bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multi-pronged human protein mimicry by SARS-CoV-2 reveals bifurcating potential for MHC detection and immune evasion

AJ Venkatakrishnan[1], Nikhil Kayal[1], Praveen Anand[2], Andrew D. Badley[3], George M. Church[4], Venky Soundararajan[1,*]

[1] nference, Cambridge, Massachusetts, USA

[2] nference Labs, Bangalore, India

[3] Department of Infectious Diseases, Mayo Clinic, Rochester, MN, USA

[4] Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

Correspondence to: VS (venky@nference.net)

The hand of molecular mimicry in shaping SARS-CoV-2 evolution and immune evasion remains to be deciphered. We identify 33 distinct 8-mer/9-mer peptides that are identical between SARS-CoV-2 and human proteomes, along similar extents of viral mimicry observed in other viruses. Interestingly, 20 novel peptides have not been observed in any previous human coronavirus (HCoV) strains. Four of the total mimicked 8-mers/9-mers map onto HLA-B*40:01, HLA-B*40:02, and HLA-B*35:01 binding peptides from human PAM, ANXA7, PGD, and ALOX5AP proteins. This mimicry of multiple human proteins by SARS-CoV-2 is made salient by the targeted genes being focally expressed in arteries, lungs, esophagus, pancreas, and macrophages. Further, HLA-A*03 restricted 8-mer peptides are shared broadly by human and coronaviridae helicases with primary expression of the mimicked human proteins in the neurons and immune cells. This study presents the first comprehensive scan of peptide mimicry by SARS-CoV-2 of the human proteome and motivates follow-up research into its immunological consequences.

Introduction

Viral infection typically leads to T cell stimulation in the host, and autoimmune response associated with viral infection has been observed[1]. SARS-CoV-2, the causative agent of the ongoing COVID-19 pandemic, has complex manifestations ranging from mild symptoms like loss of sense of smell (anosmia)[2] to severe and critical illness[3,4]. While some molecular factors governing SARS-CoV-2 infection of lung tissues, such as the ACE2 receptor expressing cells have been characterized recently[5], the mechanistic rationale underlying immune evasion and multi-system inflammation (Kawasaki-like disease) remains poorly understood[6,7].

The SARS-CoV-2 genome encodes 14 structural proteins (e.g. Spike protein) and non-structural proteins (e.g. RNA-dependent RNA polymerase), as depicted in **Figure 1a**. The non-structural ORF1ab polyprotein undergoes proteolytic processing to give rise to 15 different proteins (NSP1, NSP2, PL-PRO, NSP4, 3CL-PRO, NSP6, NSP7, NSP8, NSP9, NSP10, RdRp, Hel, ExoN, NendoU and 2'-O-MT). The human reference proteome consists of 20,350 proteins, which when alternatively spliced, result in over 100,000 protein variants (**Figure 1b**)[8].

Here, we investigate the potential for molecular mimicry-mediated escape from immune surveillance and host antigen recognition in COVID-19, by performing a systematic comparison

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of known MHC-binding peptides from humans and mapping them onto SARS-CoV-2 derived peptides (see *Methods*). For this, we compute the longest peptides that are identical between SARS-CoV-2 reference proteins and human reference proteins, thus creating a map of COVID-19 host-pathogen molecular mimicry. Based on this resource, we extrapolate the potential for HLA Class-I restricted, T-cell immune stimulation via synthesizing established experimental evidence around each of the mimicked peptides.

Methods

Computing the SARS-CoV-2 peptides that mimic MHC Class-I binding peptides on the human reference proteome

The reference proteome for SARS-CoV-2 consists of 14,221 8-mers and 14,207 9-mers, that result in 9,827 distinct 8-mers and 9,814 distinct 9-mers (see **Table S1**). The 8-mers and 9-mers are generated by using a sliding window, moving one amino acid at a time, resulting in overlapping linear peptides. The reference human proteome, on the other hand, consists of 11,211,806 peptides that are 8-mers and 11,191,459 peptides that are 9-mers, resulting in 10,275,893 unique 8-mers and 10,378,753 unique 9-mers. Including alternatively spliced variants, increases the unique peptide counts to 11,215,743 8-mers and 11,342,401 9-mers.

Thirty one 8-mer peptides and two 9-mer peptides are identical between the reference proteomes of SARS-CoV-2 and humans, after including alternatively spliced variants (**Figure 1b, Table S2**). For comparison, we analyzed the protein sequences from 9434 viral species (taxons) from NCBI RefSeq (see *Methods*). On average around 45.57 unique 8mer/9mers per viral taxon were shared with the human proteome (mean = 45.57, median = 12, S.D. = ±135.38). In order to control for the complexity and constraints of the amino acid sequences, we also analyzed the distribution of mimicked 8-mers/9-mers normalized by the total number of unique 8-mers/9-mers present in each viral taxon. On average, a fraction of 0.002 8-mers/9-mers out of all the unique 8-mers/9-mers in the virus are identical between the viral proteome and the human proteome (mean = 0.002, S.D. =± 0.011)(**Figure 1 - Supplemental Figure 1**). The fraction of 33 human-mimicking 8-mers/9-mers is proximal to the mean. Overall, this suggests that the presence of 33 human-mimicking 8-mers/9-mers in SARS-CoV-2 is not surprising compared to the number of human-mimicking 8-mers/9-mers present in other viruses.

In SARS-CoV-2, no 10-mer or longer peptides are identical between the pathogen and the host reference proteins. Of these, 29 peptides (8-mer/9-mer) mimicked by SARS-CoV-2 map onto nine of 14 SARS-CoV-2 proteins and 29 of 20,350 human proteins. By including alternative splicing, the 33 8-mers/9-mers mimicked by SARS-CoV-2 map onto 39 of 100,566 protein splicing variants. That is, 0.16% of human proteins and 0.04% of all splicing variants have 8-mer/9-mer peptides that are mimicked by the SARS-CoV-2 reference proteome. Given that MHC Class I alleles typically engage peptides that are 8-12mers[9], the analysis that follows was restricted to mapping host-pathogen mimicry from an immunologic perspective across 8-mer and 9-mer peptides only.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comparative analysis of SARS-CoV-2 peptides mimicking the human proteome with the reference SARS, MERS and seasonal HCOVs

Of the 33 peptides from SARS-CoV-2 that mimic the human reference proteome, 20 peptides are not found in any previous human-infecting coronavirus (SARS, MERS or seasonal HCoVs) (**Table 1**). The UniProt database was used to download the 15 protein reference sequences for SARS-CoV. The non-redundant set of protein sequences from other coronavirus strains (HCoV-HKU1:188; HCoV-229E: 246; HCoV-NL63: 330; HCoV-OC43: 910 and MERS: 681) was computed by removing 100% identical sequences, and the remnant sequences were all included in the comparative analysis with SARS-CoV-2 mimicked peptides. A Venn diagram depicting the overlap of mimicked peptides and identifying unique SARS-CoV-2 mimicked peptides was generated (**Figure 1c)**.

Given the zoonotic transmission potential of coronaviruses from other organisms to humans, we also considered 8-mer/9-mer peptides derived from 13,431 distinct protein sequences of all non-human coronaviridae from the VIPERdb[10].

Characterizing the SARS-CoV-2-derived 8-mers/9-mers that mimic established human MHC-binding peptides

The lengths of the human proteins mimicked by SARS-CoV-2 were considered to examine any potential bias towards the larger human proteins (**Figure 1-Supplementary Figure 1**). For instance, human Titin (TTN), despite containing 34,350 amino acids and being of the longest human proteins, does not have even one peptide mimicked by SARS-CoV-2. The longest and shortest human proteins that are mimicked by SARS-CoV-2 are MICAL3 (length = 2002 amino acids) and BRI3 (length = 125 amino acids) respectively.

The sequence conservation of each mimicked peptide was derived from all the 46,513 sequenced SARS-CoV-2 genomes available in the GISAID database (as on 06/13/2020).

The immune epitope database (IEDB)[11] was used to examine the experimentally established, in-vitro evidence for MHC presentation against human or SARS-CoV antigens. The peptides of potential immunologic interest were identified from the IEDB database using the following pair of assays. One of the assays involved purification of specific MHC-class I alleles and estimating the K_d values of specific peptide-MHC complexes through competitive radiolabeled peptide binding[12]. The other assay uses mass spectrometry proteomic profiling of the peptide-MHC complexes, where the MHC complexes were purified from the cell lines specifically engineered to produce mono-allelic MHC class I molecules. The identity of the peptide sequence bound to the Class-I MHC molecules was elucidated using mass-spectrometry[13].

Analysis of RNA expression in cells and tissues

A distribution of RNA expression across all the expressing samples collected from GTEx, Gene expression omnibus, TCGA and CCLE is created. In this distribution, a high-expression group is defined as the set of samples associated with the top 5% of expression level. Enrichment score captures the significance of the token in the high-expression group. The significance is captured based on Fisher's test along with Benjamini Hochberg correction. In comparison of gene

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression across tissue types in GTEx, the specificity of expression is computing using ‘Cohen's D’, which is an effect size used to indicate the standardised difference between two means.

Analysis of overlapping peptides between the proteome and viral proteomes

We analyzed the protein sequences from 9434 viral species (taxons) from ncbi refseq (https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/). On average around 45.57 unique 8mer/9mers per viral taxa were detected to be identical to a known human protein (mean = 45.57, median = 12, sd=±135.38). The highest number of identical peptides were detected in ‘Pandoravirus dulcis’ virus with 4423 peptides having an exact identical match to one or more human proteins.

Results

Identifying specific human peptides mimicked by SARS-CoV-2 and with in-vitro evidence for MHC binding

A set of 20 8-mer/9-mer peptides are mimicked by SARS-CoV-2 and no other human coronaviruses (see **Methods**). Of these 20 peptides, four peptides are constituted within established MHC-binding regions (**Table 1 - Panel A**). The 4 peptides with specific MHC-binding potential that are novel to SARS-CoV-2 map onto the following human proteins: Alpha-Amidating Monooxygenase (PAM), Annexin A7 (ANXA7), Peptidylglycine 6-phosphogluconate dehydrogenase (PGD), and Centromere protein I (CENPI) (**Figure 1d**). Analyzing the sequence conservation of the SARS-CoV-2-exclusive peptides shared with the above 4 human MHC binding peptides, shows that these SARS-CoV-2 peptides are largely conserved till date (**Table 2**). The previous human-infecting coronavirus strains (SARS-CoV, MERS, seasonal HCoVs) are notably bereft of these novel SARS-CoV-2 epitopes. An alternatively spliced variant of the human arachidonate 5-lipoxygenase activating protein (ALOX5AP - ENSP00000479870.1; ENST00000617770.4) contains an 8-mer peptide that is mimicked by SARS-CoV-2 as well as SARS-CoV, but not any of the seasonal HCoVs. This peptide has in-vitro evidence for positive MHC Class-I binding. Additionally, there are 4 human helicases (MCM8, DNA2, MOV10L1, ZNFX1), each containing peptides with established evidence of MHC Class-I binding that are also mimicked by SARS-CoV-2 and by previous human-infecting coronaviridae strains (**Table 1 - Panel B; Figure 1d**)(**Table 2**).

Novel mimicry of human PAM, ANXA7, and PGD by SARS-CoV-2, suggests an enrichment of mimicked peptides in lung, esophagus, arteries, heart, pancreas, and macrophages

Considering bulk RNA-seq data from over 125,000 human samples with non-zero PAM expression shows that PAM is highly expressed in pancreatic islets (enrichment score = 276.9 across 6 studies and 552 samples), artery (enrichment score = 244.5 across 3 studies and 343 samples), heart (enrichment score = 243.6 across 19 studies and 442 samples), aorta (enrichment score = 217.1 across 2 studies and 304 samples), embryonic stem cells (enrichment score = 146.7 across 2 studies and 352 samples), and fibroblasts (enrichment score = 107.7 across 28 studies and 215 samples) (**Figure 2a**). Among 54 human tissues from GTEx, PAM is particularly significant within aortic arteries (n = 432, Cohen's D = 3.1, mean = 348.2 TPM)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compared to other human tissues. Moderate specificity in gene expression is also noted for the atrial appendage of the heart (n = 429, Cohen's D = 2.2, mean = 461.3 TPM) (**Figure 2b**). Further, immunohistochemistry (IHC) data on 45 human tissues [14] shows that the PAM protein is detected at high levels in heart muscles, epididymis, and the adrenal gland (**Figure 2b**).

Exploring all available human Single Cell RNA-seq (scRNA-seq) data shows PAM is expressed in nearly 100% of pancreatic gamma cells, alpha cells, beta cells, delta cells and epsilon cells as well as between 50-90% of activated/quiescent stellate cells, acinar cells, endothelial cells, ductal cells of the pancreas (nferX scRNAseq app - Pancreas). It is also expressed in over 80% of cardiomyocytes, and 40-70% of heart fibroblasts, macrophages, endothelial cells, and smooth muscle cells (nferX scRNAseq app - Heart), as well as in 26-27% of lung pleura fibroblasts, stromal cells, and neutrophils (**Figure 2c**, nferX scRNAseq app - Lung Pleura). Moreover, analyzing the scRNA-seq data from severe COVID-19 patient's lung bronchoalveolar lavage fluid shows high PAM expression in club cells (**Figure 2d**, nferX scRNAseq app - Lung bronchoalveolar lavage fluid), which intriguingly also express the SARS-CoV-2 receptor ACE2 significantly [5]. Furthermore, esophagus scRNA-seq analysis shows esophageal mucosal cells and stromal cells as significant PAM expressors (**Figure 2e**, nferX scRNAseq app - Esophagus). Finally, rarer cell types such as pulmonary neuroendocrine cells and goblet cells of the lungs, and some common cell types like lung serous cells and respiratory secretory cells also express PAM significantly.

Similar to the expression profile of PAM, examining 130,400 human samples with non-zero ANXA7 expression shows that ANXA7 is highly expressed in pancreatic islets (enrichment score = 286.65; 543 samples; 3 studies) and artery (enrichment score = 161.68; 184 samples; 3 studies) (**Figure 3 - SupplementaryFigure1**). ANXA7 is significantly expressed in the aortic artery (n = 432, Cohen's D = 2.1, mean = 163.1 TPM) and the tibial artery (n = 663, Cohen's D = 2.6, mean = 176.4 TPM) (**Figure 3 - SupplementaryFigure2**). Analysis of the scRNA-seq data on ANXA7 confirms expression in endothelial cells across multiple tissues and organs (nferX Single Cell app - Uterus), and also indicates expression in lung type-2 pneumocytes (nferX Single Cell app - Lung), macrophages, oligodendrocytes (**Figure 3a**). Type-2 pneumocytes are noted to express the SARS-CoV-2 receptor ACE2 from scRNA-seq [5]. Analyzing the lung bronchoalveolar lavage fluid scRNA-seq data from patients with severe COVID-19 outcomes shows macrophages, lung epithelial cells, T-cells, club cells, proliferating cells, and plasma cells are significant expressors of ANXA7 (**Figure 3b**, nferX scRNAseq app - Lung Bronchoalveolar Lavage Fluid). Additionally from the study of normal lungs, activated dendritic cells and lymphatic vessel cells are noted to express ANXA7 significantly (**Figure 3c**, nferX scRNAseq app - Lungs).

