Final Revised ver#4 from June 7 2020 in response to peer reviewers at a journal Supplementary Information as PDFs can be supplied on request to corresponding author

Analysis of APOBEC and ADAR deaminase-driven Riboswitch Haplotypes in COVID-19 RNA strain variants and the implications for vaccine design

Edward J. Steele[1,2] and Robyn A. Lindley[2,3,4]

[1]CYO'Connor ERADE Village Foundation, 24 Genomics Rise, Piara Waters, 6112
[2]Melville Analytics Pty Ltd, Melbourne, Vic, AUSTRALIA
[3]GMDxCo Pty Ltd, Melbourne, Victoria, AUSTRALIA;
[4]Department of Clinical Pathology, The Victorian Comprehensive Cancer Centre, Faculty of Medicine, Dentistry & Health Sciences, University of Melbourne, Victoria, AUSTRALIA.

Running head: *Analysis of Deaminase Signatures in COVID-19*

Correspondence: A/Professor Edward J Steele, Melville Analytics Pty Ltd, 162 Collins Street, Melbourne 3000, email: e.j.steele@bigpond.com

Key words: COVID-19 genomes; Coronavirus pandemic; Single Nucleotide Variations; Cytosine and Adenosine Deaminations; AID/APOBEC and ADAR Deamination Motifs

Author Information:
Edward J. Steele e.j.steele@bigpond.com
Robyn A. Lindley robyn.lindley@gmdxgroup.com

Abbreviations

ADAR, Adenosine Deaminase that act on RNA, two main isoforms, ADAR 1, ADAR 2 mediating adenosine-to-inosine **A-to-I**) mutation predominantly seen in RNA editing in Innate Immunity to viruses; **APOBEC family**, generic abbreviation for the deoxyribonucleic acid, or dC-to-dU, deaminase family (APOBECs 1, 2,4 and 3A/B/ C/D/F/G/H) similar in DNA sequence to the "apolipoprotein B RNA editor" APOBEC1, and known to activate mutagenic cytidine deamination during transcription in somatic tissues, particularly in cancer and Innate Immunity to viruses; **Deaminase**, zinc-containing catalytic domain in ADAR and APOBEC enzymes; **MC**, mutated codon; **MC1, MC2, MC3**, respectively refer to the first, second and third nucleotide mutation target position within a mutated codon read in the 5-prime to 3- prime direction; **R**, Adenosine (A) or Guanine (G) , purines; **S**, strong base pair involving Cytosine (C) or Guanine (G); **SNV**, single nucleotide variation; **T**, Thymine; **TSM**, targeted somatic mutations : the process of targeting actively transcribed genes that results in a dominant type of mutation caused by a DBD or Inf-DBD targeting nucleotide sites at a particular codon position; **U**, uracil; **W**, weak base pair involving A or U/T; **Y,** pyrimidines T/U

Abstract

This paper reports the results of our initial analysis of APOBEC and ADAR deaminase-mediated mutation signature patterns in complete COVID-19 genomes from informative locations and times in China, USA and Spain in the 2019-2020 pandemic. We have identified a unique set of 'new' putative coordinated Riboswitches in COVID-19 genomes not previously identified, and likely generating variants of the known common strain Haplotypes now in circulation. The results reveal that COVID-19 diversifies using switching of RNA Haplotypes with minimal alteration to protein structure (the normal targets for B and T cells in conventional vaccine development). The deaminase-driven RNA Haplotypes are most likely aligned with RNA secondary structures as several studies already highlight how Riboswitches alter the ability of RNA to fold into intricate three-dimensional structures allowing them to execute their diverse cellular functions. The same functional outcomes are expected for viruses, particularly efficacy of RNA replication in new host cell environments. Thus, vaccine designs that assume that the main viral protein antigens will be the only putative protective targets could fail to produce effective and protective immunity. We conclude that understanding COVID-19 adaptation and survival strategy and identifying the host Haplotype, and which vaccine(s) is effective for each Haplotype group will be important for new vaccine design. Our study also has wider implications for the actual origins and spread of COVID-19 but these will be pursued in future publications.

INTRODUCTION

There have been a number of recent reports that are attempting to understand the mode of transmission and pathogenesis of COVID-19 (reviewed in Shi et al 2020). However our purpose in this paper has been to analyse the genetic structure of COVID-19 genomes isolated early in the first two and a half months of 2019-2020 coronavirus pandemic : from its sudden explosive origins in China (Dec 2019-Jan 2020), through early spreading to the West Coast USA (through late Jan , then Feb-to early Mar 2020) to two other informative explosive

epicenters of the pandemic in Spain and particularly New York, USA (through to March 14-22, 2020). In advance we decided the sequence data to be analysed had to be readily accessible online to other scientists, be authoritative, be curated by experts and thus accurate and reliable with GenBank linkage. We thus chose to focus exclusively on those COVID-19 sequences curated at the NIH , Bethesda, through its National Center for Bioinformatic Information, at their site *NCBI Virus*. In following this approach our aim has been to understand the adaptive genetic strategy of the virus through its deployment of putative riboswitch changes to its RNA sequence to generate robust RNA haplotypes able to replicate in the genetic background of the infected host.

Previously we applied our analyses of APOBEC (C>U) and ADAR (A>I) deaminase-mediated editing signatures in the viral RNA genomes of HCV and ZIKV during the acute- phase of innate immune responses of the host-parasite relationship (Lindley and Steele 2018); that is during the well-known host phase of the interferon-stimulated gene (or innate immune) response (Schoggins and Rice 2011, Schneider et al 2014). We reported

Table 1 Time-Line COVID-19 Patient Sample Collections and Locations

Location	Collection Dates	No. Complete COVID-19 Genomes Aligned v Hu-1 Ref.Seq.	Type of COVID-19 Outbreak
China -Wuhan and Regions	Dec 20-30 2019, Jan 2020, Feb 2-5 2020	47	Originating explosive epidemic in China centred om Wuhan, Hubei Province and immediate regions
California,USA	Jan 22- Feb 26 2020	10	Early sporadic reports and person-to-person (P-to-P) community spreads
Cruise Ship-at-Sea *Grand Princess* off CA coast,USA	Feb 17-25 2020	25	Localised yet explosive outbreak at-Sea
California,USA	Feb 27-Mar 4 2020	24	Local outbreaks and community P-to-P spreads
Washington State (Kirkland), USA	Feb 24-Mar 1	10	Local outbreak in nursing home
Spain	Feb 26- Mar5 2020,	8	Local outbreak just prior to explosive increase in Spanish cases
Spain	Mar 6- Mar 10 2020	13 +	Exponential increase phase of Spanish epidemic
New York , USA	Mar 5 - 9 2020	18	Local outbreaks and community P-to-P spreads just prior to explosive NYC outbreak
New York (overwhelmingly NYC), USA	Mar 14 - 22 2020	97	Explosive exponential phase of NYC epidemic

that the distinct signatures at known deamination motifs of cytosine to uracil (C>U read as C>T) and adenosine to inosine (A>I read as A>G) are written into the circulating viral genomes including the quasi-species of viral

variants in an individual during *Flavivirus* infection (Stoddard et al 2015). We also critically reviewed the literature showing that viral replicases themselves, the RNA-dependent RNA polymerases (RdRP), are of high replicative fidelity thus faithfully copying the deaminase-mediated mutation patterns into replicating viral progeny genomes. We concluded that this contributes to the production of the viral quasi-species observed during the acute phase of HCV disease *in vivo* (Stoddard et al 2015).

Flaviviruses possess positive single stranded RNA genomes about two thirds smaller in length than COVID-19. We now apply this same targeted somatic mutation (TSM) codon-context methodology (Lindley 2013, Lindley and Steele 2018) to the analysis of the positive single stranded COVID-19 genomes collected from patients in China, USA and Spain during the acute phase of the infection. There is now a large amount of sequence data curated at the NIH website dedicated to this virus, "NCBI Virus" (particularly for the USA, lesser extent China and very little from other countries at time of initial writing). We have, by necessity, been focused and selective as our analyses are of a different type to conventional algorithm-dependent phylogenetics which focus on global strain features (Dorp et al 2020 a,b) and which may overlook some of the important features we report here. Indeed, in our analyses there is no algorithm interface between the patterns seen and the interpretation of the same data patterns (apart from the onscreen NCBI Virus alignment tool used). We concentrate on the mutational source of trusted, reliable widely accessible data viz. GenBank accessible NCBI-curated single nucleotide variants (SNVs) at the National Institutes of Health (NIH) creating the observed genetic patterns in isolated viral genomes from infected subjects during the innate immune response (which would be the first week of disease in most patients swabbed for the virus). Thus, we need to be selective so as to allow the maximum insight into the origin (source) and spread of this newly emergent viral disease (Table 1). Apart from region and country of origin, the NIH curated COVID-19 sequences~~,~~ lack detailed patient data, age, sex, racial origin, and clinical co-morbidities and clinical outcome. We thus chose to ~~make~~ analyse sequence alignments for viral collections during the key early two and a half months of the pandemic at informative times and locations (Table 1). In the cases of Spain and New York we focused particularly on the mid-point of the exponential rising case curve from about March 14 to the end of the month in New York and March 6 -10 in Spain *versus* the COVID-19 genetic patterns from the isolates in China in January 2020 (Figure S1). This report will thus focus on comparative Variable Site Diagram (VSD) patterns across full length COVID-19 genomes (29903 nt Hu-1 ref. seq) and will be very selective. A breakdown by critical time points and regions during the early periods of the global pandemic is displayed in Table 1 (and case incident statistics from online media sites with time in the Supplementary Information FiguresS1-S6)). Future reports will analyse over 12,000 COVID-19 genomes(from non-NIH curations GISAID, Next Strain etc) to focus on a more detailed analysis to identify specific types of APOBEC and ADAR deaminases executing the observed mutational events and to provide

further insights into riboswitch adaptions as a COVID-19 mechanism to evade host innate and adaptive immunity. Later papers will also deal with the further implications of these data for the origin and global spread of this suddenly emergent pandemic disease.

MATERIALS AND METHODS

Data Source and Acquisition

The National Centre for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) curates as they come available all current complete and partial SARS-CoV-2 sequences at https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/#nucleotide-sequences particularly at the NCBI Virus site for this virus (at URL https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049). Complete genome COVID-19 sequences isolated during the phases of the epidemic can be selected and aligned with tools provided at the NCBI Virus site.

Details on Sample Access, Time Points and Regions

The collection time, for each sequence identified by GenBank Accession number is provided in the detailed major tabulated curations in Table 1 and online supplementary information Tables S1, S2, S3, S4. In some latter stages of the analyses we resorted to a screen shot record of key alignments and manual tabulation, reporting only key VSD patterns in those figures and tables discussed.

To summarise Table 1: for China, collections were from December 20 through February 5, but the great bulk of collections were through January (TableS1, Figures S1,S2). For West Coast USA during early sporadic outbreaks in that country, mainly California and the off-coast cruise ship (*Grand Princess*) collections were from January 22 through February 24, then February 27 through March 4 (TableS2, FigureS3, S4). An outbreak in an old person's hospital facility Washington State (Kirkland) were collections February 24-March 1 (TableS3, see FigureS5). For Spain there were two periods of collections examined February 26-March 5, then March 6-March 10. For New York sequence alignments were conducted on pre-epidemic collections March 5-9, then at the midpoint of the exponential rising case curve for March 14-22 in the USA which was almost exclusivel diue to thew exponential increases of COVID-19 cases in New York City (FiguresS1, S3, S5).

The analyses of data in **RESULTS and DISCUSSION** follows this chronology, to reflect more or less the order of reported key early outbreaks across the world in January, February to March14-22, 2020. However, it is likely that both Spain and USA (which is overwhelmingly data from NYC , Supplementary Information) show

very similar and overlapping case increase curves both in time and slope for March (FigureS1). The order of Spain before New York reflects the temporal order of the outbreaks reported in the mass media (following the slightly earlier explosive occurring outbreaks in Tehran/Qom and Lombardy in Italy – the latter not analysed as no significant numbers of complete genome data from these locations had been uploaded to NCBI Virus site at time of data analysis and writing of this paper).

NCBI Virus Genome Sequence Alignment and Analysis

Our data analyses involves the following steps:

1.At the NCBI Virus site for SARS-CoV-2, all selected sequences were analysed while recording the sample country, region and time period of collection in the pandemic.

2. An alignment of the selected sequence set, including the original Wuhan Hu-1 reference sequence (NC_045512.2) was made using the on-screen tools at the NCBI Virus site.

3. In the Tabulation into the excel spread sheets (TableS1, S2, S3, S4), each sequence is linked to Sequence ID or Accession Number (to GenBank), its Collection date, its Release date, on occasion the Length of the curated complete sequence (although that is in GenBank), and the Country of origin, and region where possible.

4. Each single nucleotide variation (SNV) from the Wuhan Ref. Seq. (Hu-1, NC_045512.2) was curated by position in the Multiple Alignment – viz. position in the 5' untranslated region (UTR), the protein coding (CDS), non-CDS gaps, and the 3' UTR. Most alignments gave exact sequence positions for key SNV sites, although the China alignment (TableS1) sequence positions at S Haplotype defining sites p. 8782 and p. 28144 are advanced by three to p.8785, and p.28147. Other adjustments, in our analysis, for in-frame whole codon deletions in the aligned collection were noted at times for p.11080/83 and p.25563/60. Suspect sequences in the 5' UTR or 3'UTR possibly due to sequencing technical artefacts were noted, and SNVs adjudged as genuine or not in those and other ambiguous regions (N). N runs and sequence quality were noted and reported in summary VSD patterns where appropriate as a guide to sequence quality for that alignment. Also noted were samples with truncated 5' and 3' UTR ends as these could be cause the loss of key information, such as putative "riboswitch" sites in these regions.

