Venkatakrishnan et al. *Cell Death Discovery* (2020)6:96
https://doi.org/10.1038/s41420-020-00321-y

Cell Death Discovery

ARTICLE Open Access

Benchmarking evolutionary tinkering underlying human–viral molecular mimicry shows multiple host pulmonary–arterial peptides mimicked by SARS-CoV-2

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

Abstract

The hand of molecular mimicry in shaping SARS-CoV-2 evolution and immune evasion remains to be deciphered. Here, we report 33 distinct 8-mer/9-mer peptides that are identical between SARS-CoV-2 and the human reference proteome. We benchmark this observation against other viral–human 8-mer/9-mer peptide identity, which suggests generally similar extents of molecular mimicry for SARS-CoV-2 and many other human viruses. Interestingly, 20 novel human peptides mimicked by SARS-CoV-2 have not been observed in any previous coronavirus strains (HCoV, SARS-CoV, and MERS). Furthermore, four of the human 8-mer/9-mer peptides mimicked by SARS-CoV-2 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 single-cell RNA-seq (scRNA-seq) analysis that shows the targeted genes significantly expressed in human lungs and arteries; tissues implicated in COVID-19 pathogenesis. Finally, HLA-A*03 restricted 8-mer peptides are found to be shared broadly by human and coronaviridae helicases in functional hotspots, with potential implications for nucleic acid unwinding upon initial infection. This study presents the first scan of human peptide mimicry by SARS-CoV-2, and via its benchmarking against human–viral mimicry more broadly, presents a computational framework for follow-up studies to assay how evolutionary tinkering may relate to zoonosis and herd immunity.

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 nonstructural proteins (e.g., RNA-dependent RNA polymerase), as depicted in Fig. 1a. The nonstructural ORF1ab polyprotein undergoes proteolytic processing to give rise to the following 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 (Fig. 1b)[8]. Here, we investigate the potential for molecular mimicry in the context of immune surveillance and host-antigen recognition in

Correspondence: Venky Soundararajan (venky@nference.net)
[1]nference, Cambridge, MA, USA
[2]nference Labs, Bangalore, India
Full list of author information is available at the end of the article
Edited by I. Amelio

