Manuscript Number: TBD

Running Title: Highly Distinguished Amino Acid Sequences of 2019-nCoV

Keywords: Coronavirus, Amino Acid Sequence Analysis, Communicable Diseases, Emerging

Title: Highly Distinguished Amino Acid Sequences of 2019-nCoV (Wuhan Coronavirus)

Authors: Jacob Beal[1], Thomas Mitchell[1], Daniel Wyschogrod[1], Jeff Manthey[2], Adam Clore[2]

Affiliations:

[1]Raytheon BBN Technologies, Cambridge, MA, USA (J. Beal, T. Mitchell, D. Wyschogrod)

[2]Integrated DNA Technologies, Coralville, Iowa, USA (J. Manthey, A. Clore)

Abstract

Using a method for pathogen screening in DNA synthesis orders, we have identified a number of amino acid sequences that distinguish 2019-nCoV (Wuhan Coronavirus) from all other known viruses in *Coronaviridae*. We find three main regions of unique sequence: two in the 1ab polyprotein QHO60603.1, one in surface glycoprotein QHO60594.1.

Text

The emerging coronavirus 2019-nCoV(*1*) is of significant world-wide concern as it spreads from its initial point of identification in Wuhan. Identification of significant areas of uniqueness that distinguish such an emerging pathogen may be of value in the development of methods for diagnosis, prevention, or treatment. To this end, we have identified a number of amino acid sequences that distinguish 2019-nCoV from all other known viruses within the family *Coronaviridae*. Amongst these, we find three main regions of unique sequence: two in the 1ab polyprotein QHO60603.1, one in the surface glycoprotein QHO60594.1.

To identify unique sequences, we adapted FAST-NA, a software tool for screening DNA synthesis orders for pathogens(*2*,*3*) that uses methods for automatic signature generation developed originally for cybersecurity malware detection(*4*). In particular, FAST-NA compares all k-mer sequences of a collection of target sequences to a collection of contrasting sequences in order to identify all k-mer sequences that are unique to the target population. These unique sequences are diagnostic of membership in the population, whereas shared sequences indicate structure that is conserved to some degree.

Here, we applied FAST-NA to identify all of the unique 10-mer sequences in all of the 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 used a July, 2019 snapshot of all protein sequences in family *Coronaviridae* available from NCBI, a total of 50574 sequences comprising a total of approximately 40 million residues. The resulting collection of unique 10-mer amino acids sequences were then concatenated where overlapping within the same parent sequence and trimmed to remove non-unique flanking portions.

All told, this process identifies 61 multi-amino-acid regions as significant unique sequences for 2019-nCoV, comprising a total of 1669 amino acids (3.4% unique and non-repeated), spread across 8 non-duplicative sequences (Appendix Table 1). In addition, we also identified 45 single amino-acid polymorphisms (Appendix Table 2). Figure 1 summarizes the distribution of unique sequence regions across these 8 open reading frame (ORF) sequences. Two of these have notably high amounts of unique content: the large 1ab polyprotein QHO60603.1 has much unique material, though the fraction is not large, while the surface glycoprotein QHO60594.1 has both a large amount and large fraction of unique material.

Further examination shows that the unique material in these two ORFs is strongly clustered. Taking a cluster as any sequence of at least three unique regions with no more than 50 amino acids separating them, we find that QHO60603.1 has two clusters, one spanning from residues 916 – 1294, the other from 6417 – 6715, containing 47% of the unique material in the sequence. The QHO60594.1 sequence, meanwhile, has a single large cluster, spanning from residues 9 to 883 and comprising all of the unique material in the sequence.

In summary, analysis of the amino acid sequences of 2019-nCoV identifies three large highly unique regions of the genome that distinguish it from all other *Coronaviridae*, plus several dozen other smaller regions of uniqueness. We thus hypothesize that these three large 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 diagnostic assays to understand and track the origin and spread of this disease, and therefore recommend them as a potential focus of attention.

Acknowledgments

This research was sponsored by IARPA contract 2018-17110300002 and by the Army Research Office and under Grant Number W911NF-17-2-0092. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. This document does not contain technology or technical data controlled under either U.S. International Traffic in Arms Regulation or U.S. Export Administration Regulations.

Author Bio

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.