.

5. Each curated SNV in the protein coding regions (CDS) was then classified by

a. The CDS SNV type such as C>T, G>A, A>G, T>C, G>C, G>T etc. C/G-sites are implied as APOBEC changes, A/T-sites are implied ADAR mediated changes (Lindley and Steele 2018). These will be further analysed in detail in a follow-up paper (Hall, Mamrot, Steele, Lindley In Preparation). In some cases, G>T SNVs were deduced to be more likely caused by reactive oxygen species (ROS) producing 8oxoG modified guanosines at that site. In these cases the 8oxoG modifications may well be preferred at WG sites as noted previously for cancer genomes viz. the single base substitution (SBS) signature describing this pattern is Signature 18 of Alexandrov et al (2013) and see the COSMIC website for all updated mutational signature information, https://cancer.sanger.ac.uk/cosmic/signatures . We have recently explained the rationale for diagnosis of this ROS signature in Franklin et al (2020).
b. The likely strand was identified on which the deamination event occurred: the +ve sense strand for mRNA or –ve template strand for replication of COVID-19 sequence copies) in the dsRNA Replicating Form (RF) of the virus in the putative membraneous web in the cytosol (Thimme et al 2012, Yang and Leibowitz 2015).
c. The codon context (Lindley 2013, Lindley et al 2016, Lindley and Steele 2018) of the change viz. whether in the MC1, MC2, MC3 positions or first, second and third positions of the mutated codon (by convention read 5 prime to 3 prime to allow subsequent assignment of specific codon-context deaminase associated mutation signature and motif location assignment (In Preparation).
d. The nature of the amino acid (AA) change and whether that SNV in the protein is "Conserved", "Benign" or "Radical" in its putative change of protein secondary structure and function. All nonsynonymous changes within an AA functional class are considered, like synonymous changes, as "Conservative" (black in VSD patterns). However, by definition, all observed SNVs are likely to be "benign' in terms of their likely impact on viral protein structure and replicative ability of the RNA viral genome – since the variant virus sequence has already made the "Darwinian Cut". However in this qualitative scheme a "likely benign" nonsynonymous change (green in VSD patterns) would be AA interchanges for Polar<->NonPolar, Basic<->Polar, Acidic<->Polar. A Radicle change is a full AA charge change "Basic<->Acidic" and Basic<->NonPolar, Acidic<->NonPolar (deep red in VSD patterns). In the various Variable Site Diagrams (VSD) in **RESULTS and DISCUSSION** the following colour codes and qualifications for entries are shown in Figure 1.

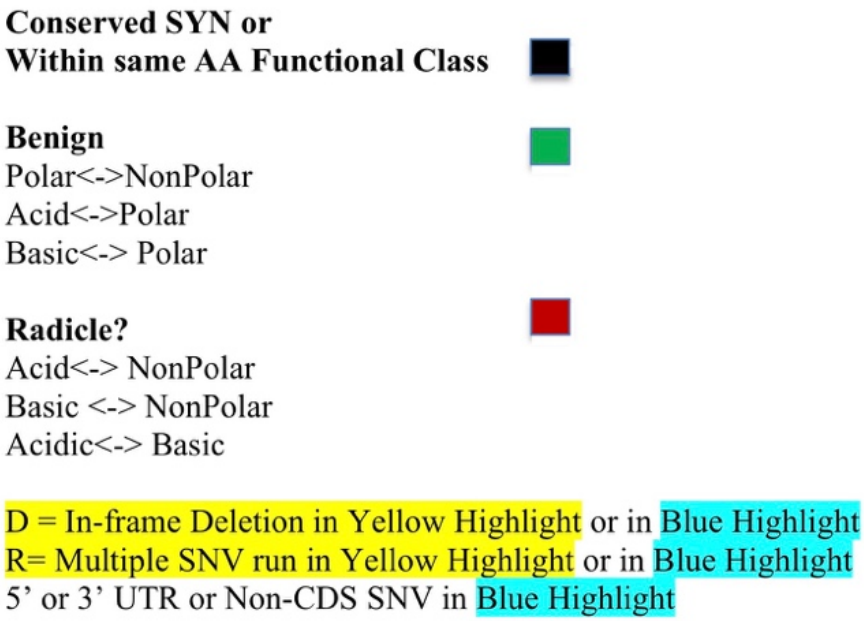

Figure 1. SNV Colour Codes for Variable Site Diagrams

6.Throughout the CDS regions, the SNVs were also analysed in terms of likely change to RNA secondary structure based on the SNV's conserved nature at the protein level (as defined, Figure 1) and whether two or more apparent co-ordinated SNV changes are required in presumptive "Haplotype" generation. As explained below we consider this could be a reflection of putative co-ordinately deaminase-targeted "Riboswitch" positions (e.g. as reviewed in Yang and Leibowitz 2015). If they occur frequently, in independent collections from different regions, and are apparently independent sequences, they were noted and the Haplotypes they appear associated with were tabulated and factored into the analysis of sequences and their mutated derivatives (Table 2). The literature on Riboswitches, RNA secondary structure and associated changes in cellular functions has been well documented (Gilbert and Fontane 2006, Tan et al 2015, Widom et al 2018).

RESULTS AND DISCUSSION

1.-A Rationale for Ordering the Data on COVID-19

Table 2 summarises our collected observations for sequences analysed from different regions. It is provisional as it may well be revised with additions, or as qualified deletions as more sequence data and patterns emerge. It shows our current assessment of Haplotypes and coordinated Riboswitch SNVs mainly at the RNA level (not predominantly at the protein level) which we consider useful in our analyses. The Colour coding in the table focuses attention on the distinction between L and L-241 RNA haplotypes as revealed between the Wuhan and

New York COVID-19 collections. In other targeted mutagenesis studies numerous RNA secondary structure variant hotspots have been revealed related to efficacy of the replicative phases of the HCV viral life cycle and other translated genes (Pirakitikulr et al 2016; also see Buhr et al 2016, Widom et al 2018), and, as indicated SARS-CoV-1 appears to have also deployed an RNA secondary structure polymorphic adaptation strategy

Table 2 Haplotypes and Sites Defining COVID-19 Common Strain Variants in China, USA, Spain

	AA class->	P<>NonP	SYN	SYN	NonP<>NonP	NonP<>NonP	P<>NonP	NonP<>NonP	P<>P	NonP<>NonP	P<>P	SYN	NonP<>NonP	Acid<>NonP	P<>Basic	NonP<>NonP	NP<>NP	P<>NonP	SYN	P<>NonP	near 3'UTR
	5'UTR	Thr<>Ile	Phe<>Phe	Ser<>Ser	Phe<>Tyr	Leu<>Phe	Ser<>Leu	Pro<>Leu	Tyr<>Tyr	Pro<>Leu	Tyr<>Cys	Leu<>Leu	Ala<>val	Asp<>Gly	Gln<>His	Gly<>Val	Gly<>Val	Leu<>Ser	Asp<>Asp	Ser<>Leu	non-CDS gap
HAP	p.241	p.1059	p.3037	p.8782	p.9477	p.11080/83	p.11916	p.14408	p.14805	p.17747	p.17858	p.18060	p.18998	p.23403	p.25563	p.25979	p.26144	p.28144	p.28657	p.28863	p.29540
L (Hu-1)	C	C	C	C	T	G	C	C	C	C	A	C	C	A	G	G	G	T	C	C	G
Ln	C	C	C	C	T	T	C	C	T	C	A	C	C	A	G	G	T	T	C	C	G
L-241a	T	T	T	C	T	G	C	T	C	C	A	C	C	G	T	G	G	T	C	C	G
L-241a.1	T	T	T	C	T	G	T	T	C	C	A	C	T	G	T	G	G	T	C	C	A
L-241b	T	T	C	C	T	G	C	T	C	C	A	C	C	G	T	G	G	T	C	C	G
L-241c	T	C	T	C	T	G	C	T	C	C	A	C	C	G	G	G	G	T	C	C	G
L-241d/s	T	T	T	C	T	G	C	C	C	C	A	C	C	G	T	G	G	T	C	C	G
L-241e	T	C	C	C	T	G	C	T	C	C	A	C	C	G	T	G	G	T	C	C	G
L-241f	T	C	T	C	T	G	C	T	C	C	A	C	C	G	G	G	G	T	C	C	G
L-241g	T	C	C	C	T	G	C	T	C	C	A	C	C	G	G	G	G	T	C	C	G
S	C	C	C	T	T	G	C	C	C	C	A	C	C	A	G	G	G	C	C	C	G
Sa	C	C	C	T	T	G	C	C	C	T	G	T	C	A	G	G	G	C	C	C	G
Sb	C	C	C	T	A	G	C	C	T	C	A	C	C	A	G	G	G	C	C	C	G
Ss	C	C	C	T	A	G	C	C	T	C	A	C	C	A	G	T	G	C	T	T	G

(Yang and Leibwitz 2015). The most notable example is the L *v* S Haplotypes as revealed first in the China data by phylogenetic relationships with SARS-CoV-1 and apparent animal variant relationships (Tang et al 2020). The current simple sequencing methods thus identify the different haplotype-defined RNA strains of L and S and their subtypes. This depends first on detecting a C at p.8782 and T at p.28144 thus identifying the L haplotypes and a T at p.8782 and a C at p.28144 thus identifying the S haplotypes. Another key site is the p.241 C>T in the 5'UTR- other, RNA-only and thus putative Riboswitch sites are presented in Table 2. No other strain RNA haplotypes have been identified using such binary (or even higher number) sequence tests. The L/S test cannot define changes implicating putative RNA secondary structure modifications in currently arising circulating strains. This is the utility of the putative riboswitched haplotypes arrayed in Table 2 – other strains can be haplotyped and much of the current sequence diversity in COVID-19 can be identified and understood in terms of haplotype diversification during global spread of the disease.

Deaminase-Driven Riboswitch Hypothesis: Haplotype variation in the initial first infection?

In support of this interpretation of the data is the fact that the two SNVs defining the S Haplotype are rarely observed by themselves - thus at the canonical S defining site p.8782 the C>T (= p.8785 in the current China alignment TableS1, is MC3, TTT AG**C** CAG) is always paired with the S defining canonical site p.28,144 of a T>C (= p. 28,147 in current Table S1 alignment MC2 TGT T**T**A CCT). However, these criteria for defining that haplotype might also apply to the other putative haplotypes identified in Table 2. Thus, it can be inferred that the COVID-19 viral diversification strategy is locked into the productive and coordinate combinations of *RNA Riboswitch modifications* which logically implies RNA secondary structure with downstream affects on

function and replication. Thus, the simplest and the most parsimonious interpretation of the data assumes the L-to-S Haplotype variations, and the others listed in Table 2 (L-to-L-241 variants), are largely deamination-driven *in vivo* during the first infection cycle by unmutated source viruses e.g. L Hu-1. That is, the variation is not expected to pre-exist in the initial source virus population prior to first infection – it makes better biological sense that the haploytype switch actually occurs during the innate immune response in the first infection cycle in that subject. Accordingly, in our view, the host-parasite interaction, and the innate immune response to the virus, ultimately determines the observed haplotype that emerges in the complete COVD-19 genome. In our observations the proportion of S Haplotypes to total sequences in any given collection alignment can range from 5-50% (for example in 206 NY samples March 14- 22, 6 sequences are S, and 10 are L, below). We have not generally observed haplotype recombinants (on scale) at this stage of our survey – although we expect it to occasionally occur. In a small number of cases a SNV site can be shared between haplotypes (p.14805, C>T site), indicative of deaminase (or reactive oxygen species mutagenesis, ROS, on oxidised guanosines, 8oxoG) activity at that site viz. it could be a hot spot for mutagenesis. If key SNVs defining that haplotype are reverted to Hu-1 reference sequence they are rare (although some undoubtedly occur on inspection of data sets, see below). It is possible that novel conversions can take place on further person-to-person transmissions. However, given that APOBEC deaminations (C>T) can in theory be reverted by ADAR deaminations (T>C) such reversions must be considered as possible, and may often occur during the innate immune response of that infection cycle in that subject. Also, important non-CDS RNA only regions, like the G>A SNV in the non-CDS gap at p.29540 may contribute to additional haplotypes in a wider data set (as that data seems to imply for the L241a.1 subhaplotype, below).

Each of the SNV-defined haplotypes identified comprises approximately 0.02% difference from the Hu-1 reference sequence. On average there are approximately 5 SNV differences from Hu-1 defining each haplotype. Thus, there is ≥99.98% identity between any haplotype and the Wuhan reference sequence whether that sequence is collected in China, Spain, the US West Coast or New York City.

So this is our operating hypothesis: the germline encoded innate immune responses in the first day or two after infection with, for example, source Hu-1 virions (L) can generate deaminase-mediated C>U and A>I changes in the replicating viral sequences, and less frequent down-stream miscopied transversions (e.g. opposite inosine template residues). Thus, a range of +ve strand RNA quasi-species are produced in an infected cell with changes at particular deaminase hot spots or riboswitch sites determining compatible RNA secondary structures

allowing rapid replication. Host-directed deaminase-mediated riboswitches are expected to create adaptive options for the virus which if then selected allows more rapid replication in that cellular environment. This hypothesis is a great simplification conceptual tool, and it has allowed us to order the complex data sets now emerging in the pandemic in a rational way.