SPRINGER NATURE CDDpress

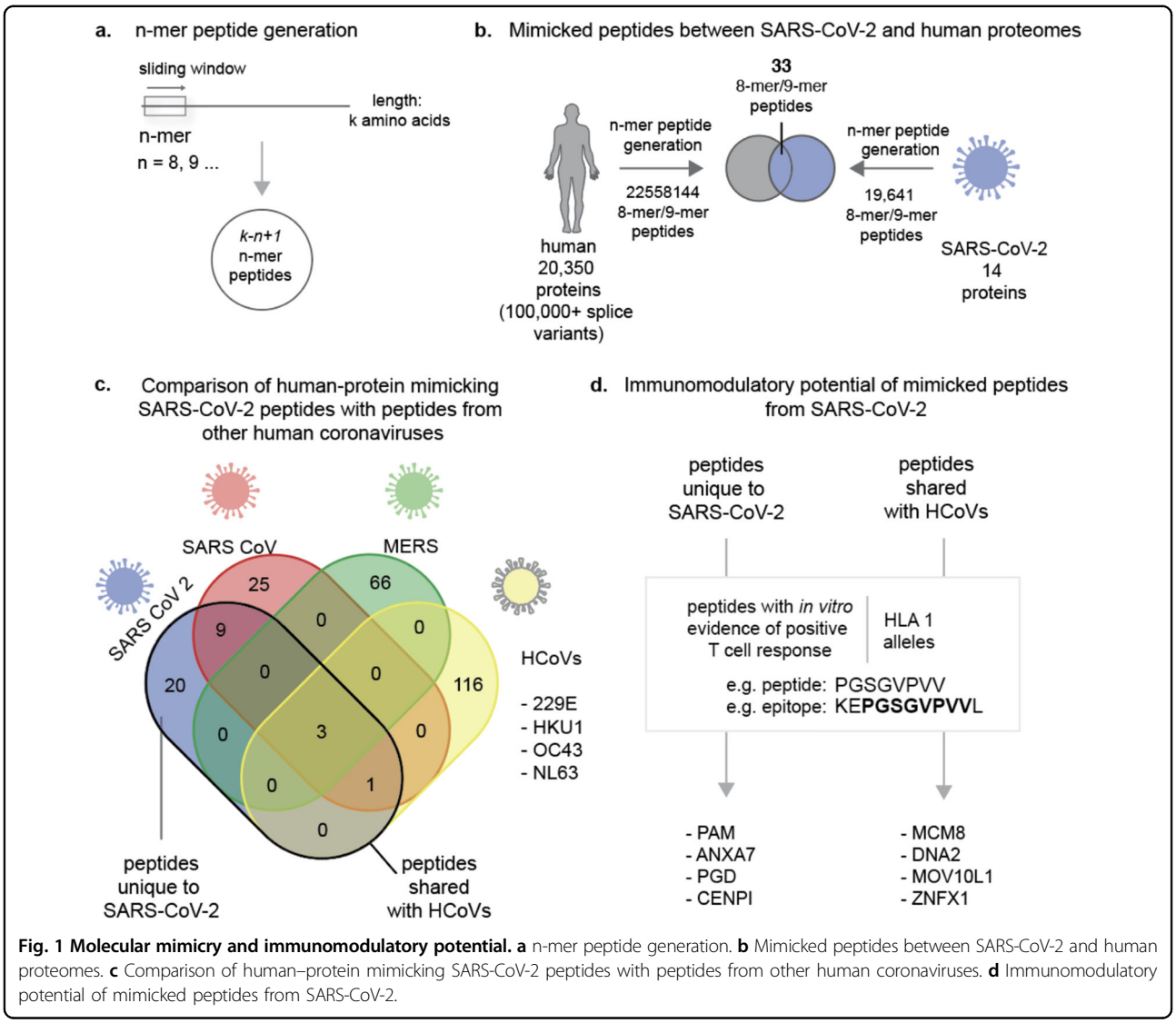

Fig. 1 Molecular mimicry and immunomodulatory potential. a n-mer peptide generation. **b** Mimicked peptides between SARS-CoV-2 and human proteomes. **c** Comparison of human–protein mimicking SARS-CoV-2 peptides with peptides from other human coronaviruses. **d** Immunomodulatory potential of mimicked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

COVID-19, by performing a systematic comparison 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 9827 distinct 8-mers and 9814 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 (Fig. 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 8-mer/9-mers per viral taxon were shared with the

human proteome (mean = 45.57, median = 12, SD = ±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, SD = ±0.011) (Fig. 1, Supplemental Fig. 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 the 14 SARS-CoV-2 proteins and 29 of the 20,350 human proteins. By including alternative splicing, the 33 8-mers/9-mers mimicked by SARS-CoV-2 map onto 39 of the 100,566 protein splicing variants. That is, 0.16% of the human proteins and 0.04% of all the known 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 8–12 mers[9], the analysis that follows was restricted to mapping host–pathogen mimicry from an immunologic perspective across 8-mer and 9-mer peptides only.

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 14 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 across different human coronaviruses was generated (Fig. 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 (Fig. 1, Supplementary Fig. 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. During the comparison of gene expression across tissue types in GTEx, the specificity of expression is computed using "Cohen's D", which is an effect size used to indicate the standardized 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 8-mer/9-mers per viral taxa were detected to be identical to a known human protein (mean = 45.57,

Table 1 SARS-CoV-2 peptides mimicking human proteins, with experimental evidence of positive MHC binding from the immune epitope database.

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
	ORF1ab polyprotein (Helicase)		DNA2 (inferred) (P51530: 1000–1007)	HLA-A*03:01 common hydrophobic amino acids at P2	53748	1000425 (RefID)

Table 1 continued

Viral peptide (coronavirus)	Viral protein (NCBI)	Human epitope	Human protein	MHC restriction (positive response)	Epitope ID (IEDB)	Pubmed ID (PMID)
NVAITRAK (SARS-CoV-2; SARS-CoV; Seasonal HCoV)	(YP_009725308.1 :561-568)	RFNVAITRAK (antigen source: SARS)		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)

The viral-human mimicked 8-mer/9-mer peptides are highlighted in green text.