Assessment of around 128,000 human samples shows that PGD is highly expressed in esophagus mucosa (enrichment = 323, n=510, 2 studies), blood (enrichment = 320.5, n=1020, 39 studies), and macrophages (enrichment = 141.6, n=202, 4 studies). (**Figure 4 - SupplementaryFigure1**). IHC data on 45 human tissues from the Human Protein Atlas [14] confirms that PGD is detected at high levels in the esophagus (**Figure 4 - SupplementaryFigure2**), and additionally in the testes, tonsils, bone marrow, gallbladder, spleen and placenta.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Unlike the expression profiles of PAM, ANXA7 and PGD, CENPI's expression is fairly non-specific and relatively negligible from available data sets. Mild to moderate expression of CENPI is seen in precursor B cells and late erythroid cells, but further studies are needed to ascertain the significance, if any, of CENPI expression, including in the context of COVID-19.

Multi-pronged mimicry of PAM, ANXA7, and PGD by SARS-CoV-2 and its potential for factoring into the pulmonary-arterial autoinflammation seen in severe COVID-19 patients

Positive HLA-B*40:02 binding has been established for the human PAM peptide ('KE**PGSGVPVV**L') and the ANXA7 peptide ('V**ESGLKTIL**') [15–17]**,** that contain the distinctive mimicking SARS-CoV-2 peptides **(Table 1)**. The closely related HLA-B*40:01 allele also binds this mimicked ANXA7 peptide [17]. The corresponding mimicking peptides of PAM and ANXA7 are from the viral RNA-dependent RNA polymerase and NSP2 protein respectively, which are constituted within the MHC-binding regions (highlighted above in **bold text**). Additionally, HLA-B*35:01 has experimental evidence for positive binding of the human PGD peptide mimicked by the SARS-CoV-2 virus (**Table 1**).

Given the high expression of ANXA7, PGD and PAM among cells of the respiratory tract, lungs, arteries, cardiovascular system, and pancreas, as well as in macrophages, their striking mimicry by SARS-CoV-2 raises the possibility of individuals with HLA-B*40 and HLA-B*35 alleles being predisposed to potential immune evasion or autoinflammation. Indeed, the potential for broad vascular/endothelial autoinflammation is consistent with the rarer multi-system inflammatory syndrome (MIS-C) or atypical Kawasaki disease noted in few COVID-19-infected children [18,19].

Alternatively spliced human protein variants analysis for mimicry by SARS-CoV-2 highlights another HLA-B*40:01 restricted protein (ALOX5AP) with autoimmune potential

A splicing variant of ALOX5AP (ENSP00000479870.1; ENST00000617770.4) containing the 8-mer peptide 'PEANMDQE' is one of 4 alternatively spliced human protein variants that are mimicked by SARS-CoV-2. The other three 8-mer peptides arising from splicing variants do not have any known class I MHC binding reported in the immune epitope database. However, SARS-CoV, which is the only other human-infecting coronavirus in addition to SARS-CoV-2 that also contains **PEANMDQE**, has been experimentally established to possess the **PEANMDQE**SF epitope that positively binds to the HLA-B*40:01 allele.

ALOX5AP is known from literature knowledge synthesis to be associated with ischemic stroke, myocardial infarction, atherosclerosis, cerebral infarction, and coronary artery disease (**Figure 5a**). Single cell RNA-seq studies show numerous types of macrophages expressing ALOX5AP, including in the lungs and brain temporal lobe. Epithelial cells and proliferating cells of the lungs also express ALOX5AP, as do other types of immune cells such as T-cells, neutrophils, and dendritic cells (**Figure 5b**). Taken together with the HLA-B*40:01 restricted binding of the PAM and ANXA7 peptides mimicked by SARS-CoV-2, the ALOX5AP splicing variant also mimicked by SARS-CoV-2 suggests the possibility of immune evasion or autoinflammation in HLA-B*40-constrained COVID-19 patients.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HLA-A*03-binding peptides are shared between helicases of all known human-infecting coronaviridae (HCoVs) and the human proteome

There are seven human protein mimicking peptides that are shared between SARS-CoV-2 and at least one other human infecting coronavirus (**Table 1C**). These proteins include DNA2, MCM8, MOV10L1, ZNFX1, which are all helicases. Analysis of single cell RNAseq suggests that the mimicked human proteins are expressed in neuronal cells and immune cells (**Figure 6c**). The HLA-A*03:01 allele has been established from in-vitro experiments to bind SARS-CoV helicase peptides that mimic an 8-mer peptides from human MOV10L1, DNA2, ZNFX1 and MCM8 helicases (summarized in **Table 1**) [20]. The HLA-A*31:01 and HLA-A*11:01 alleles, on the other hand, are known to bind peptides containing the human MOV10L1, DNA2, and ZNFX1 helicase mimics; whereas the HLA-A*68:01 allele has established in-vitro evidence of binding peptides containing the human DNA2 and MOV10L1 helicase mimics [21]. In some of these individuals carrying these HLA alleles (**Table 1C**), a positive T-cell response against their “self” cells that express and display the above coronavirus-mimicked peptides seems plausible.

Discussion

The presence of identical peptides between viruses and humans has at least two potential implications from an immunologic standpoint. On the one hand, upon presentation of the viral antigens on the surface of infected cells, the virus may evade immune response by masquerading as a host peptide and the recognition of the shared peptides by host regulatory T cells could promote a generally immunosuppressive environment. On the other hand, an autoimmune response can lead to virus-induced autoinflammatory conditions[22]. Either response requires the coupling of both the presence of the appropriate HLA allele and positive T-cell response towards the mimicked peptide epitopes[23]. It is possible that SARS-CoV-2 leverages one or both of these molecular mimicry strategies to exploit the host immune system. In a small minority of patients who happen to have the unfortunate combination of MHC restriction and T-cell receptors as mentioned above, the specific tissues and cell types harboring the mimicked protein would bear the brunt of sustained autoimmune damage. The autoimmune lung and vascular damage reported in severe COVID-19 patient mortality[24–26] necessitates hypothesis-free examination of both these mimicry strategies.

Our study suggests HLA binding of peptides based on existing literature, but existing literature is by no means exhaustive for identifying HLA binding[11]. There is no known HLA Class-I mediated positive T-cell response against certain 8-mers documented in the immune epitope database. For example, GPPGTGKS peptide is shared by the viral helicase and human VPS4A, VPS4B and SETX. This peptide is also shared with seasonal human coronaviruses (HCoV-OC43; HCoV-HKU1) and previous SARS strains (SARS-CoV; MERS). Further experiments are required to assess any potential for autoinflammation.

Although our current study focussed on human infecting coronaviruses, molecular mimicry is expected to exist beyond human infecting coronaviruses. A stringent BLAST search was also

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

performed for all the four immunomodulatory peptides specific to SARS-CoV-2 against all the sequences of Coronaviridae family in the non-redundant protein database. There were no hits found outside the orthocoronavirinae family for these peptides. An exact match for peptides - ‘PGSGVPVV’, ‘VTLIGEAV’ and ‘SLKELLQN’ was found only in either the pangolin coronavirus or the Bat coronavirus RaTG13. An exact match for ‘PGSGVPVV’ was also found in canada goose coronavirus (YP_009755895.1). The human ANXA7-mimicking peptide ‘ESGLKTIL’ is however noted only in SARS-CoV-2 sequences, with the closest known evolutionary homologs attributed to BAT SARS-like coronavirus (ESGLKTIL), the NL63-related bat coronavirus strains, and the recently sequenced pangolin coronavirus (**Figure 3 - Supplementary Figure2)**.

Our observed multi-pronged human mimicry of SARS-CoV-2, including in peptides that are notably missing from all previously human-infecting coronavirus strains, may in conclusion, owe their origins to zoonotic transmission from coronaviruses circulating within pangolins and bats as natural reservoirs, aided by genetic recombination and purifying selection[27,28]. Our hypothesis-free computational analysis of all available sequencing data, from genomic sequencing and single cell transcriptomics across the host-pathogen continuum, sets the stage for targeted experimental interrogation of immuno-evasive or immuno-stimulatory roles of the mimicked peptides within zoonotic reservoirs and human subjects alike. Such a holistic data sciences-enabled “wet lab” platform for characterizing molecular mimicry and its immunologic implications may help shine a new lens on the relentless evolutionary tinkering that propels the rise and fall of viral pandemics.

Figure Legends

Figure 1. Molecular mimicry and immunomodulatory potential (a) n-mer peptide generation. (b) Mimicked peptides between SARS-CoV-2 and human proteomes. (c) Comparison of human-protein mimicking SARS-CoV-2 peptides with peptides from other human coronaviruses (d) Immunomodulatory potential of mimicked peptides from SARS-CoV-2.

Figure 2. Multi-omics analysis of human PAM. (a) (left) Universal bulk RNA-seq analysis of all available human data shows pancreatic islets, heart, artery, aorta and embryonic stem cells harbor PAM significantly. (right) Single cell RNA-seq (scRNA-seq) confirms high PAM-expressing cells include multiple pancreatic cells, cardiomyocytes, goblet cells of the lung, bronchus and intestines, stromal cells of the digestive system, and fibroblasts of multiple organs including the lung, trachea, bronchus, intestines, and heart. **(b)** Analysis of tissue-specific expression pattern of PAM from bulk RNA-seq (GTEx) and triangulation with IHC antibody staining data (HPA) suggests artery, aorta, and myocytes of the heart muscle as significant PAM-expressing tissues. **(c)** Severe COVID-19 patient’s lung bronchoalveolar lavage fluid shows high PAM expression in club cells, which also express the SARS-CoV-2 receptor ACE2 (nferX scRNAseq app - Lung Broncheoalveolar Lavage Fluid).

Figure 3. Evidence of ANXA7 and ACE2 expression from human single cell RNA-sequencing data of the lungs. (a) List of high ANXA7-expressing cells across human tissues. **(b)** High ANXA7-expressing cells in the lungs include macrophages, proliferating cells, mast

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cells, stromal cells, Type-2 pneumocytes and endothelial cells; **(c)** Lung bronchoalveolar lavage fluid scRNA-seq shows multiple high ANXA7-expressing cells including macrophages, lung epithelial cells, T-cells, club cells, proliferating cells, and plasma cells (nferX scRNAseq app - Lung Bronchoalveolar Lavage Fluid); **(d)** From the lungs, activated dendritic cells and lymphatic vessel cells are seen to express ANXA7 significantly (nferX scRNAseq app - Lungs).

Figure 4. Evidence of PGD expression from human single cell RNA-sequencing data (a) Single cell RNA-seq (scRNA-seq) shows expression in cell types of the human lungs **(**nferX scRNAseq app - Lungs, nferX scRNAseq app - Lung Pleura, nferX scRNAseq app - Airway Epithelia), and artery (nferX scRNAseq app - Arteries).

Figure 5. Evidence for ALOX5AP from biomedical knowledge synthesis and single cell RNA-seq. (a) Knowledge synthesis suggests involvement of ALOX5AP in ischemic stroke, myocardial infarction, atherosclerosis, cerebral infarction, and coronary artery disease. **(b)** scRNA-seq shows significant expression of ALOX5AP in proliferating cells, macrophages, T-cells, and epithelial cells from the lungs (nferX scRNAseq app - Lung Bronchoalveolar Lavage Fluid, nferX - Lungs) and macrophages of the brain (nferX scRNAseq app - Brain).

Figure 6. scRNAseq based expression analysis of human helicases containing peptides mimicked by helicases in SARS-CoV-2 and other human coronaviruses.

Figure 1 - Supplemental figure 1. Mimicked 8-mers/9-mers between human and viral proteomes (a) Distribution of mimicked 8-mers/9-mers between human and viral proteomes. **(b)** Distribution of mimicked 8-mers/9-mers between human and viral proteomes normalized by the number of unique 8-mers/9-mers.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

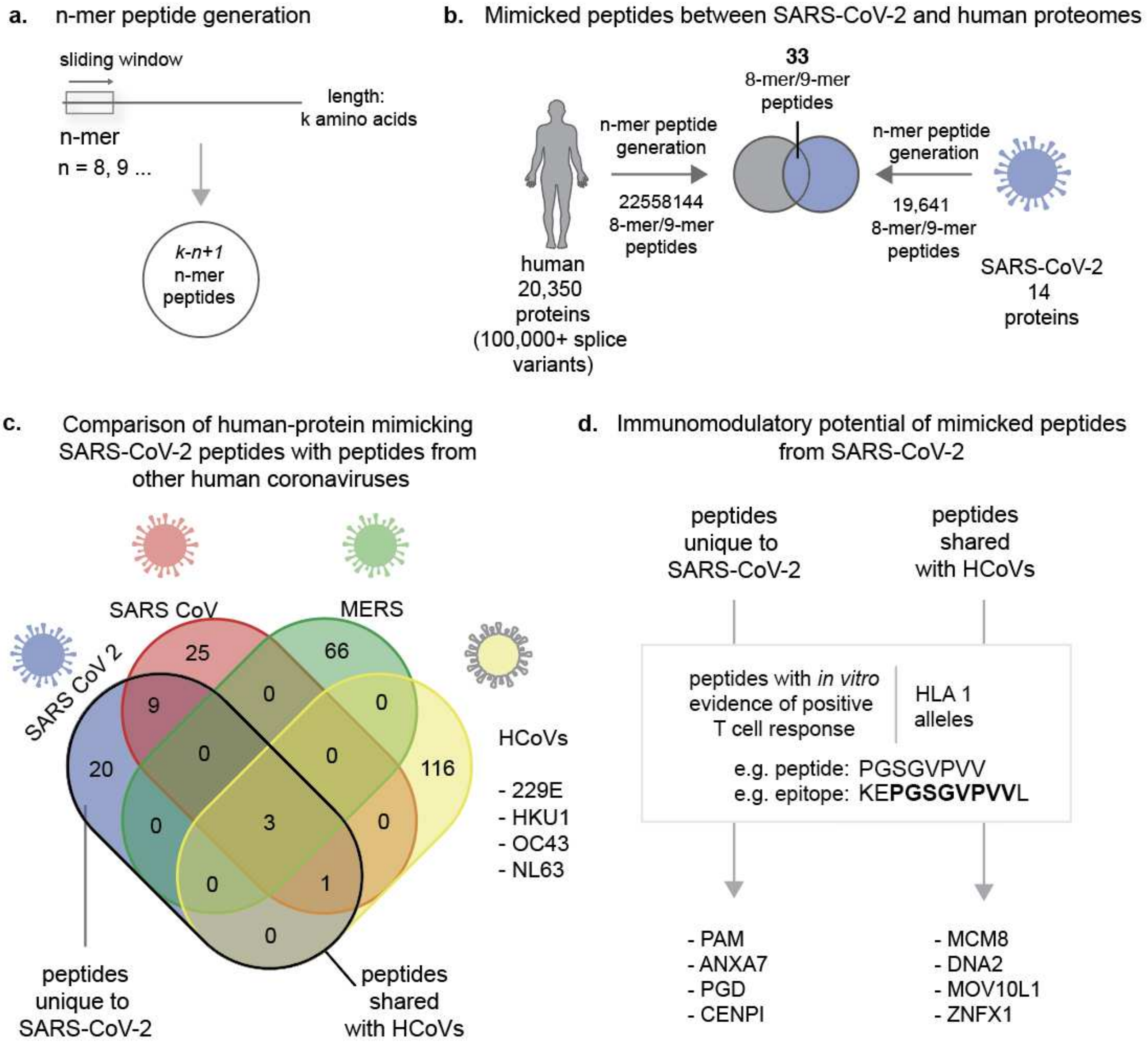
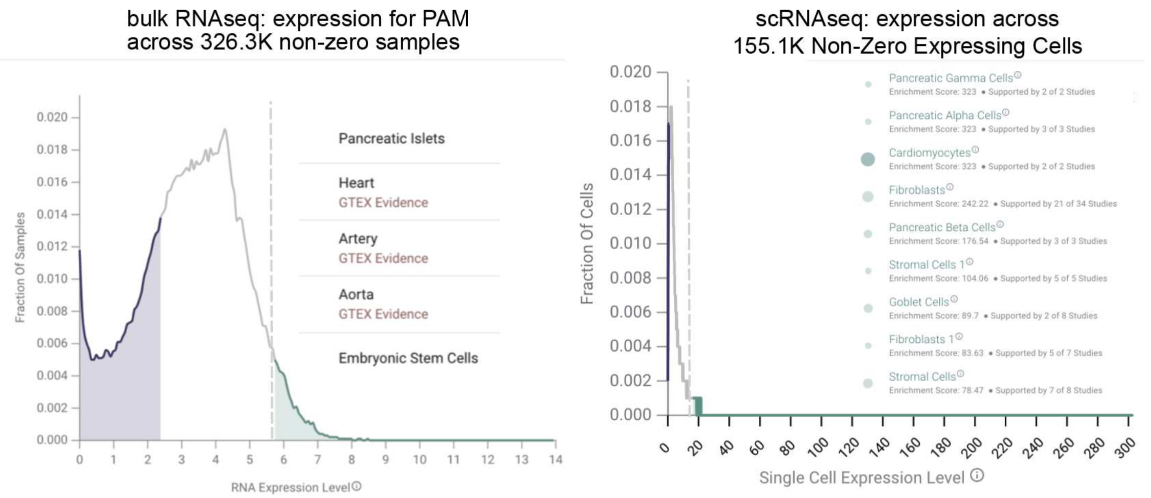


