Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

1 Manuscript Number: TBD

2	Running Title: Highly Distinguished Amino Acid Sequences of 2019-nCoV
3	Keywords: Coronavirus, Amino Acid Sequence Analysis, Communicable Diseases, Emerging
4	
5	Title: Highly Distinguished Amino Acid Sequences of 2019-nCoV (Wuhan Coronavirus)
6	Authors: Jacob Beal ¹ , Thomas Mitchell ¹ , Daniel Wyschogrod ¹ , Jeff Manthey ² , Adam Clore ²
7	Affiliations:
8	¹ Raytheon BBN Technologies, Cambridge, MA, USA (J. Beal, T. Mitchell, D. Wyschogrod)
9	² Integrated DNA Technologies, Coralville, Iowa, USA (J. Manthey, A. Clore)
10	
11	Abstract
12	Using a method for pathogen screening in DNA synthesis orders, we have identified a
13	number of amino acid sequences that distinguish 2019-nCoV (Wuhan Coronavirus) from all
14	other known viruses in Coronaviridae. We find three main regions of unique sequence: two in
15	the 1ab polyprotein QHO60603.1, one in surface glycoprotein QHO60594.1.

16

17 **Text**

Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

18	The emerging coronavirus 2019-nCoV(l) is of significant world-wide concern as it
19	spreads from its initial point of identification in Wuhan. Identification of significant areas of
20	uniqueness that distinguish such an emerging pathogen may be of value in the development of
21	methods for diagnosis, prevention, or treatment. To this end, we have identified a number of
22	amino acid sequences that distinguish 2019-nCoV from all other known viruses within the
23	family Coronaviridae. Amongst these, we find three main regions of unique sequence: two in the
24	1ab polyprotein QHO60603.1, one in the surface glycoprotein QHO60594.1.
25	To identify unique sequences, we adapted FAST-NA, a software tool for screening DNA
26	synthesis orders for pathogens $(2,3)$ that uses methods for automatic signature generation
27	developed originally for cybersecurity malware detection(4). In particular, FAST-NA compares
28	all k-mer sequences of a collection of target sequences to a collection of contrasting sequences in
29	order to identify all k-mer sequences that are unique to the target population. These unique
30	sequences are diagnostic of membership in the population, whereas shared sequences indicate
31	structure that is conserved to some degree.
2.2	

Here, we applied FAST-NA to identify all of the unique 10-mer sequences in all of the 32 33 amino acid sequences for 2019-nCoV then available from NCBI: 63 amino acid sequences available in NCBI, comprising a total of 49379 amino acids (5-8). For contrasting sequences, we 34 used a July, 2019 snapshot of all protein sequences in family Coronaviridae available from 35 NCBI, a total of 50574 sequences comprising a total of approximately 40 million residues. The 36 resulting collection of unique 10-mer amino acids sequences were then concatenated where 37 38 overlapping within the same parent sequence and trimmed to remove non-unique flanking portions. 39

Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

40	All told, this process identifies 61 multi-amino-acid regions as significant unique
41	sequences for 2019-nCoV, comprising a total of 1669 amino acids (3.4% unique and non-
42	repeated), spread across 8 non-duplicative sequences (Appendix Table 1). In addition, we also
43	identified 45 single amino-acid polymorphisms (Appendix Table 2). Figure 1 summarizes the
44	distribution of unique sequence regions across these 8 open reading frame (ORF) sequences.
45	Two of these have notably high amounts of unique content: the large 1ab polyprotein
46	QHO60603.1 has much unique material, though the fraction is not large, while the surface
47	glycoprotein QHO60594.1 has both a large amount and large fraction of unique material.
48	Further examination shows that the unique material in these two ORFs is strongly
49	clustered. Taking a cluster as any sequence of at least three unique regions with no more than 50
50	amino acids separating them, we find that QHO60603.1 has two clusters, one spanning from
51	residues $916 - 1294$, the other from $6417 - 6715$, containing 47% of the unique material in the
52	sequence. The QHO60594.1 sequence, meanwhile, has a single large cluster, spanning from
53	residues 9 to 883 and comprising all of the unique material in the sequence.
54	In summary, analysis of the amino acid sequences of 2019-nCoV identifies three large
55	highly unique regions of the genome that distinguish it from all other Coronaviridae, plus
56	several dozen other smaller regions of uniqueness. We thus hypothesize that these three large
57	regions are likely to be of significance in understanding the evolution and infectivity of 2019-

nCoV, in development of countermeasures to mitigate its effects, and in the selection of

recommend them as a potential focus of attention.

