)	551.7	143.2(.)	-47.2	4.5	-272.1	538.2
Age group: 61<=	238.2	33.7	77.5	115.7	125.8	-52.7	41.5	134.4*	117.3	1026.4	193.38*	18.3	-66.6	-434.8	1024.6
Model 3 (Age group + gender + HT + CVD + tobacco use, n=96)															
HT: noHT	63.0	-60.5*	37.7	-68.8	-259.4*	27.6	45.2	-5.8	-142.6	-316.0	48.4	27.3	73.8	27.0	-306.5
CVD: HC	-66.7	71.6*	-12.6	-44.1	116.7	127.3	-77.2**	-71.2	88.8	434.9	-8.4	-298.9	-80.9	-368.3	434.6
Gender: Male	1.2	-27.5	12.0	107.7	113.6	131.5	20.6	16.8	55.7	353.3	7.4	-490.1*	41.6	683.7(.)	301.3
Age group: 41<=x<=60	97.9	39.2	62.5	-5.1	203.8	147.3	6.1	70.5	434.7	576.3	-70.6	-416.5	190.7	-2355.2	307.5
Age group: 61<=	243.7	2.7	76.8	64.6	239.8	20.4	2.9	15.9	319.6	925.3	-178.9	-373.3	69.7	-2771.6(.)	670.4
Tobacco: No	210.6(.)	30.6	5.2	102.1	-156.6	69.0	-53.6	9.5	-11.1	-255.6	-9.5	-418.9(.)	36.8	147.1	-270.7
Tobacco: Smoking	24.6	70.9(.)	-35.3	31.6	-353.9(.)	80.9	-13.5	9.3	25.0	586.3	-104.3	-399.2	74.9	960.2(.)	551.4