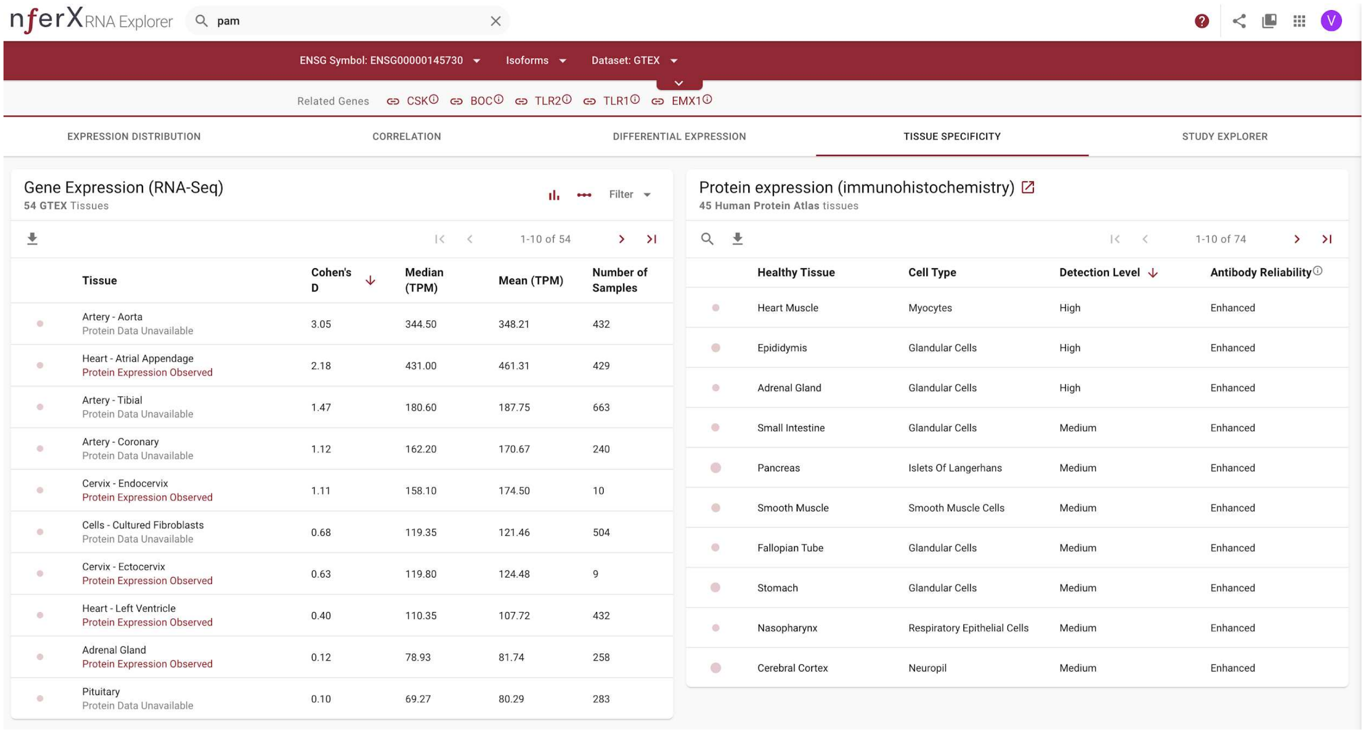
Figure 1

a

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



b



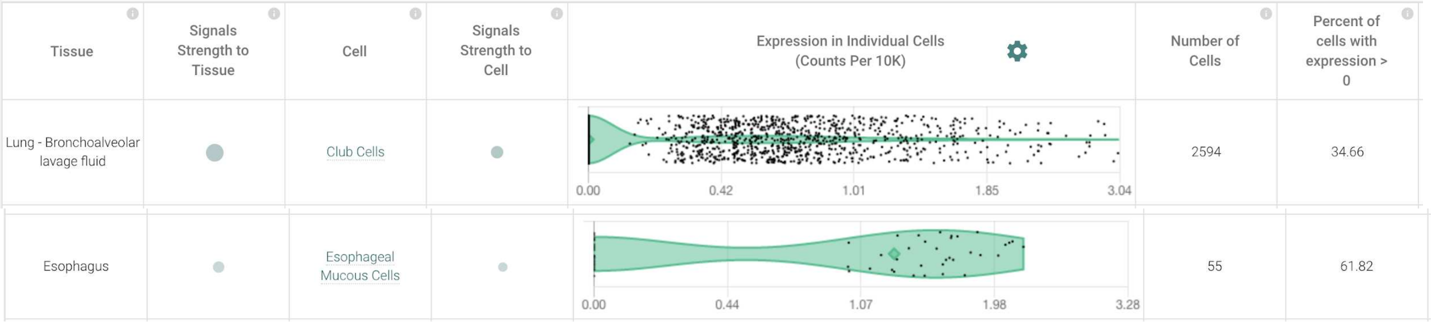
Gene Expression (RNA-Seq)
54 GTEX Tissues

Tissue	Cohen's D	Median (TPM)	Mean (TPM)	Number of Samples
Artery - Aorta Protein Data Unavailable	3.05	344.50	348.21	432
Heart - Atrial Appendage Protein Expression Observed	2.18	431.00	461.31	429
Artery - Tibial Protein Data Unavailable	1.47	180.60	187.75	663
Artery - Coronary Protein Data Unavailable	1.12	162.20	170.67	240
Cervix - Endocervix Protein Expression Observed	1.11	158.10	174.50	10
Cells - Cultured Fibroblasts Protein Data Unavailable	0.68	119.35	121.46	504
Cervix - Ectocervix Protein Expression Observed	0.63	119.80	124.48	9
Heart - Left Ventricle Protein Expression Observed	0.40	110.35	107.72	432
Adrenal Gland Protein Expression Observed	0.12	78.93	81.74	258
Pituitary Protein Data Unavailable	0.10	69.27	80.29	283

Protein expression (immunohistochemistry)
45 Human Protein Atlas tissues

Healthy Tissue	Cell Type	Detection Level	Antibody Reliability
Heart Muscle	Myocytes	High	Enhanced
Epididymis	Glandular Cells	High	Enhanced
Adrenal Gland	Glandular Cells	High	Enhanced
Small Intestine	Glandular Cells	Medium	Enhanced
Pancreas	Islets Of Langerhans	Medium	Enhanced
Smooth Muscle	Smooth Muscle Cells	Medium	Enhanced
Fallopian Tube	Glandular Cells	Medium	Enhanced
Stomach	Glandular Cells	Medium	Enhanced
Nasopharynx	Respiratory Epithelial Cells	Medium	Enhanced
Cerebral Cortex	Neuropil	Medium	Enhanced

c

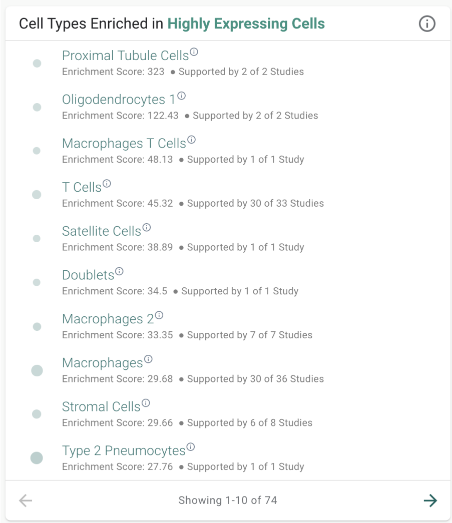


Tissue	Signals Strength to Tissue	Cell	Signals Strength to Cell	Expression in Individual Cells (Counts Per 10K)	Number of Cells	Percent of cells with expression > 0
Lung - Bronchoalveolar lavage fluid	●	Club Cells	●		2594	34.66
Esophagus	●	Esophageal Mucous Cells	●		55	61.82

Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

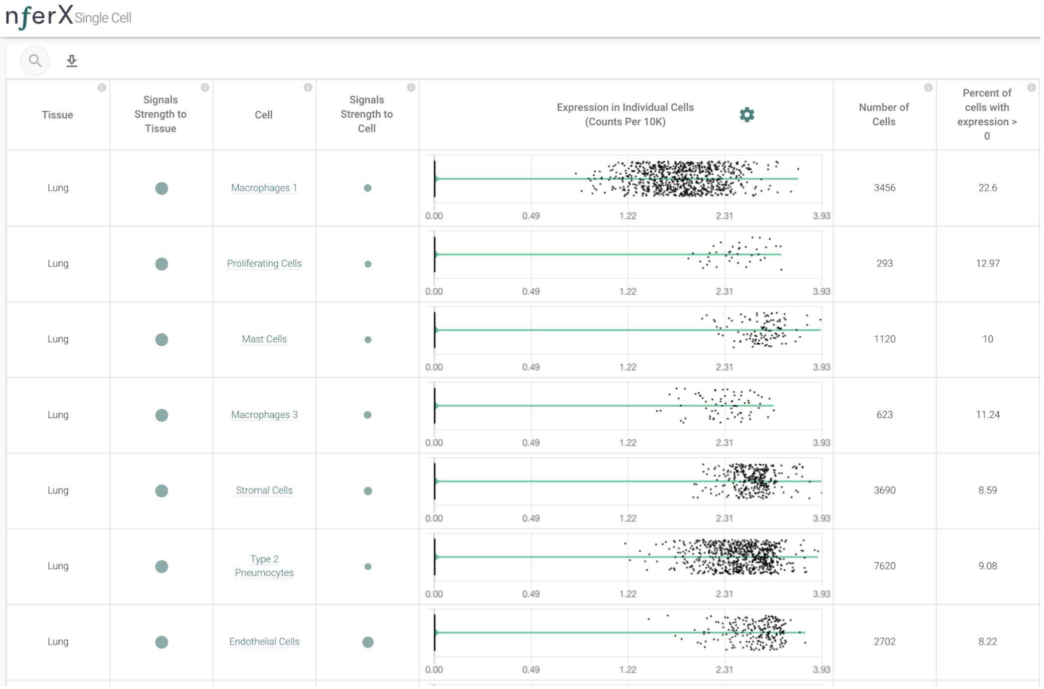
a



Cell Types Enriched in **Highly Expressing Cells**

- Proximal Tubule Cells — Enrichment Score: 323 • Supported by 2 of 2 Studies
- Oligodendrocytes 1 — Enrichment Score: 122.43 • Supported by 2 of 2 Studies
- Macrophages T Cells — Enrichment Score: 48.13 • Supported by 1 of 1 Study
- T Cells — Enrichment Score: 45.32 • Supported by 30 of 33 Studies
- Satellite Cells — Enrichment Score: 38.89 • Supported by 1 of 1 Study
- Doublets — Enrichment Score: 34.5 • Supported by 1 of 1 Study
- Macrophages 2 — Enrichment Score: 33.35 • Supported by 7 of 7 Studies
- Macrophages — Enrichment Score: 29.68 • Supported by 30 of 36 Studies
- Stromal Cells — Enrichment Score: 29.66 • Supported by 6 of 8 Studies
- Type 2 Pneumocytes — Enrichment Score: 27.76 • Supported by 1 of 1 Study

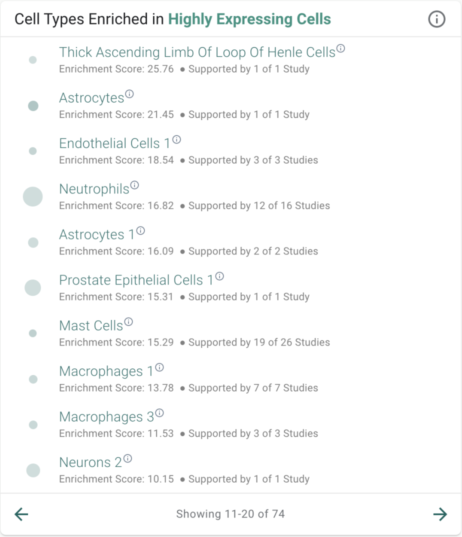
Showing 1-10 of 74



nferX Single Cell

Tissue	Signals Strength to Tissue	Cell	Signals Strength to Cell	Expression in Individual Cells (Counts Per 10K)	Number of Cells	Percent of cells with expression > 0
Lung	●	Macrophages 1	●		3456	22.6
Lung	●	Proliferating Cells	●		293	12.97
Lung	●	Mast Cells	●		1120	10
Lung	●	Macrophages 3	●		623	11.24
Lung	●	Stromal Cells	●		3690	8.59
Lung	●	Type 2 Pneumocytes	●		7620	9.08
Lung	●	Endothelial Cells	●		2702	8.22