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 human 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 four 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) (Fig. 1d). Analyzing the sequence conservation of the SARS-CoV-2-exclusive peptides shared with the above four 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, and 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. In addition, there are four human helicases (MCM8, DNA2, MOV10L1, and 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 c; Fig. 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) (Fig. 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) 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) (Fig. 2b). Further, immunohistochemistry (IHC) based data on 45 human tissues[14] shows that the PAM protein is detected at high levels in heart muscles, epididymis, and the adrenal gland (Fig. 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 and 90% of activated/quiescent stellate cells, acinar cells, endothelial cells, ductal cells of the pancreas. It is also expressed in over 80% of cardiomyocytes, and 40–70% of heart fibroblasts, macrophages, endothelial cells, and smooth muscle cells, as well as in 26–27% of lung pleura fibroblasts, stromal cells, and neutrophils (Fig. 2c). Moreover, analyzing the scRNA-seq data from severe COVID-19 patient's lung bronchoalveolar lavage fluid shows high PAM expression in club cells (Fig. 2d), 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 (Fig. 2e). 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 nonzero 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) (Fig. 3—Supplementary Fig. 1). 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) (Fig. 3—Supplementary Fig. 2). Analysis of the scRNA-seq data on ANXA7 confirms expression in endothelial cells across multiple tissues and organs, and also indicates expression in lung type-2 pneumocytes, macrophages, oligodendrocytes (Fig. 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 (Fig. 3b). Additionally from the study of normal lungs, activated dendritic cells and lymphatic vessel cells are noted to express ANXA7 significantly (Fig. 3c).

Table 2 Amino acid sequence conservation of the SARS-CoV-2 peptides mimicking human proteins.

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%]

The PGSGVPVV peptide from the NSP12 protein is present in 46,079 out of 46,513 SARS-CoV-2 sequences (99.1% conserved; mimics human PAM protein), the ESGLKTIL peptide from the NSP2 protein is present in 44750 out of 46,513 SARS-CoV-2 sequences (96.2% conserved; mimics human ANXA7), the VTLIGEAV peptide from the endoRNAase protein is present in 43,710 of 46,513 SARS-CoV-2 sequences (94% conserved; mimics human PGD); and the SLKELLQN peptide from the 3C-like proteinase is present in 45,888 of 46,513 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 45,927 out of 46,513 SARS-CoV-2 sequences (98.7% conserved; mimics human helicase MCM8 protein), the NVAITRAK peptide from the viral helicase is present in 45,834 out of 46,513 SARS-CoV-2 sequences (98.5% conserved; mimics human helicase DNA2), the RFNVAITR peptide from the viral helicase is present in 45,842 of 46,513 SARS-CoV-2 sequences (98.6% conserved; mimics human helicase MOV10L1); and the QGPPGTGK peptide from the viral helicase is present in 46,150 of 46,513 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.