References

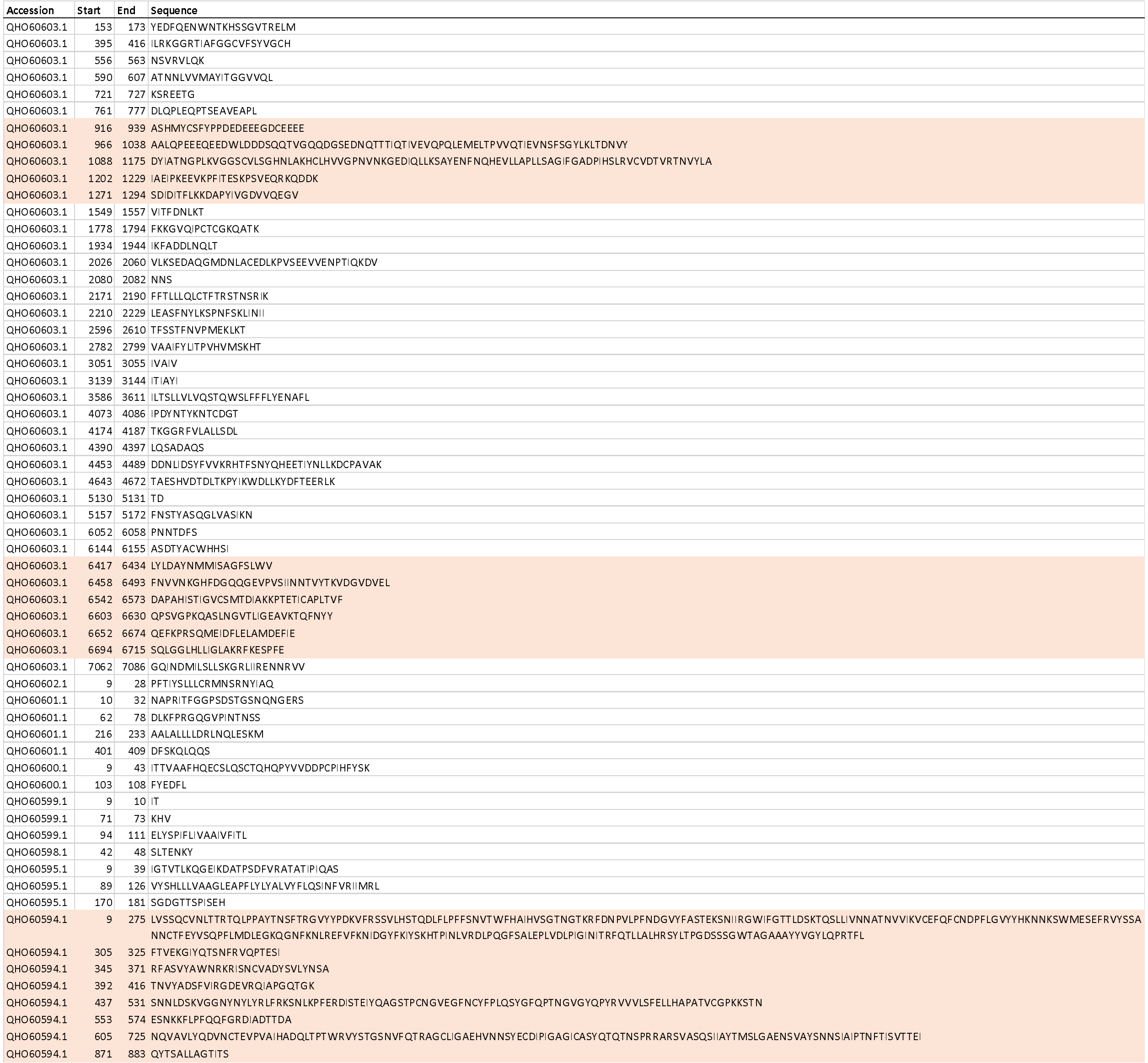
1. N Zhu, D Zhang, W Wang, X Li, B Yang, J Song, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020 Jan 24.

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

Address for correspondence: Jacob Beal, Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA, USA; email: jakebeal@ieee.org

Figure 1. Summary statistics of distinguishing amino acid sequences identified for 2019-nCoV (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 QHO60603.1 has much unique material, though the fraction is not large, while the surface glycoprotein QHO60594.1 has both a large amount and large fraction of unique material.