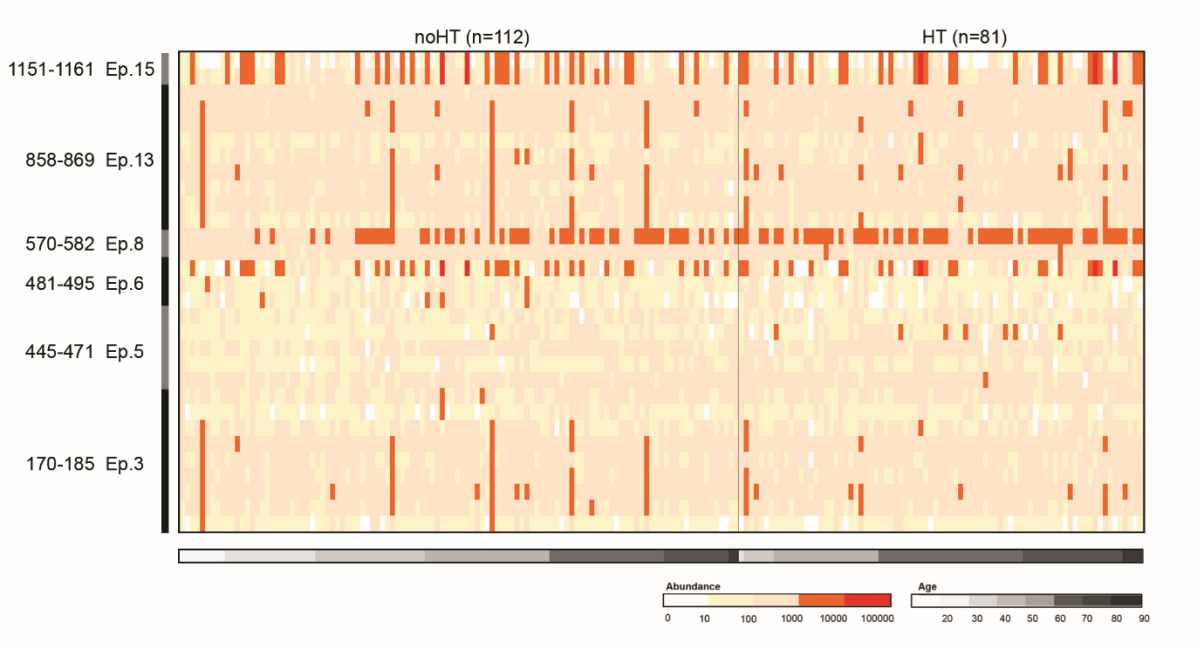
Supplementary Information

Extended Data Figure 3. Response against predicted S epitopes across age groups. Kruskal-Wallis test for overall difference of groups. Pair-wise comparisons with Mann-Whitney U test, ns p>0.05, * p<0.05, ** p<0.01, *** p<0.001. Group sizes: ≤40 years (n=186), 41≤x≤60 years (n=188), 61≤ years (n=150).



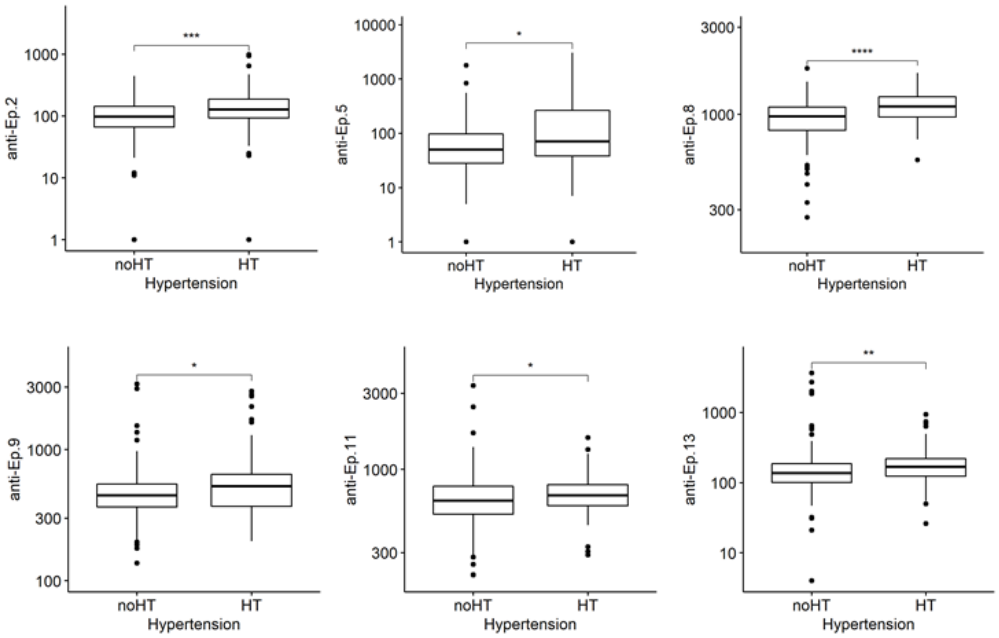
Supplementary Information

Extended Data Figure 4. Response against S epitopes in hypertension (HT) as an intensity plot. Immune response to HT-associated S epitopes was visualized across subjects (*in columns*) without HT (noHT, n=112) or with HT (n=81). Motifs aligned to predicted S epitopes are *in rows*. Motif-containing peptide antigen abundances in individual immunoprofiles are color-coded in $\log_{10}$. Subject age is color-coded under the intensity plot.