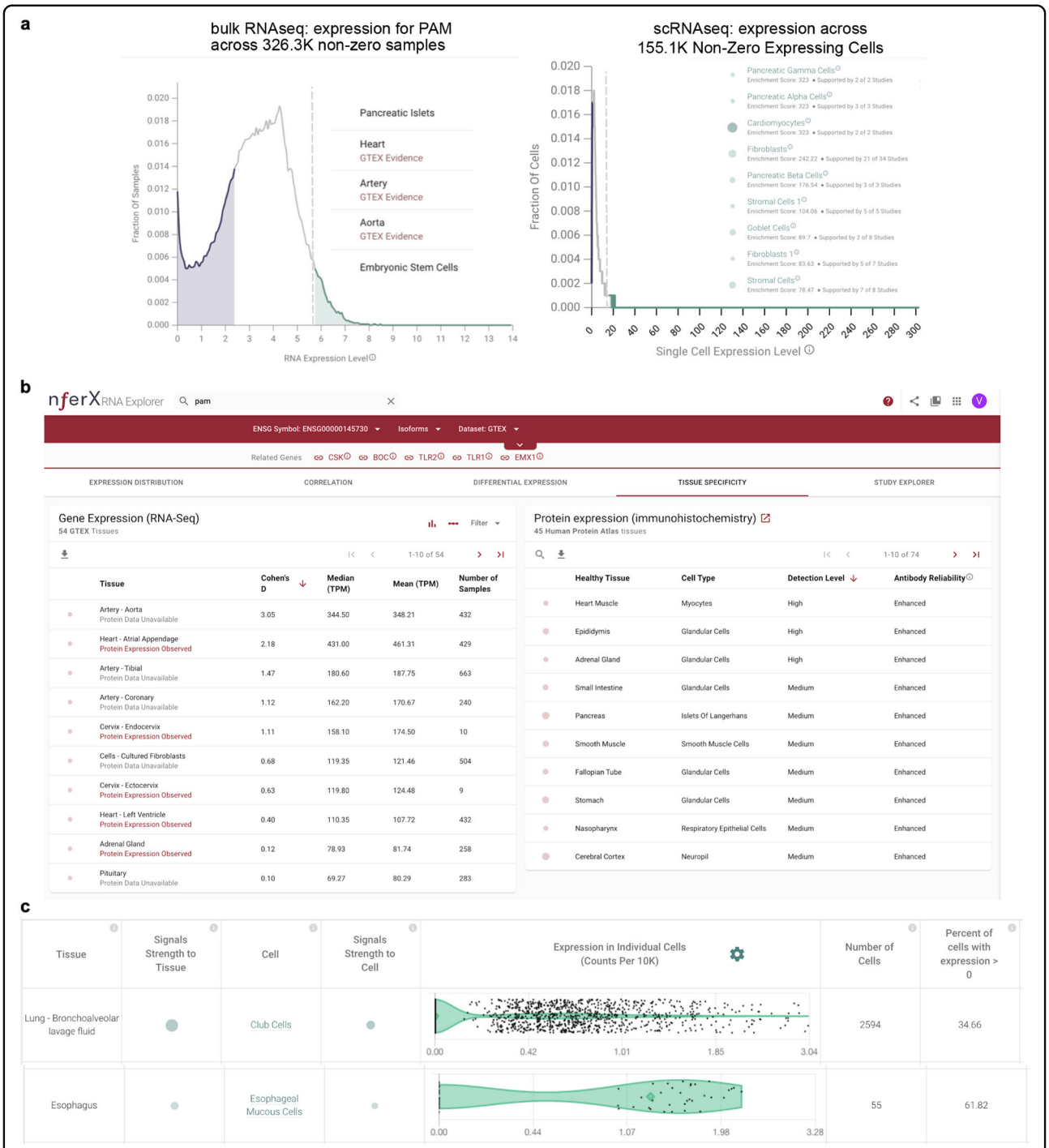

Fig. 2 Multi-omics analysis of human PAM. a (Left) Universal bulk RNA-seq analysis of all available human data shows pancreatic islets, heart, artery, aorta, and embryonic stem cells harbor PAM significantly. (Right) Single-cell RNA-seq (scRNA-seq) confirms high PAM-expressing cells, include multiple 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).

Fig. 3 Evidence of ANXA7 expression from human single-cell RNA-sequencing data. **a** List of high ANXA7-expressing cells across human tissues. High ANXA7-expressing cells in the lungs are shown on the right. These include macrophages, proliferating cells, mast cells, stromal cells, type-2 pneumocytes and endothelial cells. **b** 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. In the lungs, activated dendritic cells and lymphatic vessel cells are also seen to express ANXA7. **a**, **b** The size and transparency of the bubbles are proportional to the strength of literature-based associations (https://academia.nferx.com/).

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). (Fig. 4, Supplementary Fig. 1). IHC data on 45 human tissues from the Human Protein Atlas[14] confirms that PGD is detected at high levels in the esophagus (Fig. 4, Supplementary Fig. 2), and additionally in the testes, tonsils, bone marrow, gallbladder, spleen, and placenta.