2.-Analysis of China COVID-19 complete genomes

All China COVID-19 sequences collected from patients during December 2019 into January 2020 (to Feb 5) were selected into the alignment during a period of explosive exponential increases in COVID-19 cases in Hubei province, particularly its major city Wuhan (FiguresS1,S2,S3). These numbers account for ≥90% of all the China COVID-19 cases (and deaths) reported. However, the sequences curated at NCBI Virus do not reflect that case bias, as surrounding regions and provinces are over-represented in the collected sequence-set compared to density of case incidence as shown in FigureS2.

Table 3a. Types of SNVs observed in China outbreak

REF Base	Variant Base					
	A	T	C	G	TOTAL	%
A		1	1	**7**	9	22.5
T	3		**4**		7	17.5
C	2	**14**			16	40
G	**4**	3	1		8	20
					40	

Transitions - 72.5% Transition: Tranversion ratio 3.4:1
Transversions - 27.5%

Note MT226610 and MT019530 data excluded

Table 3b. Types SNV in California + Cruise Ship Outbreaks Jan 22- Feb 24

REF Base	Variant Base					
	A	T	C	G	TOTAL	%
A		4		**2**	6	10.7
T			**4**		4	7.14
C	1	**29**			30	53.6
G	**6**	7	3		16	28.6
					56	

Transitions - 73.2% Transition: Tranversion ratio 2.73:1
Transversions - 26.8%

Table 3c. Types SNV in CA outbreaks Feb 27-Mar 4

REF Base	Variant Base					
	A	T	C	G	TOTAL	%
A			1	**6**	7	23.3
T			**4**		4	13.3
C	1	**11**			12	40
G	**3**	4			7	23.3
					30	

Transitions - 80% Transition: Tranversion ratio 4:1
Transversions - 20%

Table 3d. Types SNV in Spain Outbreaks Mar 6-Mar 10

REF Base	Variant Base					
	A	T	C	G	TOTAL	%
A				**4**	4	14.8
T	1		**2**		3	11.1
C		**12**			12	44.4
G	**2**	6			8	29.6
					27	

Transitions - 74.1% Transition: Tranversion ratio 2.86:1
Transversions - 25.9%

Table 3e. Types of SNV in NY Outbreaks Mar 5-9

REF Base	Variant Base					
	A	T	C	G	TOTAL	%
A		2		**6**	8	20
T	1		**2**		3	7.5
C	1	**19**			20	50
G	**3**	6			9	22.5
					40	

Transitions - 75% Transition: Tranversion ratio 3:1
Transversions - 25%

Table 3f.Types of SNV in L and S Haplotypes collected NYC Mar 14-Mar 22.

REF Base	Variant Base					
	A	T	C	G	TOTAL	%
A		2		**4**	6	34.92
T	2		**5**		7	17.46
C		**16**		1	17	17.93
G	**4**	6	2		12	29.77
					42	

Transitions - 69% Transition: Tranversion ratio 2.22:1
Transversions - 31%

Table 3g. Types of SNV in collected NYC Mar 14-Mar 19.

REF Base	Variant Base					
	A	T	C	G	TOTAL	%
A		1	2	**11**	14	34.92
T	3		**10**		14	17.46
C	1	**41**		1	43	17.93
G	**9**	11	3		23	29.77
					94	

Transitions - 75.5% Transition: Tranversion ratio 3.08:1
Transversions - 24.5%
Note: 7 of 11 G>T are potentiall 8oxoG modifications at W G sites

Table 3 Types of SNVs in the different outbreaks analysed.

Caveat on all analyses in this paper

Apart from this type of bias in sequence selection, there is a major caveat on all other hidden biases present in the aligned data. The clinical decision to seek sequence information on the collected COVID-19 sample assumes that the patient had full blown disease symptoms with respiratory complications (in the main). All other analyses (below) on USA and Spain data lack specific clinical information and the interactive relationships between patients with putative sequences (Sequence IDs or GenBank Accession Numbers), the subject's age, sex, racial origin (Caucasian, Asian, African-American, Latinos etc). Absent here then is the known prior state of health (co-morbidities) or whether the subject survived the acute respiratory infection. In some cases, such as observations on specific outbreaks, we can make inferences because of the location and timing and person-to-person transfers (P-to-P, e.g. hospital outbreaks in Kirkland in Seattle, Washington State) – but that is all they are. In the case of China and Spain we can safely infer dominant Chinese ethnicity and Caucasian/Latino ethnicity of patients (in the main). However we lack key information in a putative P-to-P spreading chain such as "..who actually gave the virus to whom? ..." – that data must exist in some form, somewhere, but we do not have that data. And "community spreads" without a controversial "known link to China" is a background factor in trying to interpret P-to-P spreading. We can only plausibly infer P-to-P transfers from the mutation patterns in the sequence data. But we can say with confidence (see Figure 2, Table 3) as we did earlier for acute HCV, ZIKV infections (Lindley and Steele 2018), that the great bulk of SNVs analysed are at APOBEC (C-site) and ADAR (A-site) deamination motifs viz. APOBEC1 and ADAR1/2 deaminases (Rosenberg et al 2011, Lindley and Steele 2018). They appear to be the responsible drivers of the mutations - including the causative deaminases most likely to drive riboswitching at simultaneous (linked) deamination events by APOBEC/ADAR at functionally-coupled C-site and A-site hotspots. However, motif specificity of other APOBEC RNA C-site editors such as APOBEC3A (Sharma et al 2015, 2016a) and APOBEC3G (2016b) should also be searched for in these sequence data during acute phase COVID-19 infections, and that analysis is underway.

Variable Site Diagram (VSD) of 47 China COVID-19 Sequences (Figure 2)

This is a valuable and informative way to present the SNV data and make logical inferences on the genesis of mutational patterns and relations among sequences. Such patterns, we believe, are far more informative at the molecular and cell biology level than simple construction of phylogenetic trees – the P-to-P issue of "who gave what viral variant to whom" is a far more relevant genetic question in connecting apparently different sequences.

This variable site diagram (VSD) is displayed in Figure 2 for the 47 complete China COVID-19 genomes Dec 20 2019 through January 2020. There were originally 48 selected sequences in the alignment. Sequence

LR757997 however had to be removed as there were far too many N runs and other sequence gaps that created real problems not only for a respectable alignment but also in alignment scrolling and analysis. This sequence was thus deleted leaving 47 compete COVID-19 sequences for analysis during the height of the China epidemic.

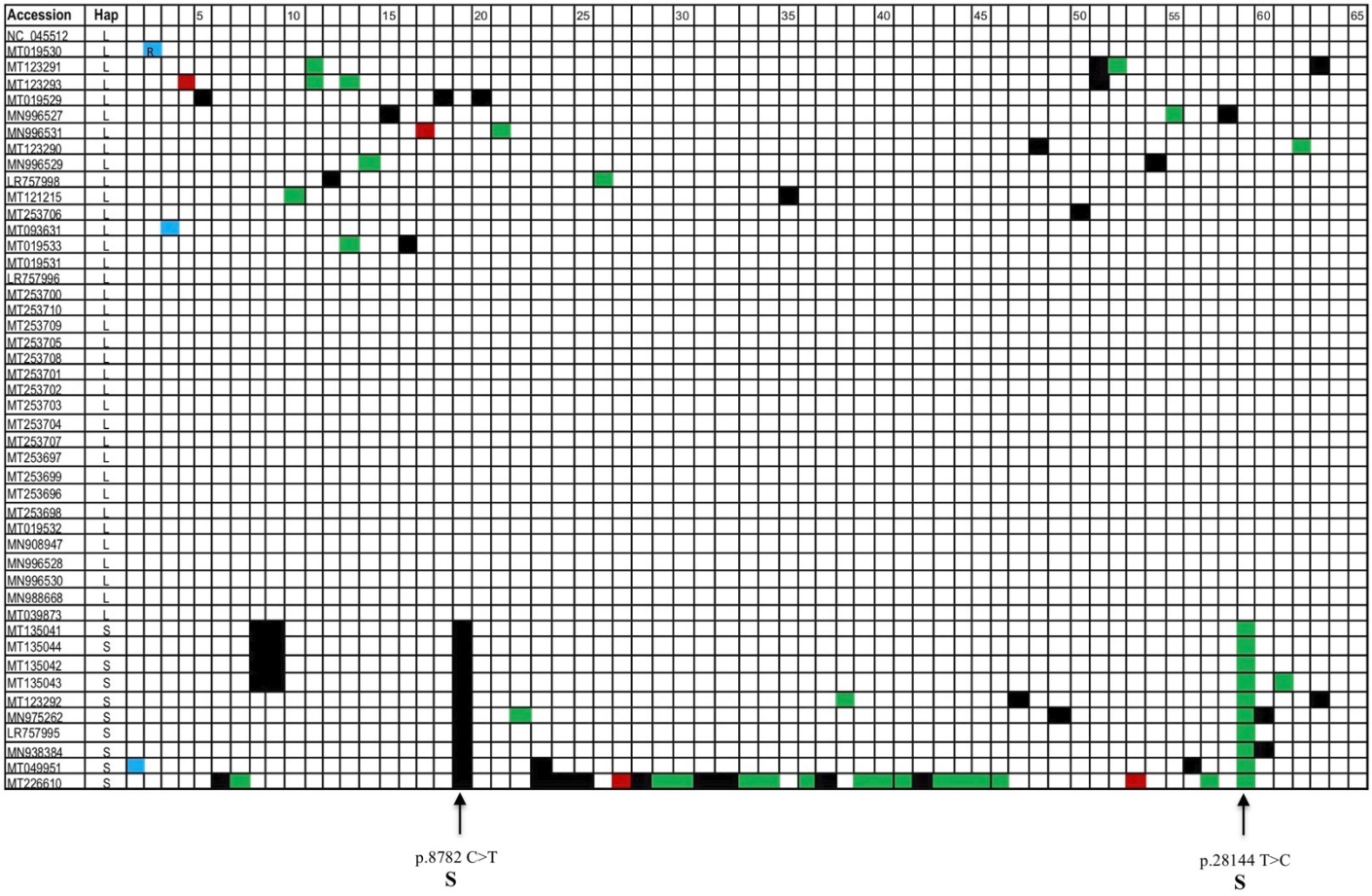

Figure 2 Variable Site Diagram of SNVs in each aligned sequence in the 47 China sequence alignment, which includes Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype assignment. Note MT226610 has 27 SNVs and is discussed separately in text. MT019530 may have corrupted sequence in 5' UTR (site 2) but included as identical in rest of sequence to Hu-1 reference. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. S Hap sites are indicated for sites 19 and 59 and arrowed (and see Table 2). The data in TableS1 should be consulted for further details. The variable site column number followed by SNV position in the alignment are : 1, p.76, C>A; 2, p107-127, T>A,T>C,T>G, C>G, T>C, G>A; 3, p.189, C>T; 4, p.657, G>A; 5, p.3781, A>G; 6, p.4291, G>T; 7, p.4310, A>C; 8, p.4405, T>C; 9, p.5065, G>T; 10, p.6029, C>T; 11, p.6822, G>T ; 12, p.6971, C>A; 13, p.6999, T>C; 14, p.7019,G>A; 15, p.7482, A>G; 16, p.7869, G>T; 17, p.8004, A>C; 18, p.8391, A>G ; 19, p.8785, C>T; 20, p.8890, T>A; 21, p.9537, C>T; 22, p.9564, C>T; 23, p.11086, G>T; 24, p.11210, G>C; 25, p.11236, T>G; 26, p.11767, T>A; 27, p.12044, G>C; 28, p.12163, G>C; 29, p.12205, G>C; 30, p.12211, G>T; 31, p.12358, G>C; 32, p.12381, G>A; 33, p.12467, G>T; 34, p.12470, G>T; 35, p.12476, C>T; 36, p.12494, G>T; 37, p.12517, G>C; 38, p.12537, C>T; 39, p.12575, G>T; 40, p.12581, G>A ; 41, p.12585, G>T; 42, p.12603, G>A; 43, p.12663, G>C; 44, p.12688, G>C; 45, p.12776, G>T; 46, p.12796, G>T; 47, p.13075, C>T; 48, p.15327, C>T; 49, p.15610, T>C; 50, p.162250, C>T; 51, p.17376, C>T; 52, p.19613, C>T; 53, p.20983, G>C; 54, p.21140, A>G; 55, p.21319, G>A; 56, p.21647,T>A ; 57, p.21787, T>A ; 58, p.24328, A>G; 59, p.28147, T>C; 60, p.29098, C>T; 61, p29304, A>T; 62, p.29306, C>T; 63, p.29530, G>A.

Overview of Mutation Pattern of 47 China COVID-19 Sequences

The most striking general patterns displayed by these data (and the California and Cruise Ship SNV data, Figure 3) are their resemblance to the similar variable site patterns seen *in vivo* among the viral quasi-species (Eigen and Schuster 1975, Andino and Domingo 2015) of HCV patients during the acute phase of HCV infection (first week or so) - as seen in the single molecule HCV sequencing of a number of such patients reported by Stoddard et al (2015). Indeed, quasi-species acute phase HCV data were used by us in the *Flavivirus* analysis previously reported (Lindley and Steele 2018). This raises the whole issue of exactly "What a COVID-19 RNA consensus sequence actually is?" Thus, future deep single-molecule sequencing should be conducted on separate COVID-19 swab or bronchial fluid collections from the *same* patient (Li et al 2012) to establish a more realistic assignment of the "consensus" sequence in some patients. The acute phase "quasi-species" like pattern – is distributed in many subjects in the Chinese COVID-19 patient population, rather than as assessed in a single patient *in vivo* by deep sequencing. The same general pattern is evident in the California and Cruise Ship data (Jan 22- Feb 24, below), and for the dominant haplotypes in the explosive New York epidemic (Mar 14- Mar 22).