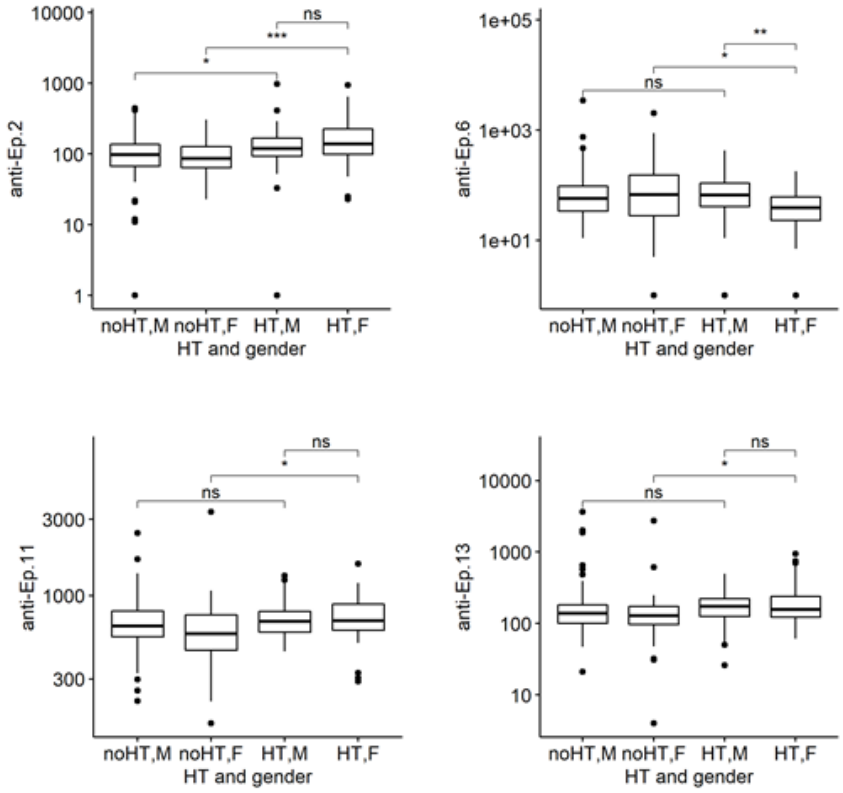
Supplementary Information

Extended Data Figure 5. Response against predicted S epitopes in hypertension (HT). Pair-wise comparisons with Mann-Whitney U test, ns p>0.05, * p<0.05, ** p<0.01, *** p<0.001, **** p<0.0001. Group sizes: noHT (n=122), HT (n=81)



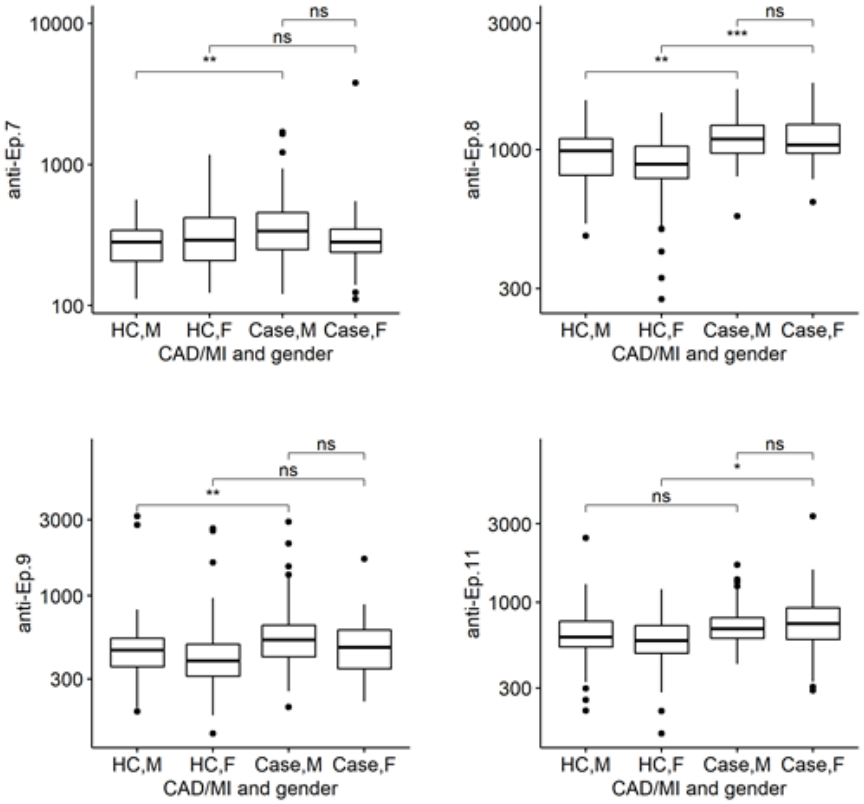
Supplementary Information

Extended Data Figure 6. Response against predicted S epitopes across hypertension and gender groups. Kruskal-Wallis test for overall difference of groups. Pair-wise comparisons with Mann-Whitney U test, ns p>0.05, * p<0.05, ** p<0.01, *** p<0.001. Group sizes: noHT, M (n=71), noHT, F (n=43), HT, M (n=48), HT, F (n=33).