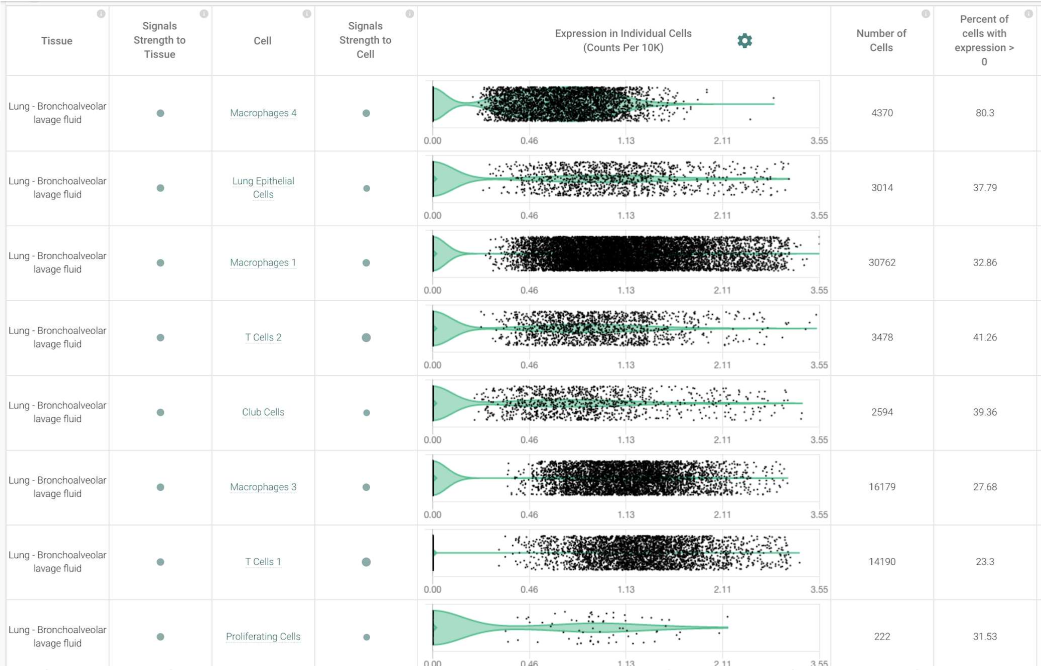
b



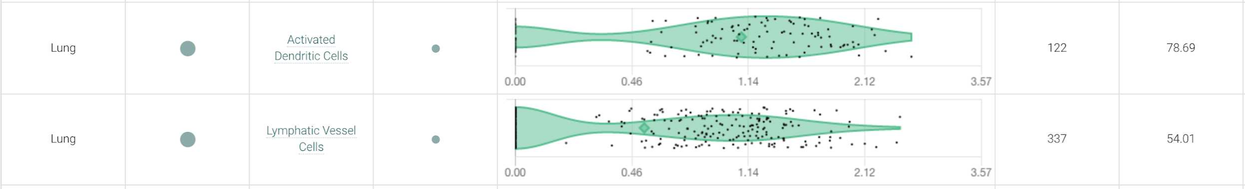
Cell Types Enriched in **Highly Expressing Cells**

- Thick Ascending Limb Of Loop Of Henle Cells — Enrichment Score: 25.76 • Supported by 1 of 1 Study
- Astrocytes — Enrichment Score: 21.45 • Supported by 1 of 1 Study
- Endothelial Cells 1 — Enrichment Score: 18.54 • Supported by 3 of 3 Studies
- Neutrophils — Enrichment Score: 16.82 • Supported by 12 of 16 Studies
- Astrocytes 1 — Enrichment Score: 16.09 • Supported by 2 of 2 Studies
- Prostate Epithelial Cells 1 — Enrichment Score: 15.31 • Supported by 1 of 1 Study
- Mast Cells — Enrichment Score: 15.29 • Supported by 19 of 26 Studies
- Macrophages 1 — Enrichment Score: 13.78 • Supported by 7 of 7 Studies
- Macrophages 3 — Enrichment Score: 11.53 • Supported by 3 of 3 Studies
- Neurons 2 — Enrichment Score: 10.15 • Supported by 1 of 1 Study

Showing 11-20 of 74



Tissue	Signals Strength to Tissue	Cell	Signals Strength to Cell	Expression in Individual Cells (Counts Per 10K)	Number of Cells	Percent of cells with expression > 0
Lung - Bronchoalveolar lavage fluid	●	Macrophages 4	●		4370	80.3
Lung - Bronchoalveolar lavage fluid	●	Lung Epithelial Cells	●		3014	37.79
Lung - Bronchoalveolar lavage fluid	●	Macrophages 1	●		30762	32.86
Lung - Bronchoalveolar lavage fluid	●	T Cells 2	●		3478	41.26
Lung - Bronchoalveolar lavage fluid	●	Club Cells	●		2594	39.36
Lung - Bronchoalveolar lavage fluid	●	Macrophages 3	●		16179	27.68
Lung - Bronchoalveolar lavage fluid	●	T Cells 1	●		14190	23.3
Lung - Bronchoalveolar lavage fluid	●	Proliferating Cells	●		222	31.53

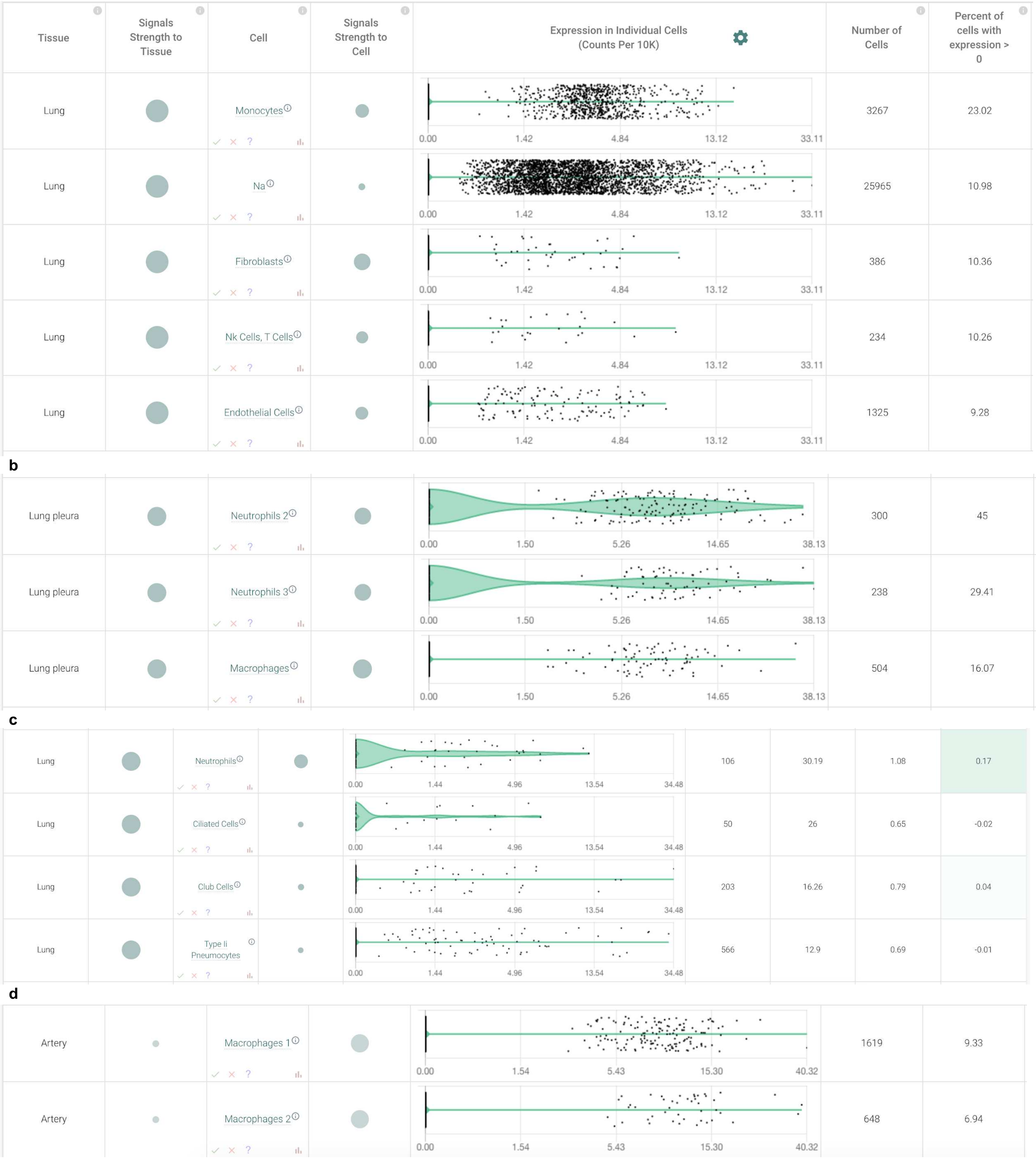


Lung	●	Activated Dendritic Cells	●		122	78.69
Lung	●	Lymphatic Vessel Cells	●		337	54.01

Figure 3

a.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Tissue	Signals Strength to Tissue	Cell	Signals Strength to Cell	Expression in Individual Cells (Counts Per 10K)	Number of Cells	Percent of cells with expression > 0
Lung	●	Monocytes	●	0.00 1.42 4.84 13.12 33.11	3267	23.02
Lung	●	Na	●	0.00 1.42 4.84 13.12 33.11	25965	10.98
Lung	●	Fibroblasts	●	0.00 1.42 4.84 13.12 33.11	386	10.36
Lung	●	Nk Cells, T Cells	●	0.00 1.42 4.84 13.12 33.11	234	10.26
Lung	●	Endothelial Cells	●	0.00 1.42 4.84 13.12 33.11	1325	9.28

b

Lung pleura	●	Neutrophils 2	●	0.00 1.50 5.26 14.65 38.13	300	45
Lung pleura	●	Neutrophils 3	●	0.00 1.50 5.26 14.65 38.13	238	29.41
Lung pleura	●	Macrophages	●	0.00 1.50 5.26 14.65 38.13	504	16.07

c

Lung	●	Neutrophils	●	0.00 1.44 4.96 13.54 34.48	106	30.19	1.08	0.17
Lung	●	Ciliated Cells	●	0.00 1.44 4.96 13.54 34.48	50	26	0.65	-0.02
Lung	●	Club Cells	●	0.00 1.44 4.96 13.54 34.48	203	16.26	0.79	0.04
Lung	●	Type Ii Pneumocytes	●	0.00 1.44 4.96 13.54 34.48	566	12.9	0.69	-0.01

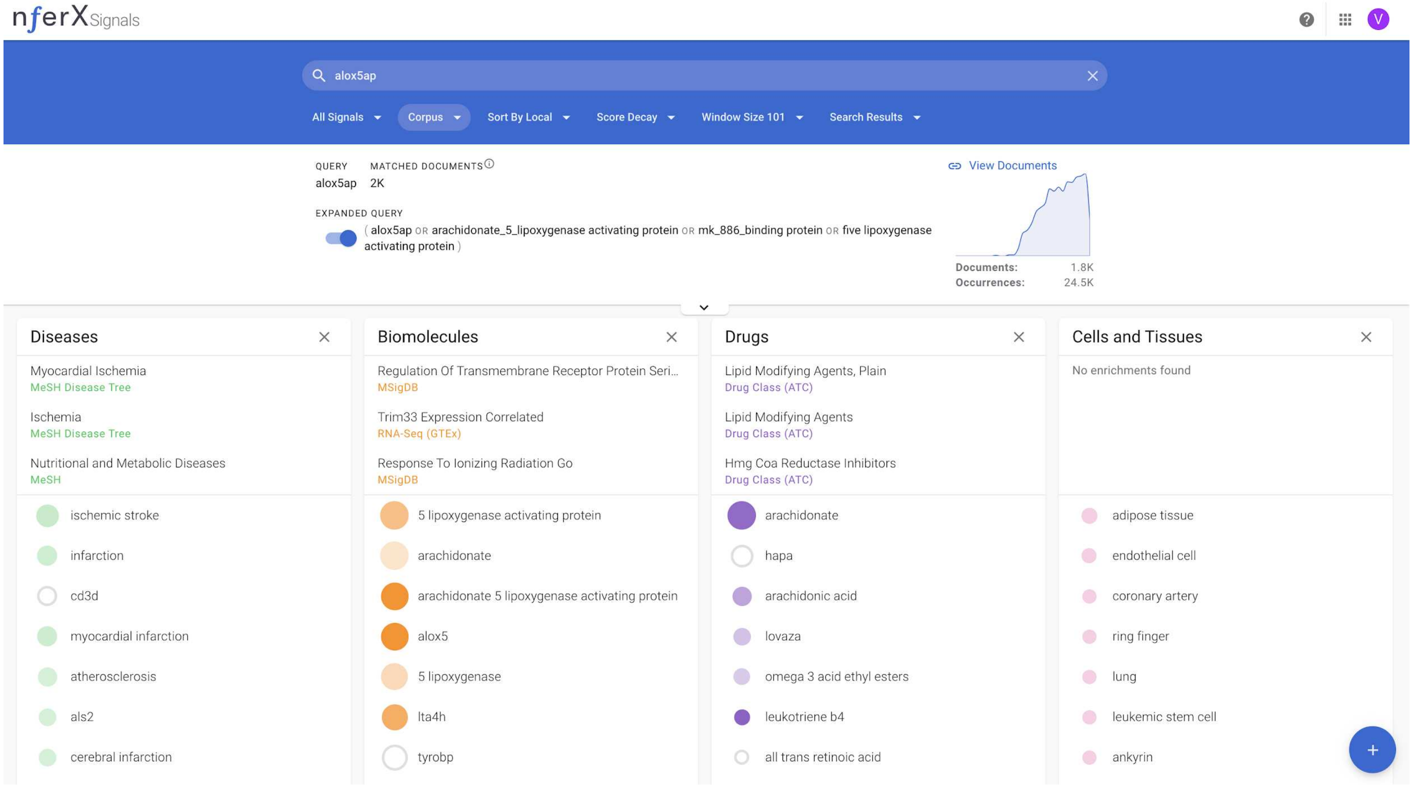
d

Artery	●	Macrophages 1	●	0.00 1.54 5.43 15.30 40.32	1619	9.33
Artery	●	Macrophages 2	●	0.00 1.54 5.43 15.30 40.32	648	6.94

Figure 4

a

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



QUERY: alox5ap
MATCHED DOCUMENTS: 2K

EXPANDED QUERY
(alox5ap OR arachidonate_5_lipoxygenase activating protein OR mk_886_binding protein OR five lipoxygenase activating protein)

Documents: 1.8K
Occurrences: 24.5K

Diseases
Myocardial Ischemia — MeSH Disease Tree
Ischemia — MeSH Disease Tree
Nutritional and Metabolic Diseases — MeSH