In Figure 2. there are 63 variable sites. For the CDS region there are 60 variable sites. Of the three sites in the 5'UTR two look legitimate #1 and #3: MT049951 C>A at p.76, MT093631 C>T at p.189. The third #2 is MT019530 and involves a cluster of 6 changes from Hu-1, most are T>C (or T>Y) and could be sequencing artefacts. Our judgement is these sequence sites be ignored, but the CDS region SNVs in MT019530 be kept in the analysis as a legitimate unmutated derivative of the Hu-1 reference sequence. Of the 46 sequences 36 are of L Haplotype and 10 are of the S Haplotype as defined Table 2.

The types of SNVs are displayed in Table 3a. Mutations at C-sites exceeds mutations at A-sites, with an excess of C>T suggesting APOBEC C>U deaminations mainly on the +ve strand either in completed COVID-19 genomic copies or in the single stranded regions of the displaced +ve strand at Transcription Bubbles during replication. ADAR A>I events occur equally on both the +ve and -ve strands suggesting A>I events on dsRNA regions of the RF form as well as in completed stem loops of completed +ve strand copies. The number of transition mutations exceeds the number of transversions more than three to one (an expectation in all deaminase-driven mutagenic systems, see Steele and Lindley 2017).

L Haplotype Analysis of the China data

Among sequences in the major L Hap group there are 26 unique variable sites (MT226610 sites are ignored- as these are all in S Hap).

• 25 L sequences are identical to Hu-1 ref i.e. unmutated

• 11 L sequences were variable from Hu-1 ref.

- 3 sites are shared MT123291, MT123293 and MT019533/MT123293. This implies either *in vivo* deamination hot spots or first generation P-to-P transfers with additional deaminase mediated mutational events laid down on that common sequence structure.
- 24 of the variable sites are thus unique singleton sites – a feature highlighted by Stoddard et al 2015 for the *in vivo* pattern among quasi species for acute phase patterns in individual HCV patients.
- In general, by the criteria for AA impact applied (Figure 1), there is much 'functional sequence conservation' or 'benignness' in these sequences. There is therefore qualified support for the supposition that these are the "Darwinian Cut" survivors of a host innate immune deaminase attack on the acute phase viral sequences (although a small portion are deduced to be the result of ROS 8oxoG modifications - indeed ROS attack is a common innate immune defence against intracellular pathogens).

Putative APOBEC and ADAR Deaminations among L Hap group

Among the L Hap SNVs in the 11 L sequence set we have : 9 C>T (one shared, presumed APOBEC (+) strand RF); 4 G>A (presumed APOBEC (-) strand RF); 5 A>G (presumed ADAR1/2 (+) strand RF); 1 T>C (shared, presumed ADAR1/2 (-) strand RF); 1 G>T (presumed 8oxoG at WG site (+) strand RF); 1 C>A (presumed APOBEC/8oxoG (-) strand RF); 1 A>C (presumed ADAR1/2 (+) strand RF); 2 T>A (presumed ADAR1/2 (-) strand RF). The identified APOBEC and ADAR putative changes are at typical motifs as observed for HCV, ZIKV (Lindley and Steele 2018). Further specific clarification is a focus of our and ongoing investigations (see also Rosenberg et al 2011, Sharma et al 2015, 2016, Eifler et al 2013).

SNVs per Sequence in L Hap Group

Among the L Hap sequences the number of unique SNV differences per sequence from Hu-1 ref. are , 4, 3, 3, 2, 2, 2, 2, 2, 1, 1, 2 ; thus a range of about 2-4 differences per sequence for the assumed first infection cycle might be concluded.

So apart from one possible single P-to-P interchange or transfer (MT123291<->MT123293) all appear to be part of the acute phase deaminase-mediated innate immune host response in the first infection with L Hu-1 viruses. So the distribution of SNV numbers in order are summarised as:

No. L Seqs.	No. SNV *v* Hu-1
25	0
2	1
5	2
2	3
2	4

There is very little P-to-P spread at the height of the epidemic in Wuhan and surrounding regions. Most COVID-19 positive subjects appear to be infected with the same virus viz. Hu-1 ref. This conclusion is consistent with the same conclusion based on the phylogenetic analysis in January 2020 during the exponential rise in COVID-19 cases in Wuhan (Anderson 2020).

An alternative, and partial, explanation is that many of those L haplotypes showing complete sequence identity to Hu-1 are actually products of P-to-P transfers with no further laying down of deaminase-mediated mutations in the recipient host from which the collection was made. The large number of unmutated COVID-19 L sequences displaying the Hu-1 reference L sequence was also observed a month later in infected patients on board the *Grand Princess* cruise ship off the Californian coast (February 18-24, Figure 3).

S Haplotype Analysis of the China data

The typical sequences in the minor S Hap group, apart from outlier MT226610, are represented by: MT049951, MT135041, MT135044, MT135042, MT135043, MT123292, MN975262, LR757995, MN938384. The SNVs are: 1 T>C (shared by MT135041, MT135044, MT135042, MT135043); 1 G>T (8oxoG ? shared by MT135041, MT135044, MT135042, MT135043); 4 C>T (one shared with L Hap, thus hotspot? MN996531, MN975262; another shared MT123262, MN938384); 1 A>T; 1 G>A (shared between MT123292, and MT123 291 a L Hap variant, thus APOBEC hotspot?); 1 G>C; 1 T>A.

After an assumed L Hu-1 primary infection there is a group of four Beijing subjects sharing a S Hap pattern: MT135041, MT135044, MT135042, MT135043. This might be indicative of P-to-P in the second/third infection cycle as one of this group MT135043 has an additional unique SNV. The other five S Hap variants have 3, 3, 3, 1, 0 differences from Hu-1 with an additional possible P-to-P transfer (MN975<-> MN938384).

However, as we highlighted in our *Caveats* section, much unknown information limits the scope of this analysis i.e. “Who met who” or “Who most likely gave the COVID-19 variant to whom?” For the limited set of S Hap variants there is evidence of P-to-P, and additional layers of deaminase-driven mutagenesis after transfer on top of the putative L to S switch. It is conceivable that on first infection with Hu-1 the sequences in MT135041,

MT135042, MT135043 were shared by P-to-P transfer. It is further conceivable that other S Hap variants were also created in the first infection in the putative quasi-species L pool, namely MT123292, MN195262, LR757995, MN938384, with transfer between MN195262 <->MN938384 and other unknown recipients (i.e. those not in our collection). Thus, three of the four S Hap variants appear to be created *in vivo* during a putative L>S deaminase-driven riboswitch then P-to-P transfer creating MN195262 <->MN938384 sequences.

Conclusions on SNV differences from presumed source virus Hu-1

Understanding what happened in the early phase of the COVID-19 pandemic at its agreed epicentre in Wuhan city and regions is important to understand before analysing the further spread of the virus during the pandemic to other countries. Most COVID-19 isolates display the unmutated sequence of the L Hu-1 reference virus. Smaller numbers display 1 to 4 SNV differences from L Hu-1. The SNVs at position "63" look like independent deamination events at a hot spot as they are of different haplotypes (MT123291, MT123292).

These mutagenic patterns in Hu-1 are consistent with host-derived deaminase-mediated mutation signatures at the Wuhan epicentre and wider Hubei region and neighbouring provinces. Among the other unknowns we have no estimate of the *magnitude of the infective dose* from the source COVID-19 virus in Wuhan/Hubei. As suggested, the many unmutated sequences might infer Wuhan patients were older co-morbids who failed to adequately mount defensive deaminase-mediated Innate Immunity. However, this interpretation is speculative without knowing patient outcomes, and patient-patient relationship for P-to-P inferences (as previously discussed). Thus, it is important to conduct further analysis of this pattern at the explosive epicentre of the COVID-19 outbreak in order to hopefully inform analysis of later outbreaks elsewhere in the USA and Europe.

The sequence MT226610 is a clear outlier. It has 27 SNVs compared with the Hu-1 reference and is of the S Hap group. It may represent serial mutagenic episodes (relapsing under clinical treatment ?) in one patient. The other interpretation is that it is the product of multiple (x8-x9) P-to-P chain transfers of infection with additional layers of mutations laid down during each infection prior to the next transfer. If so, this sequence could be an immune 'escape variant' well on the way to immune evasion and thus higher 'virulence' (?). If the latter is correct, we should note that we have not seen it crop up again in the sample sets we have analysed in Spain and USA. Alternatively as suggested by a reviewer the sequence is an artefact of poor assembly and sequencing.

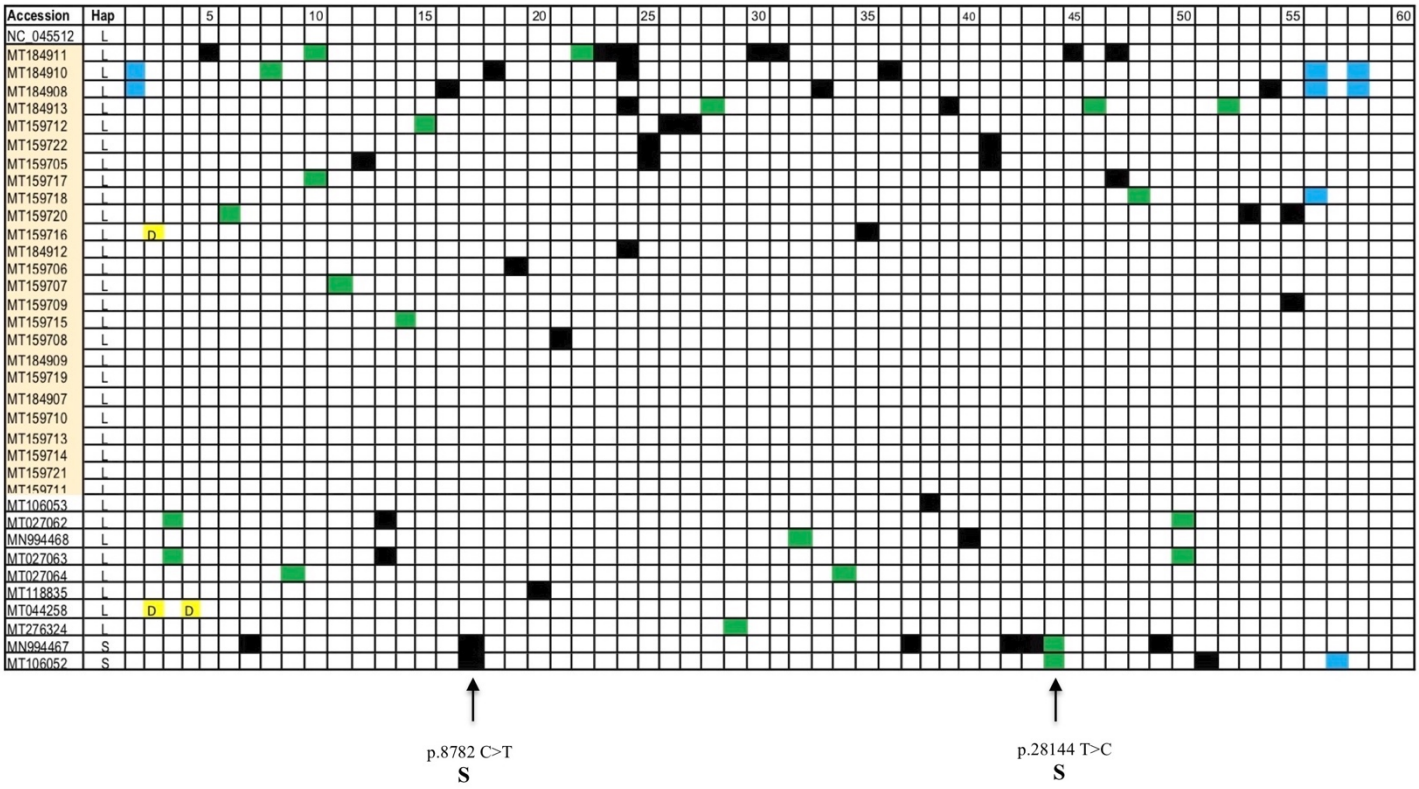

Figure 3 Variable Site Diagram of SNV in each aligned sequence in the 36 CA + Cruise Ship sequence alignment (Jan 22-Feb24) which includes Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. S Hap sites are indicated for sites 17 and 44 and arrowed (and see Table 2). Cruise Ship Accession IDs are in orange highlight and grouped. The data in TableS2 should be consulted for further details. The variable site column number followed by SNV position in the alignment are : 1, p.254, C>T, 5'UTR; 2, p.508-522, GHVM in-frame deletion; 3, p.614, G>A; 4, p.686-694, KSF in-frame deletion; 5, p.1063, C>T; 6, p.1385, C>T; 7, p.1548, G>A; 8,p.1911, C>T; 9, p.2091, C>T; 10, p.3099, C>T; 11, p.3259, G>T; 12, p.3738, C>T; 13, p.5084, A>G; 14, p.5845, A>T; 15, p.6636, C>T; 16, p.8312, A>T; 17, p.8782, C>T; 18, p.9157, T>C; 19, p.9474, C>T; 20, p.9924, C>T; 21, p.10036, C>T; 22, p.10083, C>T; 23, p.10507, C>T; 24, p.11083, G>T; 25, p.11410, G>A; 26, p.11750, C>T; 27, p.11956, C>T; 28, p.12513, C>T; 29, p.14718, G>T; 30, p.15193, G>C; 31, p.15810, C>T; 32, p.17000, C>T; 33, p.20988, T>C; 34, p.21710, C>T; 35, p.22033, C>A; 36, p.22104, G>T; 37, p.24034, C>T; 38, p.24325, A>G; 39, p.25587, C>T; 40, p.26144, G>T; 41, p.26326, C>T; 42, p.26729, T>C; 43, p.28077, G>C; 44, p.28147 (=p.28144 "S" site), T>C; 45, p.28253, C>T; 46, p.28367, C>T; 47, p.28381, G>T; 48, p.28409, C>T; 49, p.28792, A>T; 50, p.28854, C>T; 51, p.28878, G>A; 52, p.28916, G>A; 53, p.29230,C>T; 54, p.29596, A>T; 55, p.29635, C>T; 56, p.29736, G>T, 3'UTR; 57, p.29742, G>A, 3'UTR; 58, p.29751, G>C, 3'UTR.