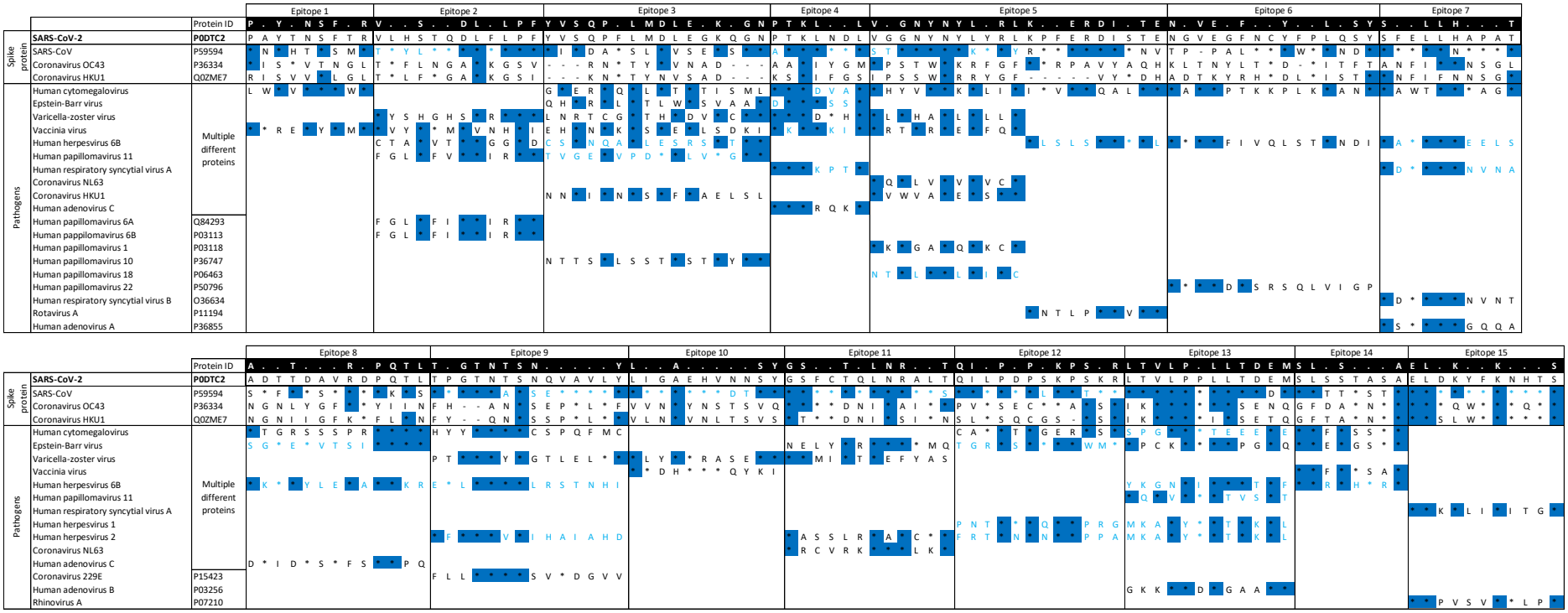
Supplementary Information

Extended Data Figure 7. Response against predicted S epitopes across cardiovascular disease (CAD or MI) and gender groups. Kruskal-Wallis test for overall difference of groups. Pair-wise comparisons with Mann-Whitney U test, ns p>0.05, * p<0.05, ** p<0.01, *** p<0.001. Group sizes: HC, M (n=40), HC, F (n=49), Case, M (n=79), Case, F (n=27)