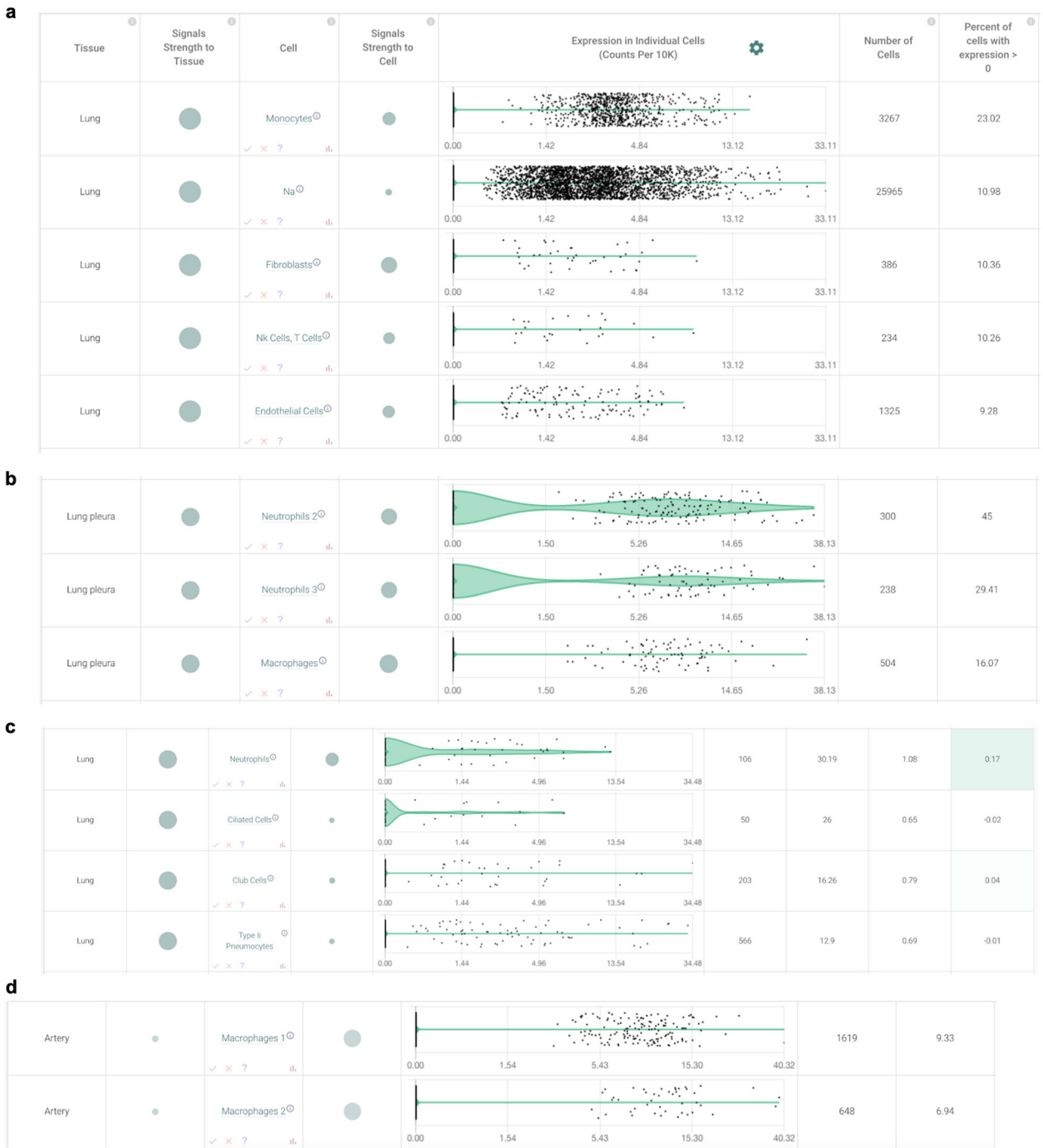

Fig. 4 Evidence of PGD expression from human single-cell RNA-sequencing data. Single-cell RNA-seq analysis based expression of PGD in cell types of **a** lungs, **b** lung pleura, **c** airway epithelia and **d** artery. The size and transparency of the bubbles are proportional to the strength of literature-based associations (https://academia.nferx.com/).

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**). In addition, 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 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 and ENST00000617770.4) containing the 8-mer peptide "PEANMDQE" is one of four 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 (Fig. 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 (Fig. 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.

Mimicked HLA-A*03-binding peptides are shared between HCoV helicases and the human helicases

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 various cell-types including neuronal cells and immune cells (Fig. 6). The HLA-A*03:01 allele has been established from in vitro experiments to bind SARS-CoV helicase peptides that mimic 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 the 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 toward 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