- ischemic stroke
- infarction
- cd3d
- myocardial infarction
- atherosclerosis
- als2
- cerebral infarction

Biomolecules
Regulation Of Transmembrane Receptor Protein Seri... — MSigDB
Trim33 Expression Correlated — RNA-Seq (GTEx)
Response To Ionizing Radiation Go — MSigDB

- 5 lipoxygenase activating protein
- arachidonate
- arachidonate 5 lipoxygenase activating protein
- alox5
- 5 lipoxygenase
- lta4h
- tyrobp

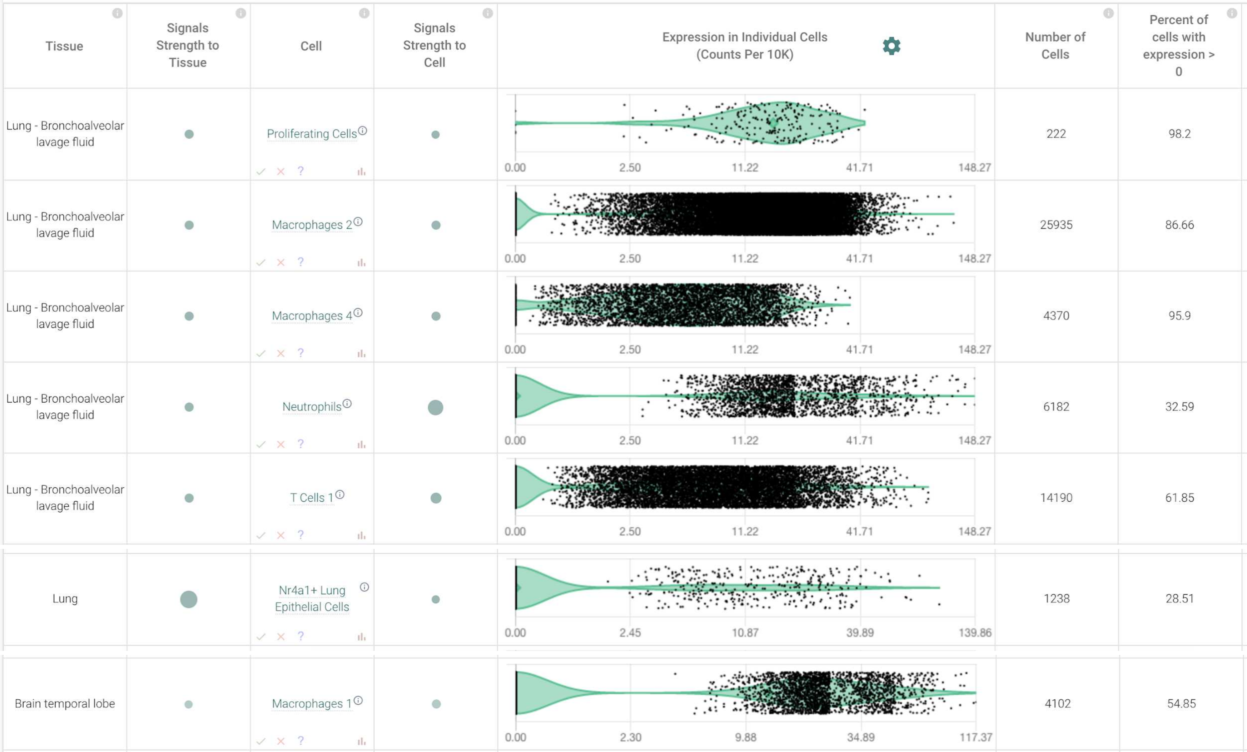
Drugs
Lipid Modifying Agents, Plain — Drug Class (ATC)
Lipid Modifying Agents — Drug Class (ATC)
Hmg Coa Reductase Inhibitors — Drug Class (ATC)

- arachidonate
- hapa
- arachidonic acid
- lovaza
- omega 3 acid ethyl esters
- leukotriene b4
- all trans retinoic acid

Cells and Tissues
No enrichments found

- adipose tissue
- endothelial cell
- coronary artery
- ring finger
- lung
- leukemic stem cell
- ankyrin

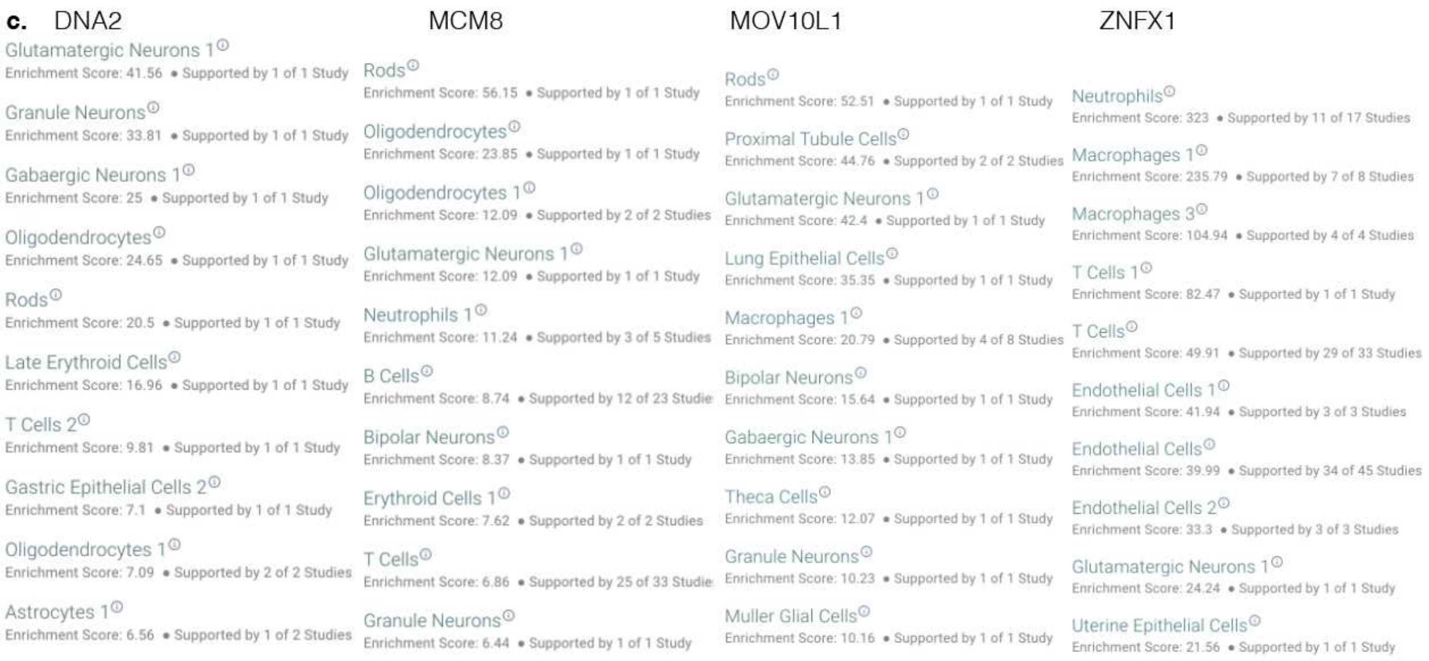
b



Tissue	Signals Strength to Tissue	Cell	Signals Strength to Cell	Expression in Individual Cells (Counts Per 10K)	Number of Cells	Percent of cells with expression > 0
Lung - Bronchoalveolar lavage fluid	●	Proliferating Cells	●	0.00 2.50 11.22 41.71 148.27	222	98.2
Lung - Bronchoalveolar lavage fluid	●	Macrophages 2	●	0.00 2.50 11.22 41.71 148.27	25935	86.66
Lung - Bronchoalveolar lavage fluid	●	Macrophages 4	●	0.00 2.50 11.22 41.71 148.27	4370	95.9
Lung - Bronchoalveolar lavage fluid	●	Neutrophils	●	0.00 2.50 11.22 41.71 148.27	6182	32.59
Lung - Bronchoalveolar lavage fluid	●	T Cells 1	●	0.00 2.50 11.22 41.71 148.27	14190	61.85
Lung	●	Nr4a1+ Lung Epithelial Cells	●	0.00 2.45 10.87 39.89 139.86	1238	28.51
Brain temporal lobe	●	Macrophages 1	●	0.00 2.30 9.88 34.89 117.37	4102	54.85

Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


c.

DNA2

Glutamatergic Neurons 1
Enrichment Score: 41.56 • Supported by 1 of 1 Study

Granule Neurons
Enrichment Score: 33.81 • Supported by 1 of 1 Study

Gabaergic Neurons 1
Enrichment Score: 25 • Supported by 1 of 1 Study

Oligodendrocytes
Enrichment Score: 24.65 • Supported by 1 of 1 Study

Rods
Enrichment Score: 20.5 • Supported by 1 of 1 Study

Late Erythroid Cells
Enrichment Score: 16.96 • Supported by 1 of 1 Study

T Cells 2
Enrichment Score: 9.81 • Supported by 1 of 1 Study

Gastric Epithelial Cells 2
Enrichment Score: 7.1 • Supported by 1 of 1 Study

Oligodendrocytes 1
Enrichment Score: 7.09 • Supported by 2 of 2 Studies

Astrocytes 1
Enrichment Score: 6.56 • Supported by 1 of 2 Studies

MCM8

Rods
Enrichment Score: 56.15 • Supported by 1 of 1 Study

Oligodendrocytes
Enrichment Score: 23.85 • Supported by 1 of 1 Study

Oligodendrocytes 1
Enrichment Score: 12.09 • Supported by 2 of 2 Studies

Glutamatergic Neurons 1
Enrichment Score: 12.09 • Supported by 1 of 1 Study

Neutrophils 1
Enrichment Score: 11.24 • Supported by 3 of 5 Studies

B Cells
Enrichment Score: 8.74 • Supported by 12 of 23 Studie

Bipolar Neurons
Enrichment Score: 8.37 • Supported by 1 of 1 Study

Erythroid Cells 1
Enrichment Score: 7.62 • Supported by 2 of 2 Studies

T Cells
Enrichment Score: 6.86 • Supported by 25 of 33 Studie

Granule Neurons
Enrichment Score: 6.44 • Supported by 1 of 1 Study

MOV10L1

Rods
Enrichment Score: 52.51 • Supported by 1 of 1 Study

Proximal Tubule Cells
Enrichment Score: 44.76 • Supported by 2 of 2 Studies

Glutamatergic Neurons 1
Enrichment Score: 42.4 • Supported by 1 of 1 Study

Lung Epithelial Cells
Enrichment Score: 35.35 • Supported by 1 of 1 Study

Macrophages 1
Enrichment Score: 20.79 • Supported by 4 of 8 Studies

Bipolar Neurons
Enrichment Score: 15.64 • Supported by 1 of 1 Study

Gabaergic Neurons 1
Enrichment Score: 13.85 • Supported by 1 of 1 Study

Theca Cells
Enrichment Score: 12.07 • Supported by 1 of 1 Study

Granule Neurons
Enrichment Score: 10.23 • Supported by 1 of 1 Study

Muller Glial Cells
Enrichment Score: 10.16 • Supported by 1 of 1 Study

ZNFX1

Neutrophils
Enrichment Score: 323 • Supported by 11 of 17 Studies

Macrophages 1
Enrichment Score: 235.79 • Supported by 7 of 8 Studies

Macrophages 3
Enrichment Score: 104.94 • Supported by 4 of 4 Studies

T Cells 1
Enrichment Score: 82.47 • Supported by 1 of 1 Study

T Cells
Enrichment Score: 49.91 • Supported by 29 of 33 Studies

Endothelial Cells 1
Enrichment Score: 41.94 • Supported by 3 of 3 Studies

Endothelial Cells
Enrichment Score: 39.99 • Supported by 34 of 45 Studies

Endothelial Cells 2
Enrichment Score: 33.3 • Supported by 3 of 3 Studies

Glutamatergic Neurons 1
Enrichment Score: 24.24 • Supported by 1 of 1 Study

Uterine Epithelial Cells
Enrichment Score: 21.56 • Supported by 1 of 1 Study

Figure 6

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

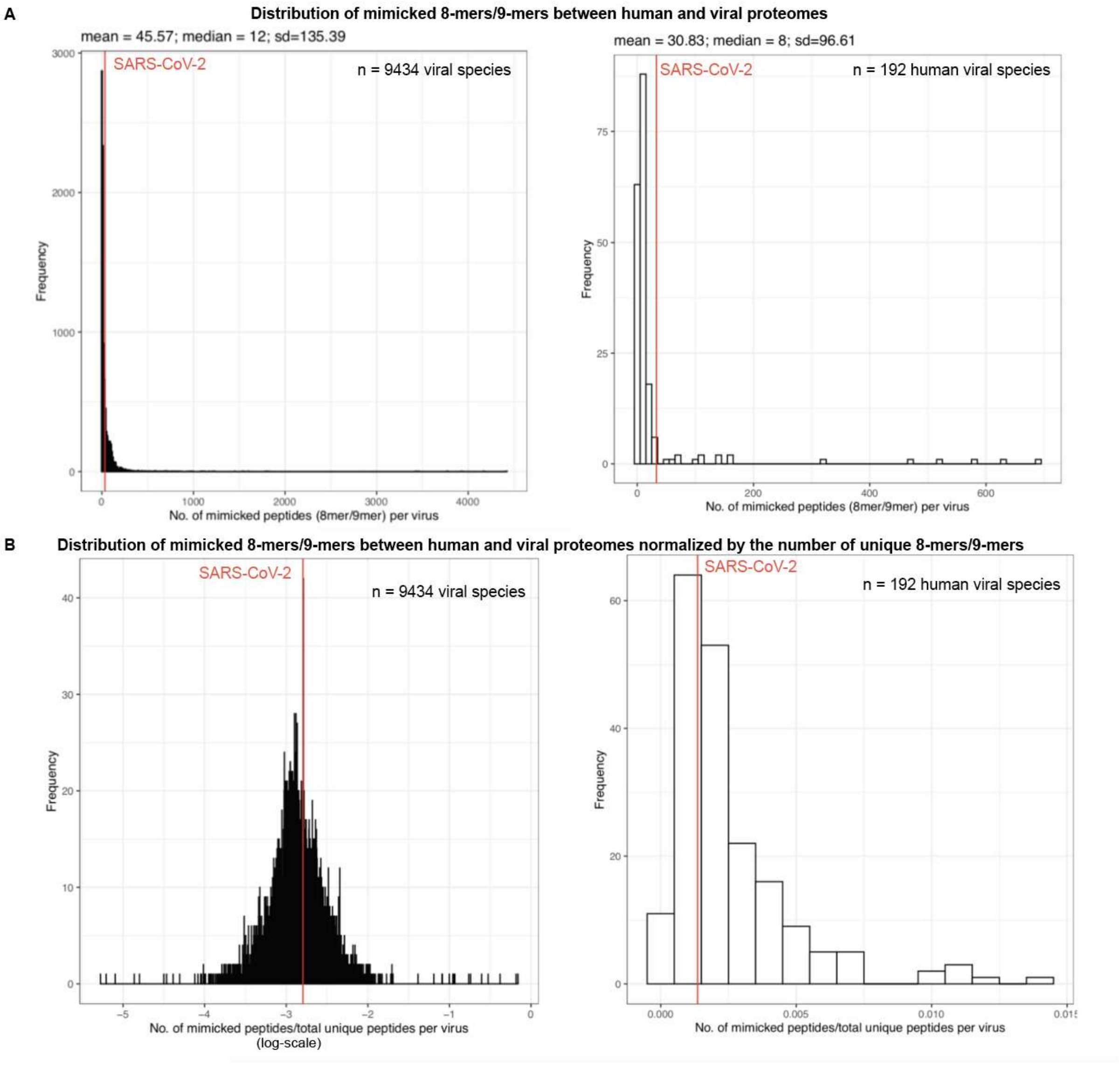


Figure 1 - Supplemental Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. SARS-CoV-2 peptides mimicking human proteins, with experimental evidence of positive MHC binding from the immune epitope database. The viral-human mimicked 8-mer/9-mer peptides are highlighted in green text.