Supplementary Information

Extended Data Figure 8. Sequence alignment of Spike glycoproteins (S) from beta-coronaviruses SARS-CoV-2, SARS-CoV, OC43 and HKU1, and the sequence similarity of predicted S epitopes with common viral pathogens. Predicted Ep.5 of S RBD, for which high immune response was associated with age and HT in males, shared sequence similarity with SARS-CoV S (70% amino acid identity) and with human cytomegalovirus (HCMV, 50% identity) and human herpesvirus 6B (HHV-6B, 50% identity). Although Ep.12 of SARS CoV-2 shares sequence identity to multiple already described immune epitopes of highly prevalent herpesviruses HCMV, Epstein-Barr virus (EBV, human herpesvirus 4), human herpesvirus 1 (HHV-1), and human herpesvirus 2 (HHV-2) with identity between 42-50%, all these share the same core pattern of P.P.KP. Predicted epitope sequences are depicted with white text on black background – letters stand for exact amino acid positions in the epitope, whereas dots (.) represent any given amino acid. Conserved amino acid positions between SARS-CoV-2 S protein and reference proteins from UniProtKB database are depicted with asterisks (*), absent amino acids with hyphen (-), and amino acids matching the S protein motifs are coloured dark blue. The known epitopes of aligned pathogen proteins from Immune Epitope Database (IEDB) are marked light blue. Aligned reference protein UniProtKB identification codes, alignment information and IEDB epitope reference IDs can be found in Extended Data Table 2.



Supplementary Information

Extended Data Table 3. Additional information about the sequence alignment of SARS-CoV-2 Spike glycoprotein and UniProtKB human virus reference proteins. The exact UniProtKB database identification codes of an aligned protein, its parental virus and virus strains are shown. The number of the first amino acid in the aligned reference protein is depicted along with identification numbers of known IEDB epitopes for the same pathogen protein. NA – not applicable, IEDB – Immune Epitope Database.

Predicted Spike glycopro-tein (S) epitope	Virus	Strain	Protein ID	Alignment start on protein	Matching epitope from IEDB (ID)
1	Human cytomegalovirus	AD169	P16799	329	
1	Vaccinia virus	Copenhagen	P20508	245	
2	Human herpesvirus 6B	NA	IEDB	NA	738622
2	Human papillomavirus 11	NA	P04014	216	
2	Human papillomavirus 6A	NA	Q84293	216	
2	Human papillomavirus 6B	NA	P03113	216	
2	SARS-CoV	NA	P59594	51	65159
2	Vaccinia virus	Copenhagen	P68441	507	
2	Varicella-zoster virus	Dumas	P09270	854	
3	Coronavirus HKU1	N5	Q0ZME7	1221	
3	Epstein-Barr virus	AG876	P0C723	260	
3	Human cytomegalovirus	AD169	P16729	1136	
3	Human herpesvirus 6B	Z29	Q9QJ26	1303	738619
3	Human papillomavirus 10	NA	P36747	378	
3	Human papillomavirus 11	NA	P04012	263	112499
3	Vaccinia virus	Copenhagen	P68694	974	
3	Varicella-zoster virus	Dumas	P09296	247	
4	Epstein-Barr virus	NA	IEDB	NA	694177
4	Human adenovirus C	Serotype 5	P04493	73	
4	Human cytomegalovirus	AD169	IEDB	NA	17424
4	Human respiratory syncytial virus A	A2	IEDB	NA	47690
4	SARS-CoV	Tor2	P59594	371	5019
4	Vaccinia virus	Copenhagen	P0CK20	44	60525
4	Varicella-zoster virus	Dumas	P09261	273	
5	Coronavirus HKU1	isolate N1	P0C6X2	2173	
5	Coronavirus NL63	NA	P0C6U6	3347	
5	Human cytomegalovirus	Merlin	Q6SWC3	19	
5	Human papillomavirus 1	NA	P03118	323	
5	Human papillomavirus 18	NA	P06463	95	111246
5	SARS-CoV	Tor2	P59594	432	66460
5	Vaccinia virus	Copenhagen	P20643	138	
5	Varicella-zoster virus	Dumas	P09263	684	
5	Human cytomegalovirus	AD169	P09695	92	
5	Human herpesvirus 6B	Z29	Q9QJ27	11	872103