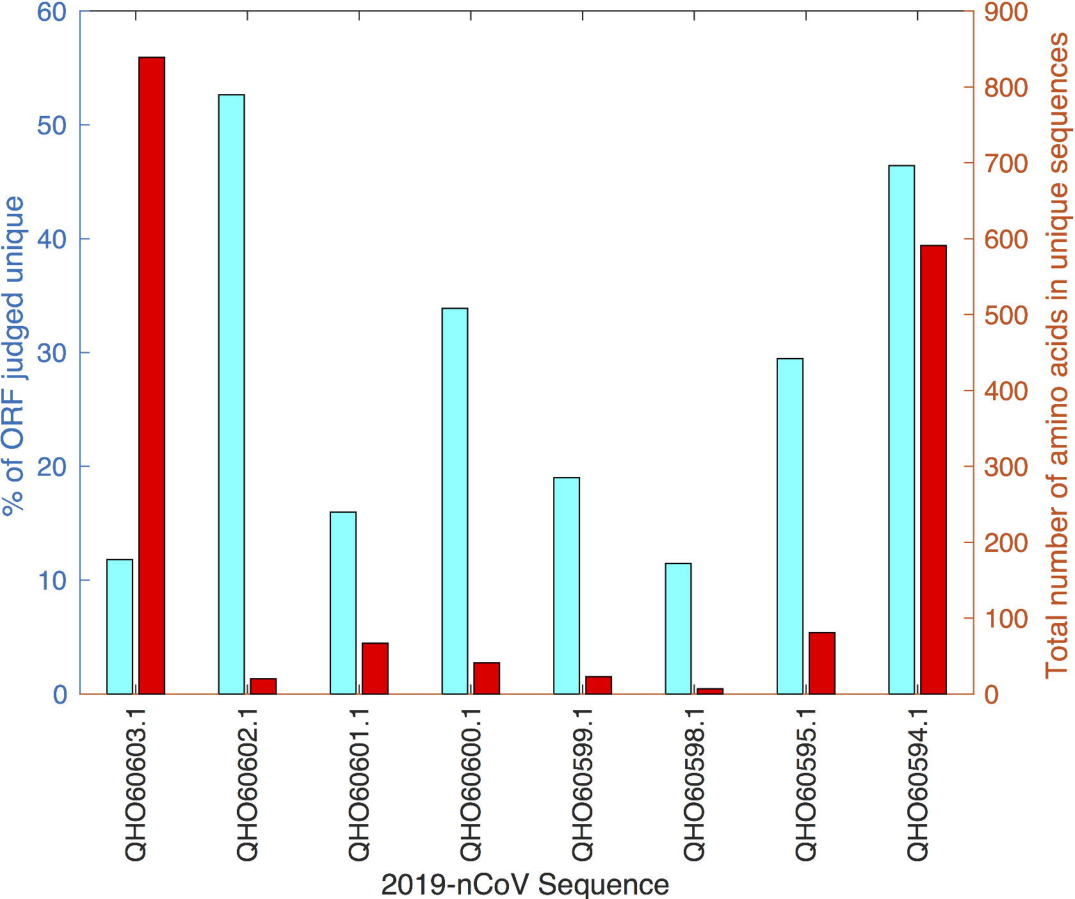
Appendix Table 1. Unique amino acid sequences of 2019-nCoV. Three clusters of unique sequences with less than 50 aa separation are highlighted in red.