2. Analyses of sporadic early USA outbreaks in California and *Grand Princess* cruise ship Jan 22-Mar 4

The early appearance of COVID-19 in the USA began in California and Washington State, and particularly the sharp outbreak on the *Grand Princess* cruise ship (off the coast of San Francisco which occurred mid to late Feb with swab collections Feb 18-24). This early appearance in the USA West coast, instance California and Washington State prior to any significant infections in the rest of the USA (FigureS1,S3) is reminiscent of the early time course of the pandemic spread to the USA from a similar China-originating pandemic outbreak of

influenza H3N2 in 1968 in the USA (FigureS6). For this reason, we have decided to conduct a detailed COVID-19 sequence analysis of isolates of the sporadic and smaller scale USA West coast outbreaks that occurred before significant outbreaks in Europe and New York city.

Analysis of the Alignment of 35 sequences from VSD Figure 3

In this alignment there are 58 variable sites. For the CDS region there are 54 variable sites
Of the 35 sequences, 33 are of L Haplotype – 25 are on the Cruise Ship, 8 on CA mainland and 2 are of the S Haplotype, both on the CA mainland. The break-down of the types of SNV are shown in Table 3b. Again C>T(U) transitions dominate the data set and the strand bias pattern is very similar to the China data (Table 3a) with the same implications as discussed. Far fewer A-site mutations are evident in these data, but appear strand balanced.

L Haplotype Analysis – e.g. Cruise Ship v CA Mainland

Among sequences in the major L Hap group there are 50 unique variable sites. Two are in-frame deletions, one shared between MT159716 and MT044258 suggestive of possible P-to-P transfer between these two subjects. Nine of 25 Cruise L sequences are identical to Hu-1 reference i.e. unmutated. No mainland L Hap sequences are unmutated. For the Cruise Ship the distribution of the number of SNV differences per sequence from Hu-1 for zero difference to 9 per sequence are 9, 6, 3, 2, 0, 1,1,1, 0, 9; for the CA mainland the corresponding numbers are 0, 3, 2,2 1, 0, 0, 1,0,0. The two shared sequences on the CA Mainland suggestive of P-to-P are between MT027062<-->MT027063. Similarly, Cruise Ship passengers MT159722 and MT159705 display evidence of sequence sharing and P-to-P transfer of MT159722 to MT159705. The common in-frame deletion p.686-694 suggests some earlier P-to-P transfer connecting MT159716 (Cruise Ship) and MT044258 (CA Mainland). It is also conceivable that Cruise Ship sequences MT184910 and MT184908 are derived (by P-to-P) from a common ancestral sequence as they have identical SNVs in the 5' and 3' UTRs. However, to qualify, in these cases the UTR changes maybe at putative riboswitch hotspots and thus indicative of an emerging new deaminase and ROS driven haplotype seeking to be established? It is observations like this that suggest that sequencers should aim for complete full length genome sequences that include both 5' UTR and 3' UTR regions (as is the case for the Hu-1 reference sequence).

Overall, the "quasi-species' acute phase infection pattern seen *in vivo* in individual subjects infected with a positive strand RNA viruses (e.g. as shown by Figure 2 in Stoddard et al 2015 for HCV), is now observed in the population of COVID-19 infected individuals (see Figures 2,3). This is supported by our observation that MT184910 <->MT184908 share three putative riboswitch changes in 5' and 3' UTRs (p.254, p.29736, p.29751

with further possible deaminase-mediated and ROS 8oxoG mutagenesis in both subjects during their separate infections with COVID-19.

On the Cruise ship the cases with putative evidence of P-to-P sequence sharing with further layers of deamination mutations in transferred infection appears evident e.g. MT159722 <-> MT159705, and MT159716 ship <-> MT044258 mainland. On the CA Mainland two shared sets of mutations are suggestive of P-to-P between MT027062<-->MT027063. Finally, as noted MT184910 <->MT184908 shared putative 5' and 3' UTR riboswitch variants.

Both S Hap variants are CA Mainland-derived carrying 4 and 7 differences from Hu-1 (the differences between the S Hap members, MN994467, MT106052). The other CA Mainland subjects display the sets of apparently random array of differences from Hu-1 per sequence (above). As with the China data (Figure 2) from 2 to 4 differences from Hu-1 in the first infection with L Hu-1 sequence seems to be the norm. The outlier MT184911 on the Cruise ship suggests that up to 9 differences can accrue in a single sequence, although the precursor sequence for this variant appears to be MT159718, based on P-to-P transfer to MT184911 and then producing the additional 7 SNV variants in that sequence during that subjects innate immune response to the virus.

In the L Hap alignment the distribution of largely deaminase driven SNVs is approximately, 29 x C>T (4 are shared), 5x G>A (2 are shared), 11x G>T (8 are shared and putative 8oxoG G>T at WG sites are noted possibles), 3x (G>C (2 are shared), 3x A>T, 2x A>G (shared), 2xT>C plus two in-frame deletions of codons in the two S Hap sequences 1x A>T, 2x T>C (shared), 3x C>T (2 shared), 3x G>A and 1x G>C.

Conclusions on CA + Cruise Ship data (Jan 22-Feb 24)
Among COVID-19 patients sequenced on the Cruise Ship we observe no mutational evidence of P-to-P spread at the height of this localised epidemic outbreak. From the limited data of confirmed cases this also applies to the surrounding CA Mainland region. Most COVID-19 positive subjects appear to be infected with the same viral strain viz. the putative Hu-1 ref initiated by "community spreads" with no obvious Patient X. This conclusion is again consistent with the phylogenetic analysis by K. Anderson in January 2020 during the exponential rise in COVID-19 cases in Wuhan (Anderson 2020).

So, the conclusions here are strikingly similar to the far larger outbreak in Wuhan, China. Most are unmutated representatives of the L Hu-1 reference virus. Smaller numbers display from 1 to 4 differences from L Hu-1.

Although the sample size is small, the proportion of the S Hap variant is just 2 out of 35 sequences (at canonical p. 8782 and p. 28144 that define L>S as in Tang et al 2020); this compares with an estimated 20% S Haplotype in China (Figure 2).

In Summary again, as in the China data analyses, these mutagenic patterns in Hu-1 are indicative of host-derived deaminase-mediated mutation signatures, particularly striking in the case of the *Grand Princess* cruise ship outbreak (where location at sea at outbreak is defined). Low mutation, low P-to-P spread (although some have likely occurred in 1 or 2 step transfers). The many unmuted sequences on the cruise ship might infer again that patients were older co-morbids who failed to adequately mount defensive deaminase-mediated innate immunity. But we believe that the data show that transfers between a couple can be inferred. That all these patients with unmuted Hu-1 sequences maybe of Chinese ethnicity can also only be inferred.

3. Analyses of sporadic USA outbreaks in California mainland February 27- March 4, 2020.

Analysis of the Alignment of 24 CA sequences from VSD Figure 4

The VSD pattern for the California outbreaks Feb 27-Mar 4 ~~(requiring COVID-19 sequencing)~~ are displayed in Figure 4 sorted into the two main haplotypes, L and Sa. Again, quasi-species type variants carrying about 4 SNV randomly distributed variants from either Hu-1 or the Sa haplotype are noted (with limited putative P-to-P sequence sharing e.g. MT419832<>MT419 830<>MT419831; MT419833<>MT419834; MT419836<>MT419835<>MT419839 <>MT419837). The Sa haplotype differs from the S haplotype detected in China (or earlier on the CA Mainland) by coordinated SNV changes seen at p.17747 (C>T), p.17858 (A>G) and p.18060 (C>T) - all these SNVs involve conserved changes (or lack of change) at the amino acid level, and because of their common strain status qualify as riboswitched haplotype changes within the S haplotype (as discussed, Table 2).

The types of SNV are displayed in Table 3c. and are similar to other collections discussed (Table 2), C>T changes dominate, on the +ve RNA strand.

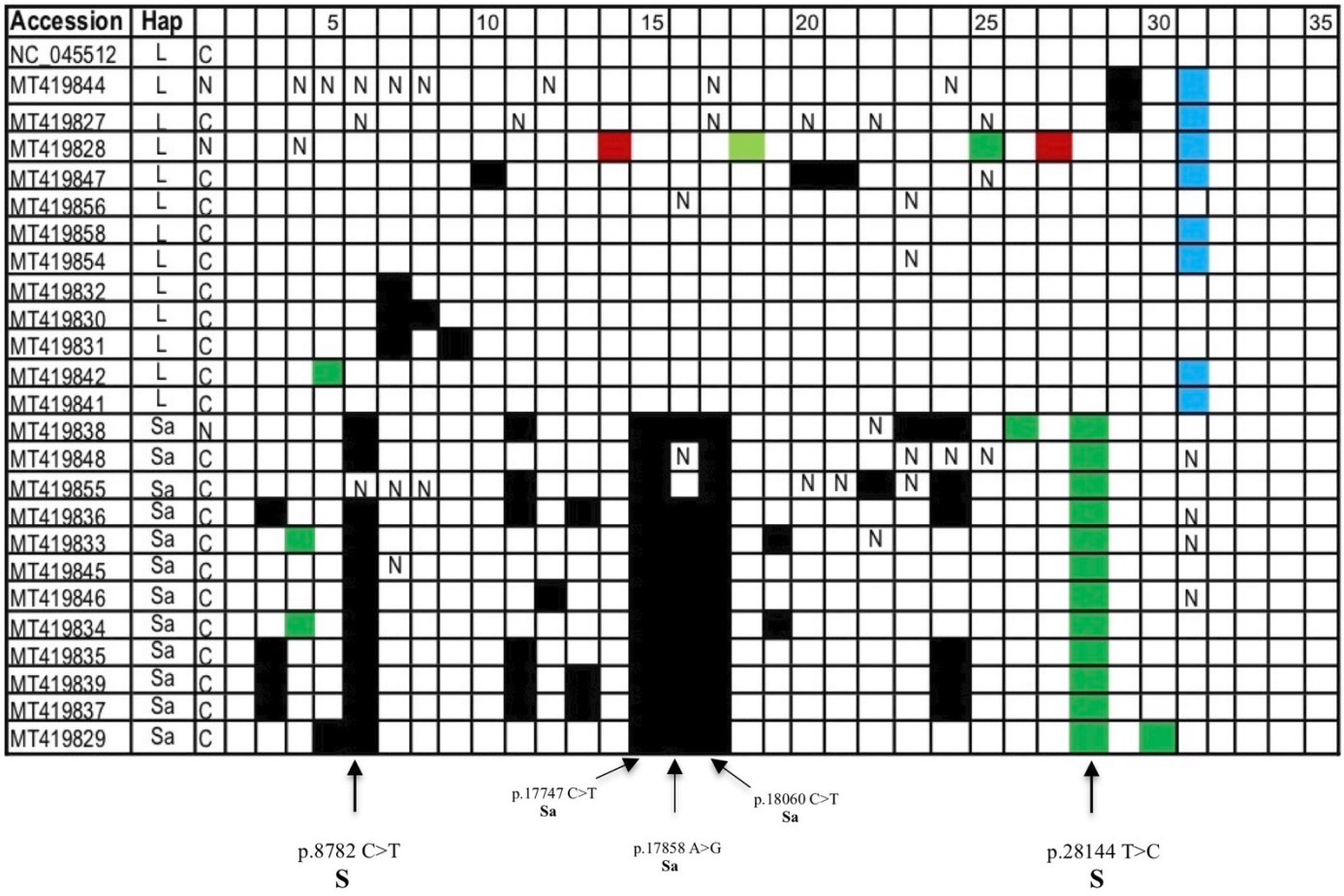

Figure 4 Variable Site Diagram of SNV in each aligned sequence in the 24 CA sequence alignment (Feb 27-Mar 4) which includes Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. S Hap sites are indicated for sites 6 and 28, and Sa subsites defined by p.177747, p.17858, p.18060 and arrowed (and see Table 2). A screen shot record of flanking sequences around each SNV in codon-context was made, and this record used to construct the pattern in Figure 4. In some cases an N or N run created uncertainty, and this is recorded here as N to reflect the quality of the sequencing in this batch of complete genomes uploaded to NCBI Virus. This qualification allows assessment of assignment of variable site 16, p.17858, A>G for MT419855 – it should be "G" by Haplotype imputation at this position. However, the generally poor sequence of MT419855 (many N runs) suggest that this assignment is in likely error as well. The variable site column number followed by SNV position in the alignment are : 1, p.241, Hu-1 ref is C, some are N, so most are unlikely of L241 Hap; 2, p.3046, A>G; 3, p.5184, C>T; 4, p.7798, G>T; 5, p.7815, C>T; 6, p.8782, C>T; 7, p.9924, C>T; 8, p.9951, C>T; 9, p.15641, A>C; 10, p.16240, G>A; 11, p.16467, A>G; 12, p.16679, C>T; 13, p.16975, G>T; 14, p.17725, A>G; 15, p.17747, C>T; 16,p.17858, A>G; 17, p.18060, C>T; 18, p.19169, T>C; 19, p.20148, C>T; 20, p.21796, G>A; 21, p.21838, T>C; 22, p.22139, G>T; 23, p.23014, A>G; 24, p.23185, C>T; 25, p.24989, C>A; 26, p.25468, T>C; 27, p.28117, A>G; 28, p.28144, T>C; 29, p.28178, G>T; 30, p.29253, C>T; 31, p.29711, G>T , 3' UTR– possibly 8oxoG ROS product at WG hotspot and thus potential riboswitch?