59 diagnostic assays to understand and track the origin and spread of this disease, and therefore

61

Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

62 Acknowledgments

63	This research was sponsored by IARPA contract 2018-17110300002 and by the Army
64	Research Office and under Grant Number W911NF-17-2-0092. The views and conclusions
65	contained in this document are those of the authors and should not be interpreted as representing
66	the official policies, either expressed or implied, of the Army Research Office or the U.S.
67	Government. The U.S. Government is authorized to reproduce and distribute reprints for
68	Governmental purposes notwithstanding any copyright notation thereon. This document does not
69	contain technology or technical data controlled under either U.S. International Traffic in Arms
70	Regulation or U.S. Export Administration Regulations.
71	Author Bio
72	Dr. Jacob Beal is a Senior Scientist at Raytheon BBN Technologies, where he leads
72 73	Dr. Jacob Beal is a Senior Scientist at Raytheon BBN Technologies, where he leads research on synthetic biology and distributed systems engineering. His work in synthetic biology
72 73 74	Dr. Jacob Beal is a Senior Scientist at Raytheon BBN Technologies, where he leads research on synthetic biology and distributed systems engineering. His work in synthetic biology includes development of methods for calibrated flow cytometry, precision analysis and design of
72 73 74 75	Dr. Jacob Beal is a Senior Scientist at Raytheon BBN Technologies, where he leads research on synthetic biology and distributed systems engineering. His work in synthetic biology includes development of methods for calibrated flow cytometry, precision analysis and design of genetic regulatory networks, engineering of biological information processing devices, standards
72 73 74 75 76	Dr. Jacob Beal is a Senior Scientist at Raytheon BBN Technologies, where he leads research on synthetic biology and distributed systems engineering. His work in synthetic biology includes development of methods for calibrated flow cytometry, precision analysis and design of genetic regulatory networks, engineering of biological information processing devices, standards for representation and communication of biological designs, and signature-based detection of
72 73 74 75 76 77	Dr. Jacob Beal is a Senior Scientist at Raytheon BBN Technologies, where he leads research on synthetic biology and distributed systems engineering. His work in synthetic biology includes development of methods for calibrated flow cytometry, precision analysis and design of genetic regulatory networks, engineering of biological information processing devices, standards for representation and communication of biological designs, and signature-based detection of pathogenic sequences.

79 1. N Zhu, D Zhang, W Wang, X Li, B Yang, J Song, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020 Jan 24.

Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

	81	2.	J Beal, D Wyschogrod, T Mitchell, S Katz, J Manthey, A Clore. Applicability of FAST-
	82		NA to Nucleic Acid Screening. Final Report on Intelligence Advanced Research Projects
	83		Activity Contract No. 2018–17110300002, Feb 2019.
	84	3.	S Adali, A Adler, J Bader, J Grothendieck, T Mitchell, A Persikov, et al. Towards
	85		Detection of Engineering in Metagenomic Sequencing Data for Yeast and Other Fungi,
	86		In International Workshop on BioDesign Automation, 2019.
	87	4.	D Wyschogrod, J Dezso. False alarm reduction in automatic signature generation for
	88		zero-day attacks. In 2nd Cyberspace Research Workshop, 2009.
	89	5.	K Queen, Y Tao, Y Li, CR Paden, X Lu, B Lynch, et al. Full genome sequence of first
	90		U.S. case of nCoV-2019. Submitted (24-JAN-2020) Division of Viral Diseases, Centers
	91		for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA
	92	6.	JF-W Chan, S Yuan, KH Kok, KK-W To, H Chu, et al. A familial cluster of pneumonia
	93		associated with the 2019 novel coronavirus indicating person-to-person transmission: a
	94		study of a family cluster. Lancet (2020) In press
	95	7.	R Buathong, S Wacharapluesadee, S Lamsirithawon, W Chaifoo, T Ponpinit, Y
	96		Joyjinda, Y, et al. Direct Submission. Submitted (19-JAN-2020) Faculty of Medicine,
	97		Chulalongkorn University, Rama IV Rd, Bangkok 10330, Thailand
	98	8.	F Wu, S Zhao, B Yu, Y-M Chen, W Wang, Y Hu, et al. A novel coronavirus associated
	99		with a respiratory disease in Wuhan of Hubei province, China. Submitted (05-JAN-2020)
1	L00		Shanghai Public Health Clinical Center & School of Public Health, Fudan University,
1	101		Shanghai, China.

Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

103 Address for correspondence: Jacob Beal, Raytheon BBN Technologies, 10 Moulton Street,

104 Cambridge, MA, USA; email: jakebeal@ieee.org

105

- **Figure 1.** Summary statistics of distinguishing amino acid sequences identified for 2019-nCoV
- 107 (Wuhan coronavirus), showing the fraction of each ORF judged to be part of unique sequences
- and the total number of amino acids in unique sequences in the ORF. The large 1ab polyprotein
- 109 QHO60603.1 has much unique material, though the fraction is not large, while the surface
- 110 glycoprotein QHO60594.1 has both a large amount and large fraction of unique material.

Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

112 Appendix Table 1. Unique amino acid sequences of 2019-nCoV. Three clusters of unique

sequences with less than 50 as separation are highlighted in red.

Accession	Start	End	Sequence
QHO60603.1	153	173	Y EDFQE NW NT KHSS GVT RE LM
QHO60603.1	395	416	ILR KG G RT IAFG GCV FSYVG CH
QHO60603.1	556	563	NSVRVLQK
QHO60603.1	590	607	AT NNLVV MAYITG G VVQ L
QHO60603.1	721	7 27	KSREETG
QHO60603.1	761	777	DLQPLEQPTSEAVEAPL
QHO60603.1	916	939	AS HMYCS FY PP DE DE EEG DC EE EE
QHO60603.1	966	1038	A A LQ PE EE QE ED W LD DDSQ QT VG QQ DGS ED NQT TT IQT IV E VQ PQLE MELT P VV QT IE V NSFSG Y LK LT DN VY
QHO60603.1	1088	1175	DY IAT NG P LKVG G SCVLSG HNLAKHC LHVVG P NVNKG E DIQLLKSAYEN F NQHEV LLAP LLSAGI FG AD PI HSLRV CVDT V RT NVY LA
QHO60603.1	1 20 2	1229	I AE IP KE EV KP FIT ES KPS VEQ.R KQ.DDK
QHO60603.1	1 27 1	1294	SDIDITFLKKDAPYIVG DV VQEGV
QHO60603.1	1549	1557	VITEDNIKT
QHO60603.1	1778	1794	FKKG VQIPCTCG KQATK
QHO60603.1	1934	1944	IKFADDINQLT
QHO60603.1	2026	2060	V LKSEDAQG MDN LACEDLKPVSEEVVEN PTIQKDV
QHO60603.1	2080	2082	NNS
QHO60603.1	2171	2190	FFTLLLQLCTFTRSTNSRIK
QHO60603.1	2210	22.29	LEASFNYLKSPNFSKLINII
QHO60603.1	2596	2610	T FSST FNV PMEKLKT
QHO60603.1	278 2	2799	V AAIFY LIT PV HV MSKHT
QHO60603.1	3051	3055	IVAIV
QHO60603.1	3139	3144	ITIAYI
QHO60603.1	3586	3611	ILTSLLV LVQSTQWSLFFFLYE NAFL
QHO60603.1	4073	4086	IPDYNTYKNTCDGT
QHO60603.1	4174	4187	T KG G RFV LALLS DL
QHO60603.1	4390	4397	LQSADAQS
QHO60603.1	4453	4489	DDNLIDSY FV VKRHT FS NYQHEETIYNLLKDCP AV AK
QHO60603.1	4643	4672	T AES HV DT DLT KPY IKW DLLKYDFTE ER LK
QHO60603.1	5130	5131	TD
QHO60603.1	5157	5172	FNSTYASOG LVASIKN
QHO60603.1	6052	6058	PNNTDFS