5	Rotavirus A	isolate RVA/Human/Italy/VA70/1975/G4P1A[8]	P11194	652	
6	Human cytomegalovirus	AD169	P16809	27	
6	Human herpesvirus 6B	Z29	P52544	124	
6	Human papillomavirus 22	NA	P50796	431	
7	Human adenovirus A	Serotype 31	P36855	182	
7	Human cytomegalovirus	AD169	P16785	371	
7	Human herpesvirus 6B	Z29	Q9QJ38	232	859354
7	Human respiratory syncytial virus A	A2	P03420	509	96860
7	Human respiratory syncytial virus B	B1	O36634	509	
8	Epstein-Barr virus	AG876	Q1HVG4	376	696380
8	Human adenovirus C	Serotype 5	P12537	413	
8	Human cytomegalovirus	NA	P07387	455	
8	Human herpesvirus 6B	Z29	Q9QJ37	254	871643
9	Coronavirus 229E	NA	P15423	58	
9	Human cytomegalovirus	AD169	P16719	95	
9	Human herpesvirus 2	HG52	P89475	234	741803
9	Human herpesvirus 6B	Z29	P52549	625	739289
9	SARS-CoV	NA	P59594	585	71190
9	Varicella-zoster virus	Dumas	P09258	63	
10	SARS-CoV	Tor2	P59594	636	1460
10	Vaccinia virus	Copenhagen	P68439	721	
10	Varicella-zoster virus	Dumas	P09245	1082	
11	Coronavirus NL63	NA	P0C6X5	1528	
11	Epstein-Barr virus	B95-8	P25939	381	
11	Human herpesvirus 2	HG52	P89466	7	154180
11	SARS-CoV	Tor2	P59594	739	39023
11	Varicella-zoster virus	Dumas	Q65ZG0	47	
12	Epstein-Barr virus	AG876	Q1HVG4	304	694044
12	Human cytomegalovirus	AD169	P14334	13	
12	Human herpesvirus 1	17	P06477	53	742197
12	Human herpesvirus 2	NA	IEDB	NA	1393
12	SARS-CoV	NA	P59594	786	17816
13	Epstein-Barr virus	AG876	Q1HVH8	547	
13	Human adenovirus B	Serotype 7	P03256	102	
13	Human cytomegalovirus	AD169	IEDB	NA	194708
13	Human herpesvirus 1	17	P10211	806	741877
13	Human herpesvirus 2	HG52	P08666	803	741773
13	Human herpesvirus 6B	Z29	P52459	24	873012
13	Human papillomavirus 11	NA	P04015	56	145356
14	Epstein-Barr virus	B95-8	P03186	701	
14	Human cytomegalovirus	AD169	P16724	259	
14	Human herpesvirus 6B	Z29	Q9PX69	140	739012
14	Vaccinia virus	Copenhagen	P21047	362	
15	Human respiratory syncytial virus A	A2	P28887	2147	

15	Human rhinovirus A	Serotype 89 (strain 41467-Gallo)	P07210	1710	
15	SARS-CoV	NA	P59594	1133	11740

Supplement References

Chowdhury, R., & Maranas, C. D. (2020). From directed evolution to computational enzyme engineering—A review. AIChE Journal, 66(3), [e16847]. https://doi.org/10.1002/aic.16847

Kassambara, A. (2016). ggcorrplot: Visualization of a Correlation Matrix using ' ggplot2'. R package version 0.1.1.9000.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., So, R.T.Y., Lv, H., Mok, C.K.P., Wilson, I.A., 2020. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV. Science. https://doi.org/10.1126/science.abb7269

Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.