4. Analyses of the outbreak in Washington State, Kirkland, Nursing Home Outbreak near Seattle February 27- March 1, 2020.

This is a small alignment (extracted from a larger alignment) relevant to the early spread pattern in the USA (FigureS6) up to and including collections Mar 4. It involved targeted collections and sequencing in the Seattle area. The tabulated data is in TableS3 and a data summary is recounted here.

The Kirkland nursing home outbreak was widely reported in the media, occurring more or less at the same time in late February as the California outbreaks and involving the at-Sea cruise ship *Grand Princess* just analysed. A Variable Site Analysis diagram is shown in Figure 5.

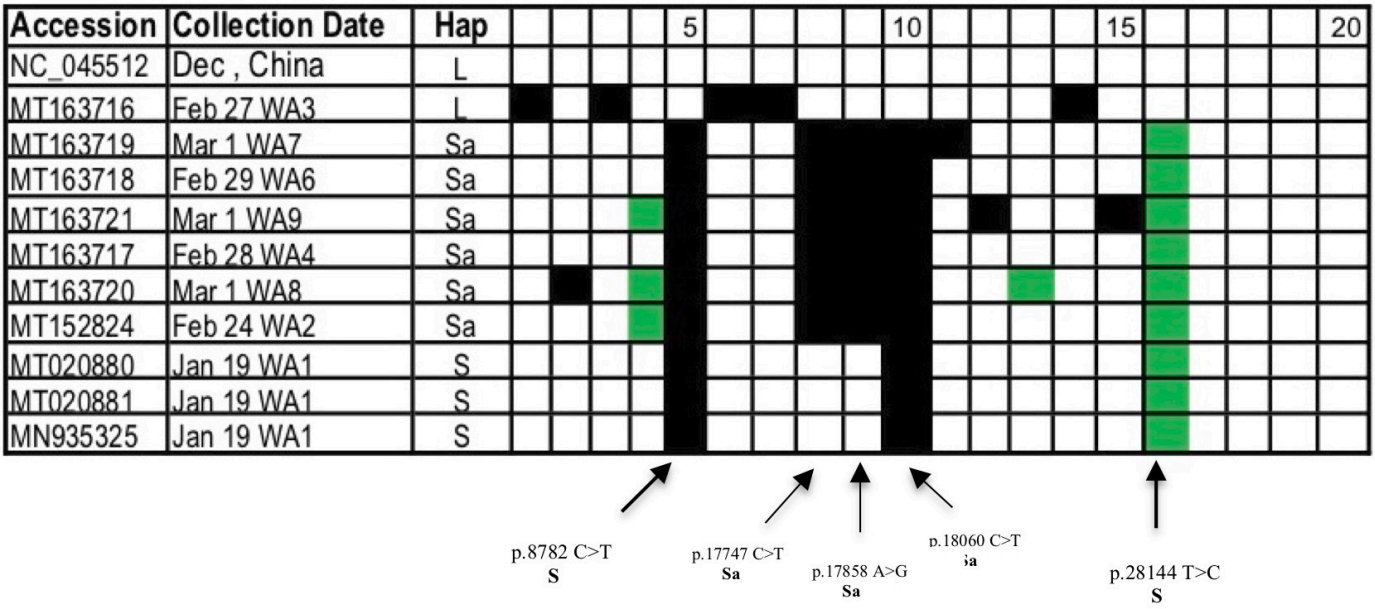

Figure 5 Variable site plot of SNV in each aligned sequence in the 10 sequence alignment for the WA State outbreak Feb 27 – Mar 1 2020 versus the Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. S, and Sa Hap designations as indicated and arrowed (and see Table 2). This alignment was part of the larger alignment done for Figure 4. Putative (speculated) P-to-P transfers are shown in Figure 6, including the Patient X, MN985325, the first case nCo-2019 in the United States collected Jan 19 2020; the others are the same virus from culture isolates (MT020880, MT020881). The variable site column number followed by SNV position in the alignment are : 1, p.2446, T>C; 2, p.3406, A>C; 3, p.5573, G>T; 4, p.5782, C>T; 5, p.8782, C>T; 6, p.11083,G>T; 7, 14085, C>T; 8, p.17747, C>T; 9, p.17858, A>G; 10, p.18060, C>T; 11,p.20282, T>C; 12, p.20580, G>T; 13, p.23528, C>T; 14, p.26147, G>T; 15, p.26733, G>A; 16, p.28147 (read as .28144), T>C.

It is evident that P-to-P transfers have occurred as Patient X (WA1) appears, from all our previous analyses, to be a patient who was infected with the L_{Hu-1} reference virus sequence which underwent a L>S>Sa haplotype switch at the two canonical S-sites,, as well as the additional three sites at p.17747, p.17858, p.18060. In this targeted, University of Washington sequencing analysis of COVID-19 patients, we can construct putative patient-to-patient transfers of the virus which lays down a further small number (about 2-3 further mutations in each infection) a pattern typical of quasi-species with mutations away from the L_{Hu-1} reference virus sequence. Thus all the data in Figure 5 can be logically explained in terms of haplotype switching, P-to-P transfers and then largely deaminase-mediated mutagenesis, that together lay down further mutational signatures in each productive infection (there are also some putative ROS 8oxoG or G>T changes that can be identified).

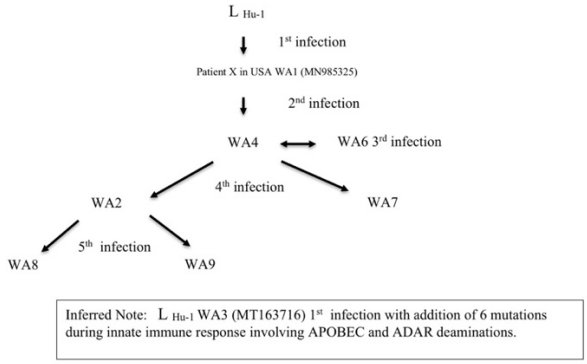

Figure 6. Putative P-to-P Transmissions in Kirkland Outbreak.
Possible P-to-P transfer chains inferred from Figure 5.

5. Analyses of COVID-19 complete genomes collected mainly in Spain Feb 26 – Mar 5, Mar 6 -10, 2020

The variable site patterns of one of two alignments of sequences largely from Spain collected Mar 6 – Mar 10 are shown in Figure 7. The haplotype variations from S as Ss is shown, including the L241c variation observed in Spain but not so much NYC (below). Apart from the small numbers of unique SNV positions there is very little mutation away from the main haplotype group. Presumably the collections were targeted on known groups suffering disease. However, the Ss haplotype assignment appears solid. It was observed with little further mutations also a week earlier in another set of Spain collections (involving an alignment of MT233519, MT233520, MT233521, MT233522, MT233523, MT198653, MT198651, MT198652). The types of SNVs are summarised in Table 3d, revealing a pattern that is similar to all other collections with a dominance of C-site deaminations on the +ve strand (C>U/G>A) over A-site deaminations, and more or less balanced to both +ve and -ve strands.

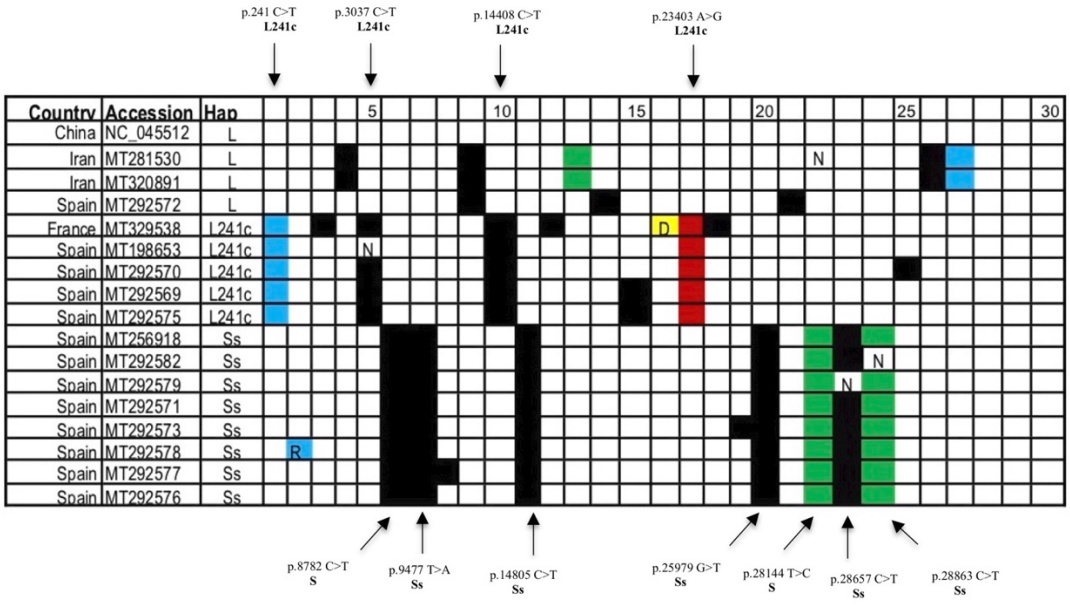

Figure 7 Variable site plot of SNV in each aligned sequence in a 17 sequence alignment for mainly Spain (13 sequences) versus the Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype assignment. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. L, L-241, S, and Ss Hap designations as indicated and arrowed (see Table 2). The variable site column number followed by SNV position in the alignment are : 1, p.241, C>T; 2, p.242,244, G>T, C>T; 3, p.618, A>G; 4, 1397, G>A; 5, p.3037, C>T; p.8782, C>T; 7, p.9477, T>A; 8, p.10156, C>T; 9, p.11083, G>T; 10, p.14408, C>T; 11, p.14805, C>T; 12, p.15324, C>T; 13, p.18377, C>T; 14, p.18835, T>C; 15, p.20268, A>G; 16, p.21880, in-frame deletion; 17, p.23403, A>G; 18, p.24025, A>G; 19, p.24928, G>T; 20, p.25979, G>T; 21, p.26144, G>T; 22, p.28144, T>C; 23, p.28657, C>T; 24, p.28863, C>T; 25, p.29144, C>T; 26, p.29374, G>A ; 27, p.29742, G>T. 3'UTR.