Viral Peptide (Coronavirus)	Viral Protein (NCBI)	Human Epitope	Human Protein	MHC restriction (positive response)	Epitope ID (IEDB)	Pubmed ID (PMID)
(a) SARS-CoV-2 mimicry of human peptides with experimental evidence for MHC-binding (unique to SARS-CoV-2)						
PGSGVPVV (SARS-CoV-2)	**ORF1ab polyprotein** (RNA-dependent RNA polymerase) (YP_009725307.1 :227-234)	KEPGSGVPVVL	**PAM** (P19021: 860-867)	**HLA-B*40:02** D or E acid at peptide position 2 (P2) and M, F, or aliphatic residues at the C terminus. (PMID:24366607)	609309	27920218
ESGLKTIL (SARS-CoV-2)	**ORF1ab polyprotein** (NSP2) (YP_009725298.1 :210-217)	VESGLKTIL	**ANXA7** (P20073: 342-350)	**HLA-B*40:01** **HLA-B*40:02** D or E acid at peptide position 2 (P2) and M, F, or aliphatic residues at the C terminus. (PMID:24366607)	579215	31844290 31530632 27841757
VTLIGEAV (SARS-CoV-2)	**ORF1ab polyprotein (**EndoRNAse) (YP_009725310.1 :165-172)	VPVTLIGEAVF	**PGD** (P52209: 278-285)	**HLA-B*35:01** P at position 2 (p2) and Y at the last position (pΩ) (and to a lesser extend F, M, L, or I) (PMID:26758079)	638710	31844290 29615400 28228285
SLKELLQN (SARS-CoV-2)	**ORF1ab polyprotein** (3C-like Proteinase) (YP_009725301.1 :267-274)	QSLKELLQNW	**CENPI** (Q92674: 496-503)	**HLA-B*57:01** **HLA-B*58:01** **HLA-B*57:03** [A,T,S] at P2; [L,F,W] at P9 (PMID: 30410026)	600524	31844290 30315122 29437277 30410026
(b) SARS-CoV-2 mimicry of other human peptides that are known MHC binders (antigen source: SARS-CoV)						
PEANMDQE (SARS-CoV-2 SARS-CoV)	**ORF1ab polyprotein (**NSP10) (YP_009742617.1)	PEANMDQESF (antigen source: SARS)	**ALOX5AP (Splicing Variant)** (ENSP00000479870 .1: 53-60)	**HLA-B*40:01** D or E acid at peptide position 2 (P2) and M, F, or aliphatic residues at the C terminus. (PMID:24366607	47238	1000425 (RefID)
c) Peptides from coronaviruses that broadly mimic human helicases and are known MHC binders (antigen source: SARS-CoV)						
YNYEPLTQ (SARS-CoV-2; SARS-CoV)	**ORF1ab polyprotein** (3C-like proteinase) (YP_009725301.1 :237-244)	RVYNYEPLTQLK	**MCM8** (inferred) (Q9UJA3: 199-206)	**HLA-A*03:01** common hydrophobic amino acids at P2 and K or R anchor residues at the C-terminus (PMID:7504010,)	624802	31844290 30315122 28228285 26992070
NVAITRAK (SARS-CoV-2; SARS-CoV;	**ORF1ab polyprotein** (Helicase)	RFNVAITRAK (antigen source: SARS)	**DNA2** (inferred) (P51530: 1000-1007)	**HLA-A*03:01**	53748	1000425 (RefID)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Seasonal HCoV)	(YP_009725308.1 :561-568)			common hydrophobic amino acids at P2 and K or R anchor residues at the C-terminus (PMID:7504010) **HLA-A*11:01** (P2-Thr;P9-Lys - PMID:31723204) **HLA-A*68:01** V, I, T, L, Y or F at P2 and K at P9 (PMID: 10449296) **HLA-A*31:01** R at P9 (PMID: 31618895)		
RFNVAITR (SARS-CoV-2; SARS-CoV; MERS; Seasonal HCoV)	**ORF1ab polyprotein** (Helicase) (YP_009725308.1 : 559-566)	RFNVAITRAK (antigen source: SARS)	**MOV10L1** (inferred) (Q9BXT6: 1130-1137)	**HLA-A*03:01** **HLA-A*11:01** (P2-Thr;P9-Lys - PMID:31723204) **HLA-A*68:01** V, I, T, L, Y or F at P2 and K at P9 (PMID: 10449296) **HLA-A*31:01** R at P9 (PMID: 31618895)	53748	1000425 (RefID)
QGPPGTGK (SARS-CoV-2; SARS-CoV; MERS; Seasonal HCoV)	**ORF1ab polyprotein** (Helicase) (YP_009725308.1 : 280-287)	LQGPPGTGK (antigen source: SARS)	**ZNFX1** (inferred) (Q9P2E3: 617-624)	**HLA-A*11:01** (P2-Thr;P9-Lys - PMID:31723204) **HLA-A*03:01** common hydrophobic amino acids at P2 and K or R anchor residues at the C-terminus (PMID:7504010) **HLA-A*31:01** R at P9 (PMID: 31618895)	38844	1000425 (RefID)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Amino acid sequence conservation of the SARS-CoV-2 peptides mimicking human proteins. The PGSGVPVV peptide from the NSP12 protein is present in 46079 out of 46513 SARS-CoV-2 sequences (99.1% conserved; mimics human PAM protein), the ESGLKTIL peptide from the NSP2 protein is present in 44750 out of 46513 SARS-CoV-2 sequences (96.2% conserved; mimics human ANXA7), the VTLIGEAV peptide from the endoRNAase protein is present in 43710 of 46513 SARS-CoV-2 sequences (94% conserved; mimics human PGD); and the SLKELLQN peptide from the 3C-like proteinase is present in 45888 of 46513 SARS-CoV-2 sequences (98.7% conserved; mimics human CENPI). Furthermore, the PGSGVPVV (NSP12 peptide mimicking PAM), ESGLKTIL (NSP2 peptide mimicking ANXA7), VTLIGEAV (endoRNAase peptide mimicking PGD), and SLKELLQN (3C-like proteinase mimicking CENPI) were not found in any of the proteins from seasonal coronavirus strains downloaded from ViPRdb as on 06/15/2020 -- HCoV-229E (756 protein sequences), HCoV-HKU1 (1310 protein sequences), HCoV-NL63 (1462 protein sequences), and HCoV-OC43 (1921 protein sequences). The YNYEPLTQ peptide from the 3C-like proteinase is present in 45927 out of 46513 SARS-CoV-2 sequences (98.7% conserved; mimics human helicase MCM8 protein), the NVAITRAK peptide from the viral helicase is present in 45834 out of 46513 SARS-CoV-2 sequences (98.5% conserved; mimics human helicase DNA2), the RFNVAITR peptide from the viral helicase is present in 45842 of 46513 SARS-CoV-2 sequences (98.6% conserved; mimics human helicase MOV10L1); and the QGPPGTGK peptide from the viral helicase is present in 46150 of 46513 SARS-CoV-2 sequences (99.2% conserved; mimics human ZNFX1). Moreover, NVAITRAK; RFNVAITR and QGPPGTGK peptides were found in 158/319 (49.5%), 161/319 (50.5%) and 69/319 (21.6%) strains of HCoV-OC43 in the NSP10 (NTPase/HEL) protein. QGPPGTGK peptide was also found in 39/236 seasonal HCoV-HKU1 strains in the NSP13 protein. YNYEPLTQ peptide was not found in any of the seasonal human coronavirus strains.

SARS-CoV-2 Mimicked Epitopes	SARS-CoV2 (GISAID)	SARS	MERS	HCoV-229E	HCoV-NL63	HCoV-OC43	HCoV-HKU1
PGSGVPVV	46079/46513 [ORF1ab/NS12 ; 99.06%]	0/659	0/572	0/293	0/478	0/319	0/236
ESGLKTIL	44750/46513 [ORF1ab/NS2; 96.21%]	0/659	0/572	0/293	0/478	0/319	0/236
VTLIGEAV	43710/46513 [ORF1ab/NS15 ; 93.97]	0/659	0/572	0/293	0/478	0/319	0/236
SLKELLQN	45888/46513 [ORF1ab/NS5; 98.66%]	0/659	0/572	0/293	0/478	0/319	0/236
YNYEPLTQ	45927/46513 [ORF1ab/NS5; 98.74%]	0/659	0/572	0/293	0/478	0/319	0/236
NVAITRAK	45834/46513 [ORF1ab/NS13 ; 98.54%]	196/659 [nsp13-pp1ab; 29.74%]	0/572	16/293 [ORF1ab|NSP13; 5.46%]	37/478 [ORF1ab|NSP13; 7.74%]	66/319 [NTPase/HEL; 20.68%]	39/236 [NSP13; 16.52%]
RFNVAITR	45842/46513 [ORF1ab/NS13 ; 98.55%]	196/659 [nsp13-pp1ab; 29.74%]	329/572 [nsp13-pp1ab; 57.51%]	16/293 [ORF1ab|NSP13; 5.46%]	28/478 [ORF1ab;NSP13; 5.85%]	69/319 [NTPase/HEL; 21.63%]	39/236 [NSP13; 16.52%]
QGPPGTGK	46150/46513 [ORF1ab/NS13 ; 99.22%]	177/659 [nsp13-pp1ab; 26.85%]	335/572 [nsp13-pp1ab; 58.56%]	0/293	0/478	69/319 [NTPase/HEL; 21.63]	39/236 [NSP13; 16.52%]

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Smatti, M. K. *et al.* Viruses and Autoimmunity: A Review on the Potential Interaction and Molecular Mechanisms. *Viruses* **11**, (2019).
2. Sayin, İ., Yaşar, K. K. & Yazici, Z. M. Taste and Smell Impairment in COVID-19: An AAO-HNS Anosmia Reporting Tool-Based Comparative Study. *Otolaryngol. Head Neck Surg.* 194599820931820 (2020).
3. Chen, T. *et al.* Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *BMJ* **368**, m1091 (2020).
4. Wagner, T. *et al.* Augmented Curation of Clinical Notes from a Massive EHR System Reveals Symptoms of Impending COVID-19 Diagnosis. doi:10.1101/2020.04.19.20067660.
5. Venkatakrishnan, A. J. *et al.* Knowledge synthesis of 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. *Elife* **9**, (2020).
6. Verdoni, L. *et al.* An outbreak of severe Kawasaki-like disease at the Italian epicentre of the SARS-CoV-2 epidemic: an observational cohort study. *Lancet* **395**, 1771–1778 (2020).
7. Caso, F. *et al.* Could Sars-coronavirus-2 trigger autoimmune and/or autoinflammatory mechanisms in genetically predisposed subjects? *Autoimmun. Rev.* **19**, 102524 (2020).
8. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
9. Trolle, T. *et al.* The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference. *J. Immunol.* **196**, 1480–1487 (2016).
10. Carrillo-Tripp, M. *et al.* VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res.* **37**, D436–42 (2009).
11. Fleri, W. *et al.* The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Front. Immunol.* **8**, 278 (2017).
12. Sidney, J. *et al.* Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Curr. Protoc. Immunol.* **Chapter 18**, Unit 18.3. (2013).
13. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315–326 (2017).
14. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
15. Ramarathinam, S. H. *et al.* Identification of Native and Posttranslationally Modified HLA-B*57:01-Restricted HIV Envelope Derived Epitopes Using Immunoproteomics. *Proteomics* **18**, e1700253 (2018).
16. Lorente, E. *et al.* Substantial Influence of ERAP2 on the HLA-B*40:02 Peptidome: Implications for HLA-B*27-Negative Ankylosing Spondylitis. *Mol. Cell. Proteomics* **18**, 2298–2309 (2019).
17. Pearson, H. *et al.* MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126**, 4690–4701 (2016).
18. Belhadjer, Z. *et al.* Acute heart failure in multisystem inflammatory syndrome in children (MIS-C) in the context of global SARS-CoV-2 pandemic. *Circulation* (2020) doi:10.1161/CIRCULATIONAHA.120.048360.
19. Viner, R. M. & Whittaker, E. Kawasaki-like disease: emerging complication during the COVID-19 pandemic. *The Lancet* vol. 395 1741–1743 (2020).
20. Sidney, J., Botten, J., Neuman, B., Buchmeier, M. & Sette, A. Epitopes Described in-Immune Epitope Database (IEDB). (2006).
21. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405–12 (2015).
22. Getts, D. R., Chastain, E. M. L., Terry, R. L. & Miller, S. D. Virus infection, antiviral immunity, and autoimmunity. *Immunol. Rev.* **255**, 197–209 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

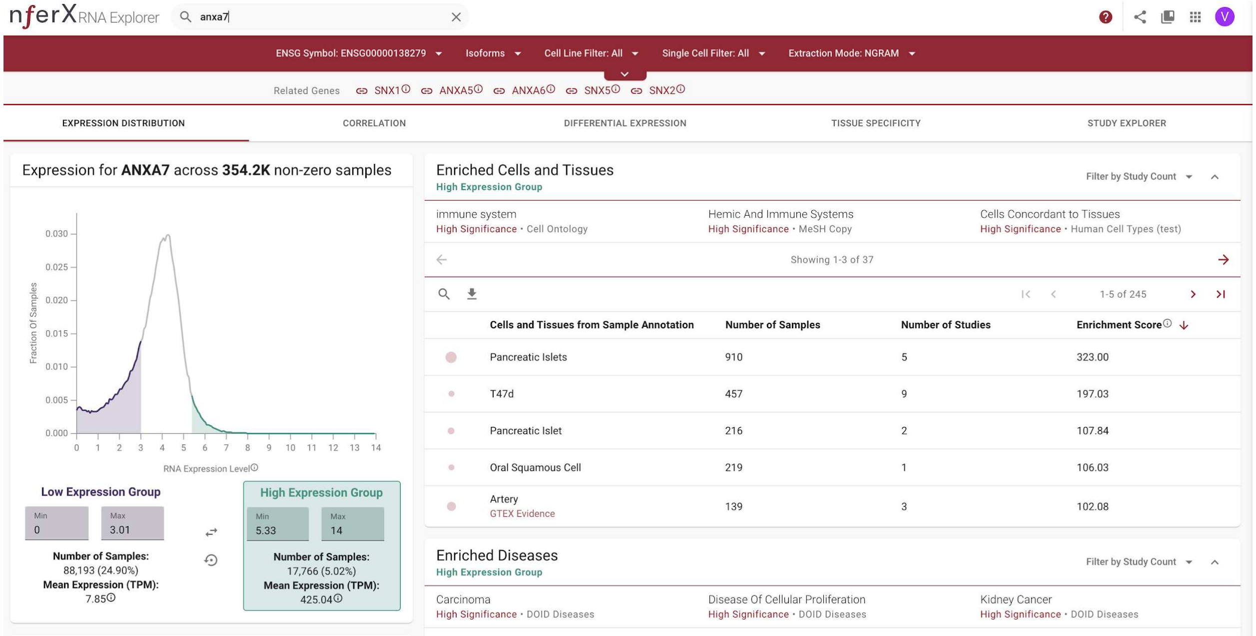
23. Fujinami, R. S., von Herrath, M. G., Christen, U. & Whitton, J. L. Molecular mimicry, bystander activation, or viral persistence: infections and autoimmune disease. *Clin. Microbiol. Rev.* **19**, 80–94 (2006).
24. Carsana, L. *et al.* Pulmonary post-mortem findings in a series of COVID-19 cases from northern Italy: a two-centre descriptive study. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30434-5.
25. Varga, Z. *et al.* Endothelial cell infection and endotheliitis in COVID-19. *Lancet* **395**, 1417–1418 (2020).
26. Ackermann, M. *et al.* Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19. *N. Engl. J. Med.* (2020) doi:10.1056/NEJMoa2015432.
27. Li, X. *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci.Adv.* eabb9153 (2020).
28. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* **30**, 1346–1351.e2 (2020).