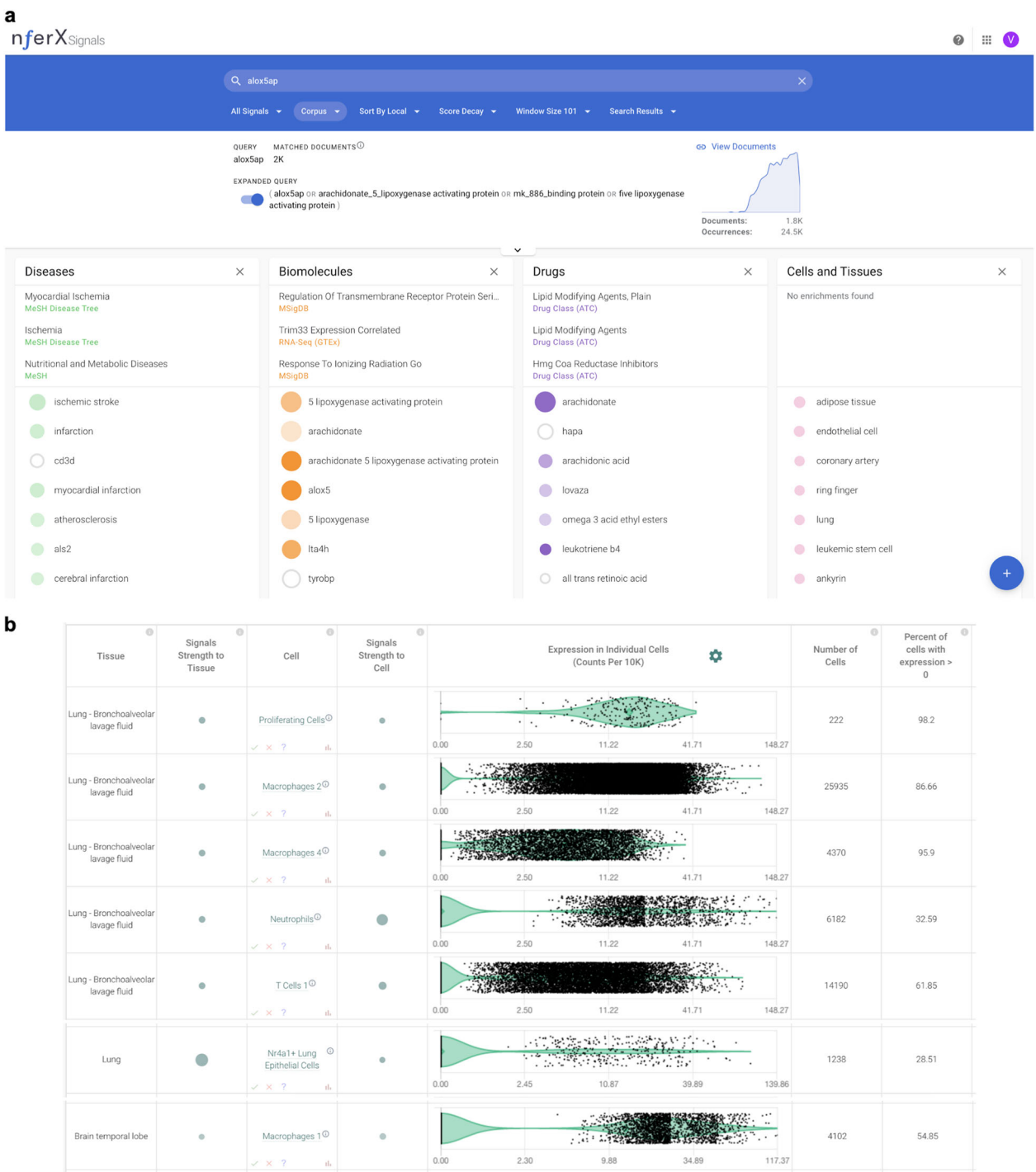

Fig. 5 Evidence for ALOX5AP from biomedical knowledge synthesis and single-cell RNA-seq. a Knowledge synthesis suggests involvement of ALOX5AP in ischemic stroke, myocardial infarction, atherosclerosis, cerebral infarction, and coronary artery disease. **b** scRNA-seq shows significant expression of ALOX5AP in proliferating cells, macrophages, T-cells, and epithelial cells from the lungs and macrophages of the brain.

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 and HCoV-HKU1) and previous SARS strains (SARS-CoV and MERS). Further experiments are required to assess any potential for autoinflammation.