6. Analyses of COVID-19 complete genomes collected in New York Mar 5 – 9.
7.

The genomes of a small number of collections for COVID-19 sequencing in New York Mar 5 – 9 just prior to the exponential increase in cases (from about March 14 onwards) were aligned against the Hu-1 reference. The VSD pattern is shown in Figure 8, and the types of SNV recorded in Table 3e. It appears that overt COVID-19 cases were targeted for sequencing and that the sample is clinically biased. The genomes appear harvested from small groups of subjects where putative P-to-P transfers was suspected - several groups share a common COVID-19 sequence with additional unique SNVs added following transfer to suspected recipients. For example, MT143800 may have transferred its COVID-19 sequence to MT434786, with one additional SNV added. In the L Hap group of sequences (MT434807, MT434787, MT434783, MT434784) putative transfers may have been MT434787-> MT434783<>MT434784 with a further additional 2-4 unique SNVs laid down after transfer. The common SNVs in four sequences at sites 17 (p. 148805, C>T, synonymous Tyr<>Tyr) and 18 (p.17247, T>C, synonymous Arg<>Arg) could suggest both P-to-P transfers and/or deamination hot spot

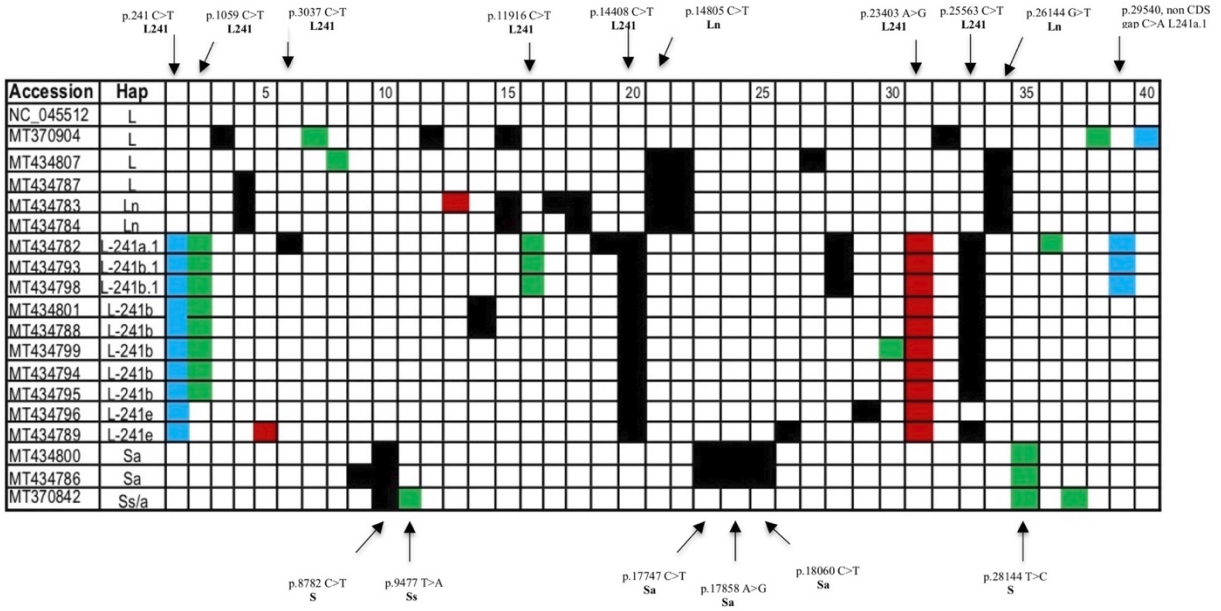

Figure 8 Variable site plot of SNV in each aligned sequence in an 18 sequence alignment for collections New York Mar 5 – 9 versus the Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. L, S, and Sa Hap designations are indicated (see Table 2) and the main S and L haplotype sites are in sites 10 and 35. The other main sites for the L241 haplotypes (Table 2) are also highlighted in bold and arrowed. The variable site column number followed by SNV position in the alignment are : **1, p.241, C>T**; **2, p.1059, C>T**; 3, p.1397, G>A; 4, p. 1625, C>T; 5, p.2592, C>A; **6, p.3037**, C>T; 7, p.3242, G>A; 8, p.5730, C>T; 9, p.6639, A>G; 10, p.8782, C>T; 11, p.9477, T>A; 12, p.9514, A>G; 13.p.10155, A>G; 14, p.10851, C>T; 15, p.11083/11080, G>T; **16, 11916, C>T**; 17, p.12992, C>T; 18, p.13265, A>T; 19, p.13536, C>T; **20, p.14408, C>T**; 21, p.14805, C>T; 22, p.17237, T>C; 23, p.17747, C>T; 24, p.17858, A>G; 25, p.18060, C>T; 26, p.18877, C>T; 27, p.18985, G>T; **28, p.18998, C>T**; 29, p.20268, A>G; 30, p.21846, C>T; **31, p.23403, A>G**; 32, p.25215, C>T; **33, p.25563, G>T**; 34, p.26144, G>T; 35, p.28144, T>C; 36, p.28989, A>T; 37, p.28863, C>T; 38, p.29027, G>T; **39, p.29540, G>A, non-CDS gap**; 40, p.29742, G>T, 3'UTR. Note that sequence MT370842 was collected Mar 4, and sequence MT370904 was collected Feb 29.

changes. Overall the numbers of common and unique SNVs among the L group is similar to that observed in the Wuhan outbreak and the earlier outbreaks on the West coast USA (CA + Cruise ship). Among the L-241 haplotype series there are also examples of P-to-P transfers and further additions of SNVs after transfer. Novel unique mutations per infection are low (but MT370904 has no SNV from the Hu-1 L sequence) as observed earlier for first or second infections (2-6 SNVs per sequence per P-to-P transfer). Again C>T transition SNVs on the +ve strand dominate the sample numbers (Table 3e).

8. Analyses of COVID-19 complete genomes collected in New York Mar 14 – 22.

The genomes of 206 COVID-19 subjects were selected for the period Mar 14- 22. A screen shot tabulation was created of alignments against Hu-1 in approximate groups of 10. It was noted, in an initial survey, that most sequences were of the L-241 haplotype with the distinctive C>T at p.241 in the 5' UTR Our analyses of NY sequences during the explosive exponential rise of confirmed COVID-19 cases proceeded in several steps. We separately assess the types of L and S haplotypes in these 206 sequences. We analysed 16 L + S sequences that were clearly not of the L-241 haplotype. We then analysed a sample of the first 58 in sequences in their temporal order of curation/upload to NCB Virus.

7.a. Analysis of L and S Haplotype derivatives collected in New York Mar 14 – 22

The VSD pattern for the set of these 16 sequences, seven S and nine L is displayed in Figure 9. The types of SNV are tabulated in Table 2f. Among L we only see the Ln variant haplotype, defined by SNVs from Hu-1 at

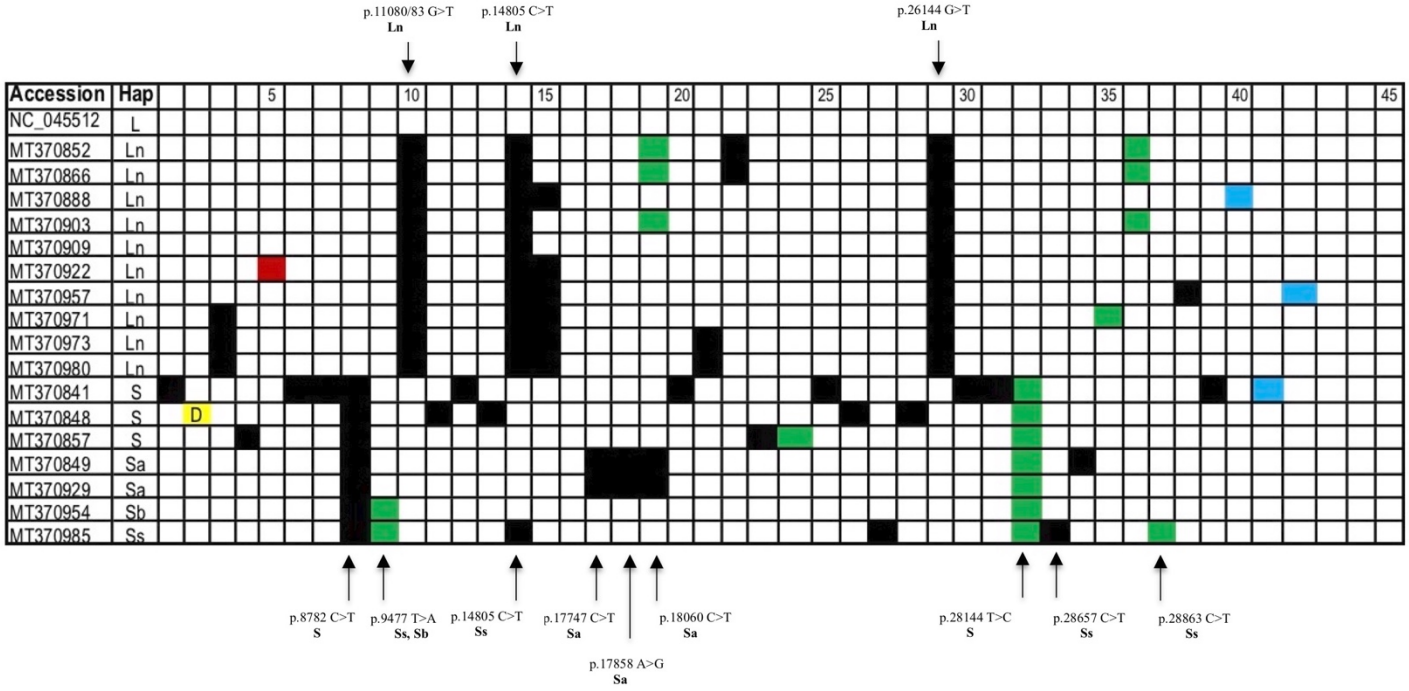

Figure 9 Variable site plot of SNV in each aligned sequence in a 16 sequence alignment for collections New York Mar 5 – 9 versus the Hu-1 ref NC_045512.2 focusing on L and S haplotype derivatives. Variable site number across the top, and Sequence ID down left hand side and Haplotype. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. L, S Hap designations as indicated (see Table 2) and the main S and L haplotype sites are indicated (site in bold and arrowed) – for S p.8782 site 8, p. 28144 site 32; for Ln sites p. 11080/83 site 10, p. 14805 site 14, p. 26144 site 29. The variable site column number followed by SNV position in the alignment are : 1, p.490, T>A; 2, p.1600-1616, in frame deletion LNDNL; 3, p.1625, C>T; 4, p.2676, C>T; 5, p.2745, A>T; 6, p.3177, C>T; 7, p.6040, C>T; **8, p.8782, C>T**; **9, p.9477, T>A**; **10, p.11080/83, G>T**; 11, p.12274, G>A; 12, p.12478, G>A; 13, p.13115, C>T; **14, p.14805, C>T**; 15, p.17247, T>C; **16, p.17747,C>T; 17, p.17858, A>G; 18, p.18060, C>T**; 19, p.18086, C>T; 20, p.18735, T>C; 21, p.19166, A>G; 22, p.21137, A>G; 23, p.22606, A>T; 24, p.23525, C>T; **25, p.24034, C>T**; 26, p.25541, T>C; **27, p.25979, G>T**; 28, p.26087, C>T; **29, p.26144, G>T**; 30, p.26729, T>C; 31, p.28077, G>C; **32, p.28144, T>C**; **33, p.28657, C>T**; 34, p.28708, C>T; 35, p.28739, G>T; 36, p.28842, G>T; **37, p.28863, C>T**; 38, p.28878, G>A; 39, p.28896, C>G; 40, p.29543, G>C, non-CDS gap; 41, p.29700, A>G, 3'UTR; 42, p.29742, G>A, 3'UTR.

p.11080/83, p.14805, p.26144 (Table 2). The Wuhan L variant is not seen in this sample, unlike the week earlier. Thus, Ln is the dominant haplotype and likely sharing of sequences are evident e.g. MT370852, MT370866, MT370903, indicative of P-to-P transfers. However little further mutation is observed on transfer.

Similar P-to-P patterns are evident among MT370971, MT370973, MT370980. In addition, the quasi-species type patterns of apparently random SNVs are also evident in these data- as commented on above for the Wuhan and Cruise Ship patterns (Figures 2, 3).

Among S Haplotype mutational derivatives, the S in Spain (Ss) is found in one case (MT370985). Other examples of sequence sharing (P-to-P) among S are evident in these data.

7.b. Analysis of a sample of 58 sequences collected in New York Mar 14 – 19

The VSD pattern for the set of these 58 sequences, grouped by Haplotype is displayed in Figure 10, and the tabulated data in TableS4. The types of SNV are shown in Table 2g. The striking pattern of apparent mutational diversity compared with the that in Wuhan is apparent. In Wuhan there was a dominant, largely unmutated (or lightly mutated) L haplotype of the Hu-1 sequence (Figure 2). The pattern in New York in the exponential phase is complex and diverse. However, as discussed above, the great bulk of this diversity resides at key sites that determine the main haplotypes (Table 2), particularly for the L-241 series haplotypes, and some L-241a Hap which is dominant (much like the L in Wuhan). As discussed already the L and S haplotypes form a minor component of the NY variable site pattern (Figure 9). The L-241b Hap which was the major one present in the small sample in NY in the week before case numbers began to explode (Figure 8) has been replaced in this sample, by L-241a which was a minor haplotype identified in that earlier period.