Acknowledgements

The authors thank Patrick Lenehan, Travis Hughes, and Murali Aravamudan for their thoughtful feedback.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Material



nferX RNA Explorer anxa7

ENSG Symbol: ENSG00000138279 · Isoforms · Cell Line Filter: All · Single Cell Filter: All · Extraction Mode: NGRAM

Related Genes SNX1 ANXA5 ANXA6 SNX5 SNX2

EXPRESSION DISTRIBUTION · CORRELATION · DIFFERENTIAL EXPRESSION · TISSUE SPECIFICITY · STUDY EXPLORER

Expression for **ANXA7** across **354.2K** non-zero samples

Low Expression Group

Min 0 · Max 3.01

Number of Samples: 88,193 (24.90%)
Mean Expression (TPM): 7.85

High Expression Group

Min 5.33 · Max 14

Number of Samples: 17,766 (5.02%)
Mean Expression (TPM): 425.04

Enriched Cells and Tissues
High Expression Group · Filter by Study Count

immune system
High Significance • Cell Ontology

Hemic And Immune Systems
High Significance • MeSH Copy

Cells Concordant to Tissues
High Significance • Human Cell Types (test)

Showing 1-3 of 37

1-5 of 245

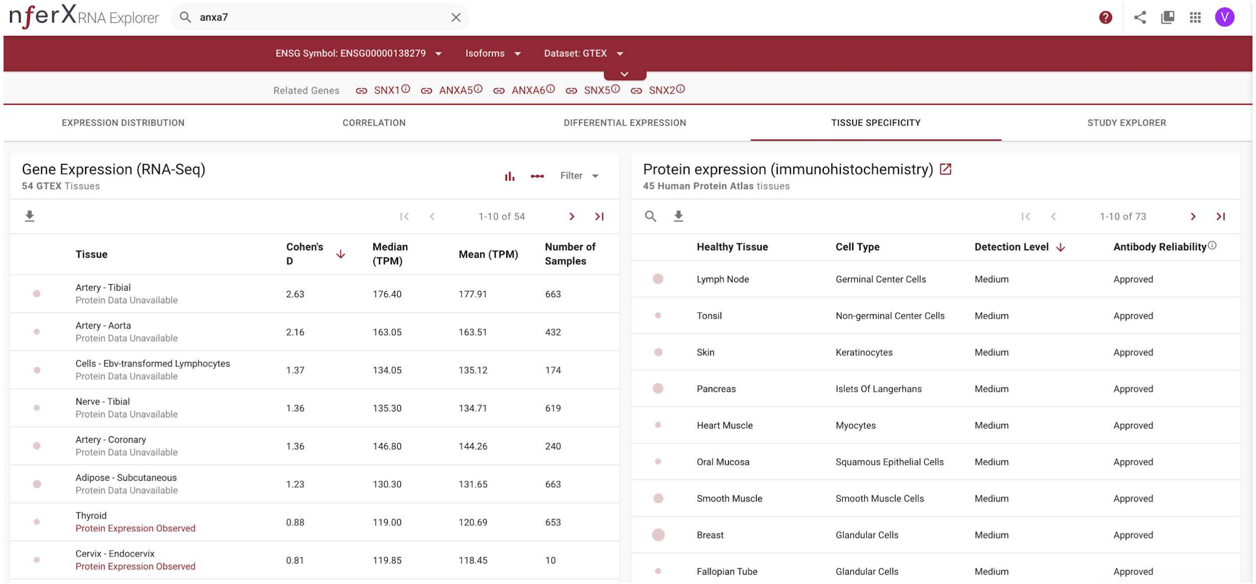
Cells and Tissues from Sample Annotation	Number of Samples	Number of Studies	Enrichment Score
Pancreatic Islets	910	5	323.00
T47d	457	9	197.03
Pancreatic Islet	216	2	107.84
Oral Squamous Cell	219	1	106.03
Artery GTEX Evidence	139	3	102.08

Enriched Diseases
High Expression Group · Filter by Study Count

Carcinoma
High Significance • DOID Diseases

Disease Of Cellular Proliferation
High Significance • DOID Diseases

Kidney Cancer
High Significance • DOID Diseases



nferX RNA Explorer anxa7

ENSG Symbol: ENSG00000138279 · Isoforms · Dataset: GTEX

Related Genes SNX1 ANXA5 ANXA6 SNX5 SNX2

EXPRESSION DISTRIBUTION · CORRELATION · DIFFERENTIAL EXPRESSION · TISSUE SPECIFICITY · STUDY EXPLORER

Gene Expression (RNA-Seq)
54 GTEX Tissues · Filter

1-10 of 54

Tissue	Cohen's D	Median (TPM)	Mean (TPM)	Number of Samples
Artery - Tibial Protein Data Unavailable	2.63	176.40	177.91	663
Artery - Aorta Protein Data Unavailable	2.16	163.05	163.51	432
Cells - Ebv-transformed Lymphocytes Protein Data Unavailable	1.37	134.05	135.12	174
Nerve - Tibial Protein Data Unavailable	1.36	135.30	134.71	619
Artery - Coronary Protein Data Unavailable	1.36	146.80	144.26	240
Adipose - Subcutaneous Protein Data Unavailable	1.23	130.30	131.65	663
Thyroid Protein Expression Observed	0.88	119.00	120.69	653
Cervix - Endocervix Protein Expression Observed	0.81	119.85	118.45	10

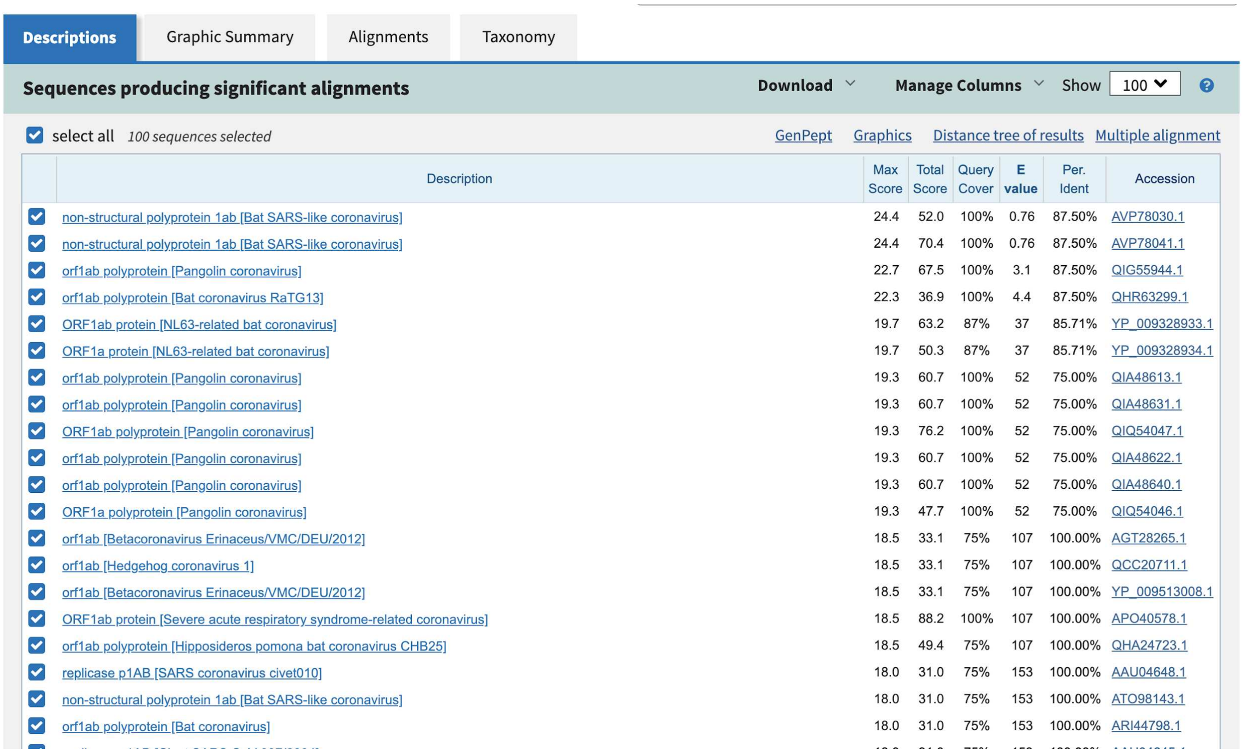
Protein expression (immunohistochemistry)
45 Human Protein Atlas tissues

1-10 of 73

Healthy Tissue	Cell Type	Detection Level	Antibody Reliability
Lymph Node	Germinal Center Cells	Medium	Approved
Tonsil	Non-germinal Center Cells	Medium	Approved
Skin	Keratinocytes	Medium	Approved
Pancreas	Islets Of Langerhans	Medium	Approved
Heart Muscle	Myocytes	Medium	Approved
Oral Mucosa	Squamous Epithelial Cells	Medium	Approved
Smooth Muscle	Smooth Muscle Cells	Medium	Approved
Breast	Glandular Cells	Medium	Approved
Fallopian Tube	Glandular Cells	Medium	Approved

Figure 3 - Supplementary Figure1. Single cell expression of ANXA7

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	non-structural polyprotein 1ab [Bat SARS-like coronavirus]	24.4	52.0	100%	0.76	87.50%	AVP78030.1
☑	non-structural polyprotein 1ab [Bat SARS-like coronavirus]	24.4	70.4	100%	0.76	87.50%	AVP78041.1
☑	orf1ab polyprotein [Pangolin coronavirus]	22.7	67.5	100%	3.1	87.50%	QIG55944.1
☑	orf1ab polyprotein [Bat coronavirus RaTG13]	22.3	36.9	100%	4.4	87.50%	QHR63299.1
☑	ORF1ab protein [NL63-related bat coronavirus]	19.7	63.2	87%	37	85.71%	YP_009328933.1
☑	ORF1a protein [NL63-related bat coronavirus]	19.7	50.3	87%	37	85.71%	YP_009328934.1
☑	orf1ab polyprotein [Pangolin coronavirus]	19.3	60.7	100%	52	75.00%	QIA48613.1
☑	orf1ab polyprotein [Pangolin coronavirus]	19.3	60.7	100%	52	75.00%	QIA48631.1
☑	ORF1ab polyprotein [Pangolin coronavirus]	19.3	76.2	100%	52	75.00%	QIQ54047.1
☑	orf1ab polyprotein [Pangolin coronavirus]	19.3	60.7	100%	52	75.00%	QIA48622.1
☑	orf1ab polyprotein [Pangolin coronavirus]	19.3	60.7	100%	52	75.00%	QIA48640.1
☑	ORF1a polyprotein [Pangolin coronavirus]	19.3	47.7	100%	52	75.00%	QIQ54046.1
☑	orf1ab [Betacoronavirus Erinaceus/VMC/DEU/2012]	18.5	33.1	75%	107	100.00%	AGT28265.1
☑	orf1ab [Hedgehog coronavirus 1]	18.5	33.1	75%	107	100.00%	QCC20711.1
☑	orf1ab [Betacoronavirus Erinaceus/VMC/DEU/2012]	18.5	33.1	75%	107	100.00%	YP_009513008.1
☑	ORF1ab protein [Severe acute respiratory syndrome-related coronavirus]	18.5	88.2	100%	107	100.00%	APO40578.1
☑	orf1ab polyprotein [Hipposideros pomona bat coronavirus CHB25]	18.5	49.4	75%	107	100.00%	QHA24723.1
☑	replicase p1AB [SARS coronavirus civet010]	18.0	31.0	75%	153	100.00%	AAU04648.1
☑	non-structural polyprotein 1ab [Bat SARS-like coronavirus]	18.0	31.0	75%	153	100.00%	ATO98143.1
☑	orf1ab polyprotein [Bat coronavirus]	18.0	31.0	75%	153	100.00%	ARI44798.1

Figure 3 - Supplementary Figure2. Human ANXA7 mimicking peptide ESGLKTIL is only present in SARS-CoV-2, with the closest known evolutionary homologs being from BAT SARS-like coronavirus (ESGLKTIL), pangolin coronavirus, and the NL63-related bat coronavirus strains.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S1. Reference SARS-CoV-2 proteome from UniProt

SARS-CoV-2 Protein Name	SARS-CoV-2 Gene Symbol	UniProt ID (Length)
R1AB_SARS2 Replicase polyprotein 1ab	rep	P0DTD1 (7096)
SPIKE_SARS2 Spike glycoprotein	S	P0DTC2 (1273)
R1A_SARS2 Replicase polyprotein	rep	P0DTC1 (4405)
NS7A_SARS2 Protein 7a	7a	P0DTC7 (121)
AP3A_SARS2 Protein 3a	3a	P0DTC3 (275)
VME1_SARS2 Membrane protein	3	P0DTC5 (222)
NCAP_SARS2 Nucleoprotein	N	P0DTC9 (419)
ORF9B_SARS2 Protein 9b	3	P0DTD2 (97)
VEMP_SARS2 Envelope small membrane protein	E	P0DTC4 (75)
NS6_SARS2 Non-structural protein 6	6	P0DTC6 (61)
NS8_SARS2 Non-structural protein 8	8	P0DTC8 (121)
NS7B_SARS2 Protein non-structural 7b	7b	P0DTD8 (43)
Y14_SARS2 Uncharacterized protein 14	ORF14	P0DTD3 (73)
A0A663DJA2_SARS2 ORF10	ORF10	A0A663DJA2 (38)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S2: All 33 peptides that are shared between SARS-CoV-2 and the human proteome