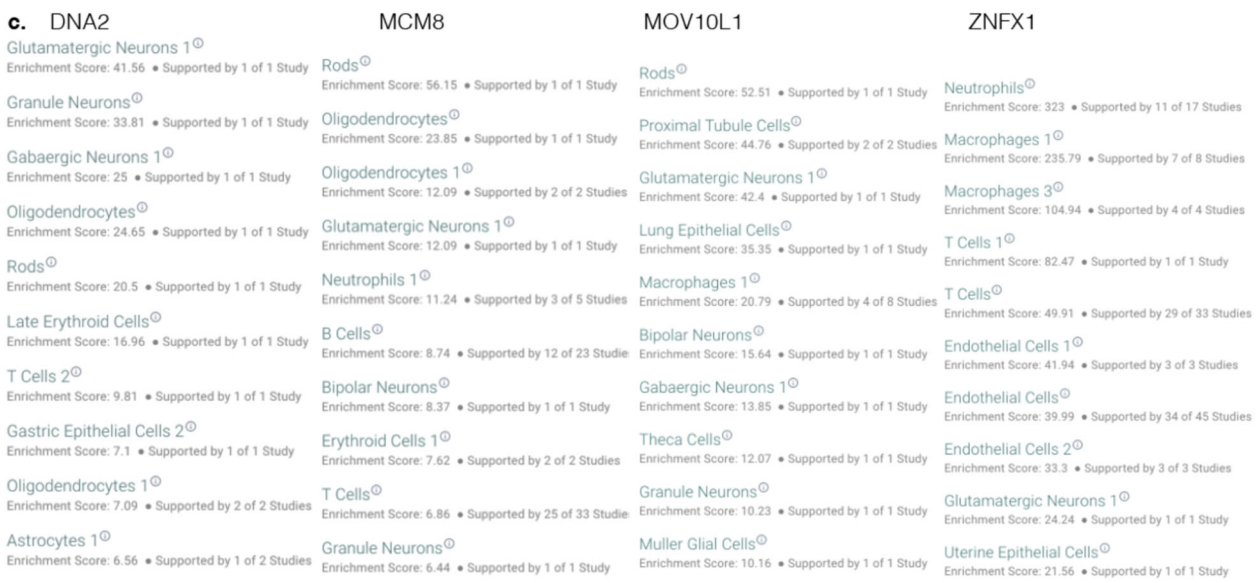

Fig. 6 Expression of mimicked human helicases. scRNAseq based expression analysis of human helicases containing peptides mimicked by helicases in SARS-CoV-2 and other human coronaviruses (https://academia.nferx.com/).

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 performed for all the four immunomodulatory peptides specific to SARS-CoV-2 against all the sequences of Coronaviridae family in the nonredundant protein database. There were no hits found outside the orthocoronavirinae family for these peptides. An exact match for peptides—"PGSGVPVV", "VTLI-GEAV", 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 (Fig. 3, Supplementary Fig. 2). Our observed multi-pronged human mimicry of distinct SARS-CoV-2 peptides may 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, 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 light on the relentless evolutionary tinkering that propels the rise and fall of viral pandemics.

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

Author details
[1]nference, Cambridge, MA, USA. [2]nference Labs, Bangalore, India. [3]Department of Infectious Diseases, Mayo Clinic, Rochester, MN, USA. [4]Department of Genetics, Harvard Medical School, Boston, MA, USA

Conflict of interest
The authors declare that they have no conflict of interest.

The online version of this article (https://doi.org/10.1038/s41420-020-00321-y) contains supplementary material, which is available to authorized users.

Received: 18 August 2020 Accepted: 1 September 2020
Published online: 02 October 2020

References

1. Smatti, M. K. et al. Viruses and Autoimmunity: A Review on the Potential Interaction and Molecular Mechanisms. *Viruses* **11**, 762. https://doi.org/10.3390/v11080762 (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 2020* **163**, 473–479, https://doi.org/10.1177/0194599820931820 (2020).
3. Chen, T. et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *Br. Med. J.* **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. *eLife* **9**, e58227. https://doi.org/10.7554/eLife.58227 (2020).
5. Venkatakrishnan, A. et al. Knowledge synthesis of 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. *eLife* **9**, e58040. https://doi.org/10.7554/eLife.58040 (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. https://academic.oup.com/nar/article/47/D1/D506/5160987 (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–D442 (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. Proteom.* **18**, 2298–2309 (2019).
17. Pearson, H. et al. MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Investig.* **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* https://doi.org/10.1161/CIRCULATIONAHA.120.048360 (2020).
19. Viner, R. M. & Whittaker, E. Kawasaki-like disease: emerging complication during the COVID-19 pandemic. *Lancet* **395**, 1741–1743 (2020).
20. Vita, R. et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
21. Vita, R. et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405–D412 (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).
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.* https://doi.org/10.1016/S1473-3099(20)30434-5 (2020).
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.* https://doi.org/10.1056/NEJMoa2015432 (2020).
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).