Accession	Start	End	Sequence
QHO60603.1	153	173	YEDFQENWNTKHSSGVTRELM
QHO60603.1	395	416	ILRKGGRTIAFGGCVFSYVGCH
QHO60603.1	556	563	NSVRVLQK
QHO60603.1	590	607	ATNNLVVMAYITGGVVQL
QHO60603.1	721	727	KSREETG
QHO60603.1	761	777	DLQPLEQPTSEAVEAPL
QHO60603.1	916	939	ASHMYCSFYPPDEDEEEGDCEEEE
QHO60603.1	966	1038	AALQPEEEQEEDWLDDDSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSGYLKLTDNVY
QHO60603.1	1088	1175	DYIATNGPLKVGGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTVRTNVYLA
QHO60603.1	1202	1229	IAEIPKEEVKPFITESKPSVEQRKQDDK
QHO60603.1	1271	1294	SDIDITFLKKDAPYIVGDVVQEGV
QHO60603.1	1549	1557	VITFDNLKT
QHO60603.1	1778	1794	FKKGVQIPCTCGKQATK
QHO60603.1	1934	1944	IKFADDLNQLT
QHO60603.1	2026	2060	VLKSEDAQGMDNLACEDLKPVSEEVVENPTIQKDV
QHO60603.1	2080	2082	NNS
QHO60603.1	2171	2190	FFTLLLQLCTFTRSTNSRIK
QHO60603.1	2210	2229	LEASFNYLKSPNFSKLINII
QHO60603.1	2596	2610	TFSSTFNVPMEKLKT
QHO60603.1	2782	2799	VAAIFYLITPVHVMSKHT
QHO60603.1	3051	3055	IVAIV
QHO60603.1	3139	3144	ITIAYI
QHO60603.1	3586	3611	ILTSLLVLVQSTQWSLFFFLYENAFL
QHO60603.1	4073	4086	IPDYNTYKNTCDGT
QHO60603.1	4174	4187	TKGGRFVLALLSDL
QHO60603.1	4390	4397	LQSADAQS
QHO60603.1	4453	4489	DDNLIDSYFVVKRHTFSNYQHEETIYNLLKDCPAVAK
QHO60603.1	4643	4672	TAESHVDTDLTKPYIKWDLLKYDFTEERLK
QHO60603.1	5130	5131	TD
QHO60603.1	5157	5172	FNSTYASQGLVASIKN
QHO60603.1	6052	6058	PNNTDFS
QHO60603.1	6144	6155	ASDTYACWHHSI
QHO60603.1	6417	6434	LYLDAYNMMISAGFSLWV
QHO60603.1	6458	6493	FNVVNKGHFDGQQGEVPVSIINNTVYTKVDGVDVEL
QHO60603.1	6542	6573	DAPAHISTIGVCSMTDIAKKPTETICAPLTVF
QHO60603.1	6603	6630	QPSVGPKQASLNGVTLIGEAVKTQFNYY
QHO60603.1	6652	6674	QEFKPRSQMEIDFLELAMDEFIE
QHO60603.1	6694	6715	SQLGGLHLLIGLAKRFKESPFE
QHO60603.1	7062	7086	GQINDMILSLLSKGRLIIRENNRVV
QHO60602.1	9	28	PFTIYSLLLCRMNSRNYIAQ
QHO60601.1	10	32	NAPRITFGGPSDSTGSNQNGERS
QHO60601.1	62	78	DLKFPRGQGVPINTNSS
QHO60601.1	216	233	AALALLLLDRLNQLESKM
QHO60601.1	401	409	DFSKQLQQS
QHO60600.1	9	43	ITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSK
QHO60600.1	103	108	FYEDFL
QHO60599.1	9	10	IT
QHO60599.1	71	73	KHV
QHO60599.1	94	111	ELYSPIFLIVAAIVFITL
QHO60598.1	42	48	SLTENKY
QHO60595.1	9	39	IGTVTLKQGEIKDATPSDFVRATATIPIQAS
QHO60595.1	89	126	VYSHLLLVAAGLEAPFLYLYALVYFLQSINFVRIIMRL
QHO60595.1	170	181	SGDGTTSPISEH
QHO60594.1	9	275	LVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFL
QHO60594.1	305	325	FTVEKGIYQTSNFRVQPTESI
QHO60594.1	345	371	RFASVYAWNRKRISNCVADYSVLYNSA
QHO60594.1	392	416	TNVYADSFVIRGDEVRQIAPGQTGK
QHO60594.1	437	531	SNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTN
QHO60594.1	553	574	ESNKKFLPFQQFGRDIADTTDA
QHO60594.1	605	725	NQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEI
QHO60594.1	871	883	QYTSALLAGTITS

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

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


% of ORF judged unique
0
10
20
30
40
50
60
Total number of amino acids in unique sequences
0
100
200
300
400
500
600
700
800
900
QHO60603.1
QHO60602.1
QHO60601.1
QHO60600.1
QHO60599.1
QHO60598.1
QHO60595.1
QHO60594.1
2019-nCoV Sequence