QHO60603.1	6144	6155	ASDTYACWHHSI
QHO60603.1	6417	6434	LYLDAY NIMIS AG FSLWV
QHO60603.1	6458	6493	F NVV NKG HF DG QQG EVPVS IIN NT VYT KV DG V DV EL
QHO60603.1	6542	6573	DAPAHISTIGVCSMTDIAKKPTETICAPLTVF
QHO60603.1	6603	6630	QPSVGPKQASLNGVTLIS EAVKTQFNYY
QHO60603.1	6652	6674	QEFKPRSQ.MEIDF.LE LAMDEF.LE
QHO60603.1	6694	6715	SQLGGLHLLIGLAKRFKESPFE
QHO60603.1	7062	7086	g Q IN DMILSLISK G R LIIR EN NRVV
QHO60602.1	9	28	P FT IY SLLLCR MNSR NY IAQ
QHO60601.1	10	32	NAPRIT FG GPSDS TGS NQNG ERS
QHO60601.1	62	78	DLKFPRGQGVPINTNSS
QHO60601.1	216	233	AALALLLIDRINQLESKM
QHO60601.1	401	409	DFSKQLQQS
QHO60600.1	9	43	ITT VAAFHQECS LQSCTQHQPYVVDDpCPHFYSK
QHO60600.1	103	108	FYEDFL
QHO60599.1	9	10	
QHO60599.1	71	73	KHV
QHO60599.1	94	111	ELYSPIFLIVARIVEITL
QHO60598.1	42	48	SLTENKY
QHO60595.1	9	39	IGT VT LKQGE IKDATPS DFV RATAT IPIQAS
QHO60595.1	89	1 26	VYSHLLIV AAG LEAPF LY LYA LYY FLOS IN FY RI IMR L
QHO60595.1	170	181	SG DG TTSPISEH
QHO60594.1	9	275	LUSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTW FHAIHVSG TNG TKRFDNPVLPFNDG VYFASTEKSNIIRG WIFGTTLDSKTQSLLIV NNAT NVVIKVCEFQFCNDPFLG VYHKNNKSW MESEFRVYSSA NNCTEFYSQPFLMDLEGKQQNFKNLREFVKNDG YFKIYSKHTPINLVRDLPQFSALEPLVDLPG INITRFOTLLAI HRYYTPG FXSSG WTAGAAAYYVGY I OPRTFI
QHO60594 1	30.5	3 25	FTVEKGIVOTSNERVOPTESI
QHO60594 1	345	371	R F AS VY AW N R KR ISNC VADYSV LYNSA
QHO60594 1	39.2	416	T NYY ADSFY IRG DEVROI APG OTG K
QHO60594 1	437	531	S N NLDS KV G G N Y NY LY RLFRKSN LKPFERDISTE I YO AGSTPCNG VEG FNCYFPLOSYGFOPT NG V G Y OPYR V V LSFELLHAPATV CG PKKST N
QHO60594 1	553	574	ESNKKFLPFQQFgRDIADTTDA
QHO60594 1	60.5	7 25	N V V V V V V V V V V V V V V V V V V V
OHO60594 1	871	883	0/154/467/15
2.1000004.1	0,1	000	

115 Appendix Table 2. Additional single-amino-acid polymorphisms of 2019-nCoV.

Reserved space. Do not place any text in this section. Include the mandatory author checklist or your manuscript will be returned. Use *continuous* line numbering in your manuscript.

Accession	Single AA Polymorphisms location=value
QHO60603.1	37=V 92=E 113=I 279=I 337=K 375=S 444=G 497=A 858=A 1392=V 1439=D 1732=S 1821=T
	1861=P 1897=N 2006=T 2129=V 2264=G 2452=V 2876=T 3085=L 3668=M 3846=V 3956=F
	4114=S 4275=A 5038=S 5938=V 6023=E 6100=N 6217=T 6243=A 6298=S 6361=V 6520=V
QHO60601.1	102=D 127=D 333=T 378=T
QHO60600.1	64=A
QHO60594.1	844=A 1083=D 1132=V
QHN73809.1	3098=L
QHD43422.1	83=L