Peptide	Viral Protein	Accession No.	Start	End	Human Protein	Description	Uniprot ID	Start	End
AKKNNLPF	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	2732	2739	LPGAT1	Lysophosphatidylglycerol Acyltransferase 1	Q92604	198	205
DEDDSEPV	SPIKE_SARS2	P0DTC2	1256	1263	MYO16	Myosin XVI	Q9Y6X6	1403	1410
DEDEEEGD	R1AB_SARS2 Replicase polyprotein 1ab	PODTD1	927	934	GMCL1	Germ Cell-Less 1, Spermatogenesis Associated	Q96IK5	68	75
DIQLLKSA	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	1126	1133	EML1	EMAP Like 1	O00423	50	57
DTSLSGFK	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	3670	3677	SLC12A7	Solute Carrier Family 12 Member 7	Q9Y666	994	1001
ELPDEFVV	ORF9B_SARS2 Protein 9b	P0DTD2	85	92	MROH2B	Maestro Heat Like Repeat Family Member 2B	Q7Z745	102	109
ESGLKTIL	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	389	396	ANXA7	Annexin A7	P20073	403	410
EVEKGVLP	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	54	61	NDST1	N-Deacetylase And N-Sulfotransferase 1	P52848	213	220
GPPGTGKS	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	5605	5612	SETX	Senataxin	Q7Z333	1962	1969
					VPS4A	Vacuolar Protein Sorting 4 Homolog A	Q9UN37	166	173
					VPS4B	Vacuolar Protein Sorting 4 Homolog B	O75351	173	180
KDKKKKAD	NCAP_SARS2 Nucleoprotein	P0DTC9	369	376	MICAL3	Microtubule Associated Monooxygenase, Calponin And LIM Domain Containing 3	Q7RTP6	1748	1755
KKDKKKKA	NCAP_SARS2 Nucleoprotein	P0DTC9	368	375	MICAL3	Microtubule Associated Monooxygenase, Calponin And LIM Domain Containing 3	Q7RTP6	1747	1754
KKDKKKKAD	NCAP_SARS2 Nucleoprotein	P0DTC9	368	376	MICAL3	Microtubule Associated Monooxygenase, Calponin And LIM Domain Containing 3	Q7RTP6	1747	1755
LALITLAT	NS7A_SARS2 Protein 7a	P0DTC7	6	13	HTR1B	5-Hydroxytryptamine Receptor 1B	P28222	55	62
LVDPQIQL	ORF9B_SARS2 Protein 9b	P0DTD2	13	20	VARS2	Valyl-TRNA Synthetase 2, Mitochondrial	Q5ST30	988	995
NVAITRAK	R1AB_SARS2 Replicase polyprotein 1ab	PODTD1	5885	5892	DNA2	DNA Replication Helicase/Nuclease 2	P51530	1000	1007
PDEDEEEG	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	926	933	CC2D1A	Coiled-Coil And C2 Domain Containing 1A	Q6P1N0	83	90
PGSGVPVV	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	4618	4625	PAM	Peptidylglycine Alpha-Amidating Monooxygenase	P19021	859	866
QGPPGTGK	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	5604	5611	HELZ2	Helicase With Zinc Finger 2	Q9BYK8	2172	2179
					UPF1	UPF1 RNA Helicase And ATPase	Q92900	501	508

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

					ZNFX1	Zinc Finger NFX1-Type Containing 1	Q9P2E3	617	624
RFNVAITR	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	5883	5890	MOV10L1	Mov10 Like RISC Complex RNA Helicase 1	Q9BXT6	1130	1137
RRARSVAS	SPIKE_SARS2 Spike glycoprotein	P0DTC2	681	688	SCNN1A	Sodium Channel Epithelial 1 Subunit Alpha	P37088	200	207
RRSFYVYA	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	2430	2437	TPRA1	Transmembrane Protein Adipocyte Associated 1	Q86W33	224	231
RYPANSIV	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	6315	6322	BRI3	Brain Protein I3	O95415	65	72
SLKELLQN	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	3529	3536	CENPI	Centromere Protein I	Q92674	495	502
SRSSSRSR	NCAP_SARS2 Nucleoprotein	P0DTC9	183	190	CCNL2	Cyclin L2	Q96S94	462	469
					CLASRP	CLK4 Associating Serine/Arginine Rich Protein	Q8N2M8	395	402
					LUC7L2	LUC7 Like 2, Pre-MRNA Splicing Factor	Q9Y383	304	3011
SSRSSSRS	NCAP_SARS2 Nucleoprotein	P0DTC9	182	189	CLASRP	CLK4 Associating Serine/Arginine Rich Protein	Q8N2M8	390	397
					LUC7L2	LUC7 Like 2, Pre-MRNA Splicing Factor	Q9Y383	304	311
SSRSSSRSR	NCAP_SARS2 Nucleoprotein	P0DTC9	182	190	CLASRP	CLK4 Associating Serine/Arginine Rich Protein	Q8N2M8	394	402
					PGD	Phosphogluconate Dehydrogenase	P52209	277	284
VNSVLLFL	VEMP_SARS2 Envelope small membrane protein	P0DTC4	13	20	RANBP6	RAN Binding Protein 6	O60518	408	415
VTLIGEAV	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	6616	6623	PGD	Phosphogluconate Dehydrogenase	P52209	277	284
YNYEPLTQ	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	3499	3506	MCM8	Minichromosome Maintenance 8 Homologous Recombination Repair Factor	Q9UJA3	198	205
EVLLAPLL	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	1141	1148	ARL6IP4	ADP Ribosylation Factor Like GTPase 6 Interacting Protein 4	ENSP000004389 69.1	175	182
PEANMDQE	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	4311	4318	ALOX5AP	Arachidonate 5-Lipoxygenase Activating Protein	ENSP000004798 70.1	53	60
GGSCVLSG	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	1099	1106	SNX27	Sorting Nexin 27	ENSP000004967 75.1	111	118
REETGLLM	R1AB_SARS2 Replicase polyprotein 1ab	P0DTD1	723	730	ESRRG	Estrogen Related Receptor Gamma	ENSP000004663 43.1	723	730

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S3. Distinctive peptides from SARS-CoV-2, not present in previously sequenced human coronavirus strains, that do mimic human proteins. There is no compelling positive T-cell immune response against either the human or viral proteins to warrant further discussion in the current study, but these will be the topic of follow-up experimental studies into SARS-CoV-2-based immunologic modulation in humans.

SARS-CoV-2 Peptide	**SARS-CoV-2 protein**	**Mimicked Human Protein**
AKKNNLPF (SARS-CoV-2)	NSP3 (YP_009725299.1 : 1914-1921; P0DTD1:2732-2739)	**LPGAT1** (Q92604:198-205)
DEDEEEGD (SARS-CoV-2)	NSP3 (YP_009725299.1 : 109-116; P0DTD1: 927-934)	**GMCL1** (Q96IK5: 68-75)
DIQLLKSA (SARS-CoV-2)	NSP3 (YP_009725299.1 : 50-57; P0DTD1:1126-1133)	**EML1** (O00423: 50-57)
DTSLSGFK (SARS-CoV-2)	NSP6 (YP_009725302.1 :101-108; P0DTD1: 3670-3677)	**SLC12A7** (Q9Y666:994-1001)
EVEKGVLP (SARS-CoV-2)	Leader protein (YP_009725297.1 :54-61; P0DTD1:54-61)	**NDST1** (P52848:213-220)
KKDKKKKAD (SARS-CoV-2)	Nucleocapsid phosphoprotein (YP_009724397.2 : 368-37; P0DTC9:368-376)	**MICAL3** (Q7RTP6: 368-376**)**
LALITLAT (SARS-CoV-2)	ORF7a protein (YP_009724395.1 : 6-13; P0DTC7:6-13)	**HTR1B** (P28222: 55-62)
PDEDEEEG (SARS-CoV-2)	NSP3 (YP_009725299.1 :108-115; P0DTD1:926-933)	**CC2D1A** (Q6P1N0:83-90)
RRARSVAS (SARS-CoV-2)	Surface glycoprotein (YP_009724390.1 :681-688; P0DTC2:681-688)	**SCNN1A** (P37088: 200-207)
RRSFYVYA (SARS-CoV-2)	NSP3 (YP_009725299.1 :1612-1619; P0DTD1:2430-2437)	**TPRA1** (Q86W33:224-231)
RYPANSIV (SARS-CoV-2)	3'-to-5' exonuclease (YP_009725309.1 : 390-397; P0DTD1:6315-6322)	**BRI3** (O95415: 65-72)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S4. Reference SARS-CoV proteome from UniProt

SARS-CoV Protein Name	SARS-CoV Gene Symbol	UniProt ID
R1AB_CVHSA Replicase polyprotein 1ab	rep	P0C6X7 (7073)
SPIKE_CVHSA Spike glycoprotein	S	P59594 (1255)
R1A_CVHSA Replicase polyprotein	rep	P0C6U8 (4382)
NS7A_CVHSA Protein 7a	7a	P59635 (122)
AP3A_CVHSA Protein 3a	3a	P59632 (274)
VME1_CVHSA Membrane protein	3	P59596 (221)
NCAP_CVHSA Nucleoprotein	N	P59595 (422)
NS3B_CVHSA Non-structural protein 3b	3b	P59633 (154)
ORF9B_CVHSA Protein 9b	3	P59636 (98)
VEMP_CVHSA Envelope small membrane protein	E	P59637 (76)
NS6_CVHSA Non-structural protein 6	6	P59634 (63)
NS8B_CVHSA Non-structural protein 8b	8b	Q80H93 (84)
NS8A_CVHSA Non-structural protein 8	8a	Q7TFA0 (39)
NS7B_CVHSA Protein non-structural 7b	7b	Q7TFA1 (44)
Y14_CVHSA Uncharacterized protein 14	ORF14	Q7TLC7 (70)

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S5. Seasonal human coronavirus (HCoV) peptide mimicry of human proteins with experimental evidence of positive T-cell assays with specific MHC restriction. The MHC-TCR-peptide assays conducted include: (Assay 2.1) Cellular MHC/mass spectrometry, ligand presentation {ref}, (Assay 2.2)

Viral Peptide	Viral Protein (HCoV strain: Uniprot)	Human Epitope	Human Protein	MHC restriction (Assay - T-cell stimulation)	Epitope ID (IEDB)	Pubmed ID (PMID)
GPPGTGKS (HKU1;OC43)	Nsp13 (**HCoV-HKU1** - YP_459942.1 :280-287) ; Nsp10 (**HCoV-OC43** - YP_009555254.1:280-287)	GPPGTGKSYLAKAVATEAN	SKD1 (167-185)	HLA-DRA*01:01 HLA-DRB1*08:01 **(Assay 2.1 - Positive)**	433968	21654843
GRIVTLIS (HKU1)	Nsp6 (YP_460019.1: 142-149)	GRIVTLISF	MCL-1 (262-270)	HLA-B*27:05 HLA-B*27:09 HLA-B*27:04 HLA-B*27:01 HLA-B*27:02 HLA-B*27:06 HLA-B*27:07 HLA-B*27:08 **(Assay 2.1 - Positive)**	241225	31844290 31154438 29632046 29393594 28188227 28063628 27920218 26811146 27846572 26992070 26929215 25469497 26154972 25418920 25645385 20112406
SLLRTSIM (NL63)	Replicase polyprotine 1ab (HCoV-NL63 YP_003766.2 :1920-1927)	SLLRTSIMSK	CCT8 (162-171)	HLA class I cellular MHC/mass spectrometry ligand presentation **Positive**	625570	26992070
TCNSKLTL (OC43)	Spike glycoprotein (YP_009555241.1: 254-261)	STCNSKLTLK	LIM (H0Y592:112-121)	HLA class I mass spectrometry ligand presentation **Positive**	884003	30429286
VVGSTEEVK (229E)	Replicase polyprotine 1ab (NP_073549.1: 524-532)	HLPFAVVGSTEEVKIGNK	SEPTIN11 (Q9NVA2:241-258)	HLA-B*27:05	799404	29393594

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S6. SARS-CoV peptide mimicry of human proteins with experimental evidence of positive T-cell assays with specific MHC restriction. The MHC-TCR-peptide assays conducted include: (Assay 2.1) Cellular MHC/mass spectrometry, ligand presentation {ref}, (Assay 2.2)

Viral Peptide	Viral Protein (SARS-CoV: Uniprot)	Human Epitope	Human Protein	MHC restriction (Assay - T-cell stimulation)	Epitope ID (IEDB)	Pubmed ID (PMID)
GPPGTGKS (SARS-CoV; MERS; HCoV-OC43; HCoV-HKU1)	Nsp13 (NP_828870.1 :281-288)	GPPGTGKSYLAKAVATEAN	**VPS4A** (Q9UN37 :167-185)	HLA-DRA*01:01 HLA-DRB1*08:01 **(Assay 2.1 - Positive)**	433968	21654843
YNYEPLTQ (SARS-CoV; SARS-CoV-2)	Nsp5 (NP_828863.1 :236-243)	RVYNYEPLTQLK	**MCM8** (Q9UJA3: 197-208)	HLA-A*03:01 cellular MHC/mass spectrometry ligand presentation **Positive**	624802	31844290 30315122 28228285 26992070

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.19.161620; this version posted August 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S7. MERS peptide mimicry of human proteins with experimental evidence of positive T-cell assays with specific MHC restriction. The MHC-TCR-peptide assays conducted include: (Assay 2.1) Cellular MHC/mass spectrometry, ligand presentation {ref}, (Assay 2.2)

Viral Peptide	Viral Protein (MERS-CoV: Uniprot)	Human Epitope	Human Protein	MHC restriction (Assay - T-cell stimulation)	Epitope ID (IEDB)	Pubmed ID (PMID)
DGKPISAY (MERS-CoV)	Nsp2 (YP_009047214.1:12-19)	LENFYPLEGGRVLLDGKPISAYD	ABCB9 (Q9NP78:552-574)	HLA-A*30:02 (mass spectrometry ligand presentation **Positive**)	1028898	31844290
GPPGTGKS (SARS-CoV-2; SARS-CoV; MERS-CoV; HCoV-OC43; HCoV-HKU1)	Nsp13 (YP_009047224.1: 281-288)	GPPGTGKSYLAKAVATEAN	VPS4A (Q9UN37:167-185)	HLA-DRA*01:01 HLA-DRB1*08:01 **(Assay 2.1 - Positive)**	433968	21654843
LLGSIAGV	Spike glycoprotein (YP_009047204.1: 950:957)	GLLGSIAGV	CLCF1 (Q9UBD9: 142-150)	HLA-Class I cellular MHC/mass spectrometry ligand presentation **Positive**	923359	31222486 31154438