Whilst there are some cases of sharing of sequences (putative P-to-P transfers), an important finding is that the L-241a haplotype set (defined by changes from Hu-1, at p.241, p.1059, p.3037, p.14408, p.23403 and p.25563) has a very low number of mutations. If P-to-P transfers are ongoing then the recipients are not laying down a significant deaminase-mutagenic pattern in their own infection prior to transfer as one might expect given the

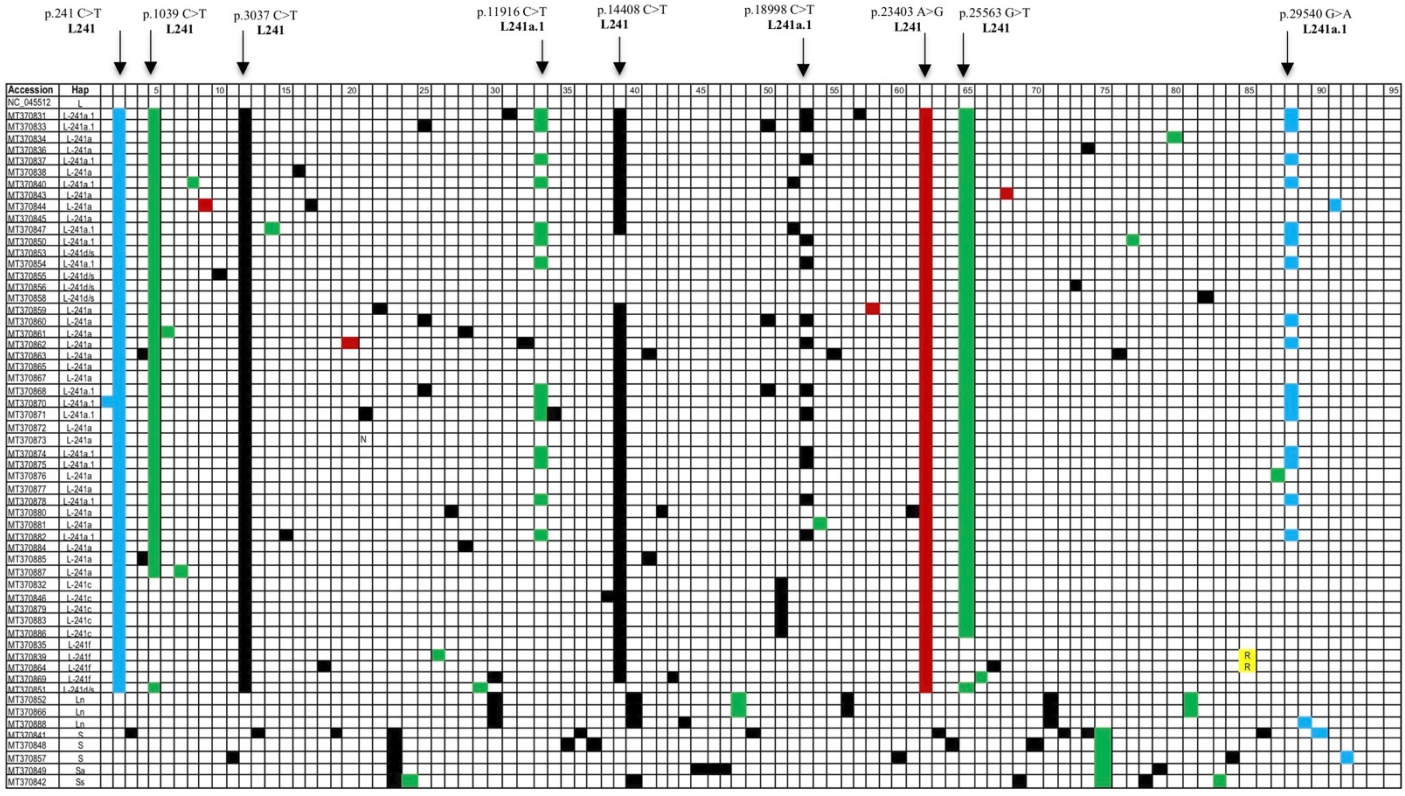

Figure 10 Variable site plot of SNV in each aligned sequence in a 58 sequence alignment for collections New York Mar 14 – 19 versus the Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype. The

source data is in the tabulation in Table S4. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. L, S Hap and L-241 subset Hap designations as indicated (see Table 2). The variable site column number followed by SNV position in the alignment are in order with the L-241 sites listed in Table 2 in bold and arrowed. Ln, S, Sa, Sb, Ss already summarised for NYC collection as in Figure 9. Other features whether in UTR, non-CDS region or G>T most likely caused by oxidation of G (8oxoG) are added given the complexity of the data set : 1, p.199, G>T, 5'UTR; **2, p.241, C>T, 5' UTR**; 3, p490, T>A; 4, p.619, C>T; **5, p.1059, C>T**; 6, 1820, G>A; 7, p.1917, C>T; 8, 2091, C>T; 9, p.2165, A>G; 10, 2632, G>T, 8oxoG at WG; 11, p.2676, C>T; **12, p.3037, C>T**; 13, 3175, C>T; 14, p.4035, T>C; 15, p.4113, C>T; 16, p.4456, C>T; 17, p.4810, C>T; 18, p.5140, C>T; 19, p.6040, C>T; 20, p.6324, A>G; 21, 7291, A>G; 22, p.7770, A>G; **23, p.8782, C>T**; **24, p.9477, T>A**; 25, p.10015, C>T; 26, p.10265, G>A; 27, p.10369, C>T; 28, p.10851, C>T; 29, p.11003, C>T; **30, p.11083/80, G>T, 8oxoG at WG** ; 31, 11101, A>G; 32, 11191, T>A; **33,p.11916, C>T**; 34, p.12153, C>T; 35, p.12274, G>A; 36, 12478, G>A; 37, 13110, C>T; 38, p.14104, T>C; **39, p.14408, C>T** ; **40, p.14805, C>T**; 41, p.14912, A>G; 42, p.15324, C>T; 43, p.16293, C>T; 44, 17247, T>C; **45, p.17747, C>T**; **46, p.17858, A>G**; **47, p.18060, C>T**; 48, p.18086, C>T; 49, p.18736, T>C; 50, p.18744, C>T; 51, p.18877, C>T; 52, p.18988, C>T; **53, p.18998, C>T**; 54, p.20755, A>C; 55, p.20844, C>T; 56, p.21137, A>G; 57, p.21830, G>T, 80xoG at WG; 58, p.22455, A>G; 59, p.22468, G>T; 60, p.22606, A>T; 61, p.23284, T>C; **62, p.23403, A>G** ; 63, p.24034, C>T; 64, p.25541, T>C; **65, p.25563, G>T, 8oxoG at WG** ; 66, 25575, A>C; 67, 25644, G>T, 8oxoG at WG ; 68, p.25688, C>A; 69, p.25979, G>T; 70, p.26088, C>T; **71, p.26144, G>T**; 72, p.26729, T>C; 73, p.28061, T>C; 74, p.28077, G>C; **75, p.28144, T>C** ; 76, p.28199, T>C; 77, p.28472, C>T; **78, p.28657, C>T**; 79, p.28708, C>T; 80, p.28836, C>T; 81, p.28842, G>T, 8oxoG at WG ; 82, p.28849, C>T; 83, p.28863, C>T; 84, p.28878, G>A; 85, p.28881-3, G>A,G>A,G>C; 86, p.28896, C>G; 87, p.29274, C>T; **88, p.29540. G>A, non-CDS gap**; 89, p.29543, G>C, non-CDS gap; 90, p.29700, A>G, 3'UTR; 91, p.29711, G>T, 8oxoG at WG, 3'UTR; 92, p.29742, G>A, 3'UTR.

nature of the deaminase driven host-parasite relationship. There are a small number of putative ROS mediated 8oxoG modifications found at WG sites that may contribute to the G>T SNVs. Thus, 6 of the L-241a Hap set are unmutated from Hu-1 (MT370845, MT370865, MT370867, MT370872, MT370873, MT370877; 8 within this haplotype set have one SNV difference from Hu-1 viz. MT370834 , MT370836, MT370838, MT370843, MT370876, MT370881, MT370884, MT370887; three have 2 SNV differences from Hu-1 (MT370859, MT370861, MT370885); and MT370880 and MT370863 have 3 and 4 SNV differences respectively.

There are also patterns within the VSD in Figure 10 showing probable P-to-P transfer of a lightly mutated sub-haplotype viz. defined by SNV differences from Hu-1 at p.11916, p.18998, and the change in the RNA only non-CDS gap at site p.29540 near the 3' end of the genome. Other cases of sequence conservation (L-241c) and likely sharing of sequences (indicative of P-to-P) and the addition of one SNV on transfer can be seen in the sequences MT370832, MT370846, MT370879, MT370883, MT370886.

The L-241d Hap set lacks the C>T SNV at p.14408. Even among this set there is very low further mutation, MT370853 is unmutated within the haplotype from Hu-1; and MT370858, MT370856, MT370855 have only one SNV difference from Hu-1 within the haplotype.

These patterns of very low mutation and sequence conservation among individual subjects is reminiscent of that observed in the major Wuhan epidemic, and on a smaller scale, on the *Grand Princess* cruise ship.

To further check on and confirm these observations of very low mutation and haplotype conservation a set of 22 sequences collected in NYC Mar 19-22 were aligned against Hu-1. The VSD plot is show in Figure 11. Once again, the L241a and variant L241a.1 are dominant members of this set. Five L241a sequences have no mutation (MT3701001, MT3701006, MT3701007, MT3701008, MT3701009) and the rest have one (MT3701010, MT3701003) or two to three mutations (MT370990, MT370995, MT3701005, MT370991). A similar haplotype mutation pattern applies to the L241a.1 subset, where four show no mutation (MT370997, MT370998, MT3701000, MT370993) and two one or two mutations (MT370989, MT3701004).

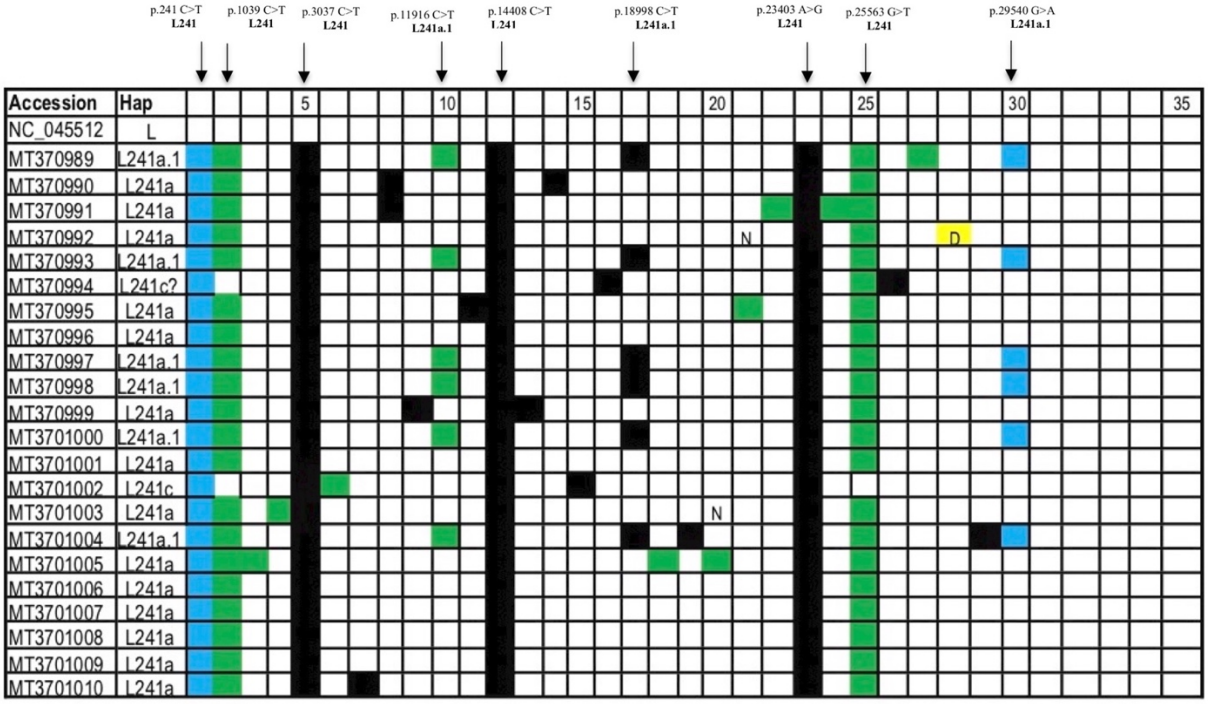

Figure 11 Variable site plot of SNV in each aligned sequence in a 22 sequence alignment for collections New York Mar 19 – 22 versus the Hu-1 ref NC_045512.2. Variable site number across the top, and Sequence ID down left hand side and Haplotype. The source data are from a screen shot record. The SNV key with respect to putative impact on protein structure of each SNV is discussed in text and Figure 1. L-241 and subset Hap designations as indicated (see Table 2). The variable site column number followed by SNV position in the alignment are in order with the L-241 sites listed in Table 2 in bold and arrowed. Other features whether in UTR, non-CDS region or G>T most likely caused by oxidation of G (8oxoG) are added given the complexity of the data set : **1, p.241, C>T, 5'UTR**; **2, p.1059, C>T**; 3, p.1917, C>T; 4, p.2222, T>C : **5, p.3037, C>T**; 6, p.4575, C>T; 7, p.10831, T>C; 8, p.10851, C>T; 9, p.11781, A>G; **10, p.11916, C>T;** 11, p.13548, C>T; **12, p.14408, C>T**; 13, p.16381, G>A; 14, p.18395, C>T; 15, p.18486, C>T; 16, p.18877, C>T; **17, p.18998, C>T**; 18, p.20005, G>A; 19, p.20553, A>G; 20, p.21458, T>C; 21, p.21485, G>T, 8oxoG at WG site?; 22, p.22530, C>T; **23, p.23403, A>G**; 24, p.25305, G>T, 8oxoG at WG site?; **25, p.25560/63, G>T, 8oxoG at WG site?;** 26, p.26681, C>T; 27, p.28115, T>C; 28, p.29367-29384, in-frame deletion (PTNPKKD); 29, p.28957, C>T; **30, p.29540, G>A, non-CDS gap near 3' UTR.**

SUMMARY and CONCLUSIONS

The mutational patterns presented here are for the origin of the COVID-19 virus (Dec-Jan 2020 China, mainly Wuhan), early spread to West Coast USA (mid to late February 2020), the exponential case rises in Spain (Mar 6-10, 2020), and New York just before (Mar 4-9, 2020) and during its key exponential rise in infections (Mar

14-19, 2020). The patterns evident within COVID-19 samples are thus for collections at key informative times and locations during the pandemic.

Our main finding has been the identification of a set of nucleotide sequence sites defining new COVID-19 RNA haplotypes - which we consider to have been created during the first infections with the Hu-1 sequence or its close relative. These key riboswitch sites (Table 2), in our view, are driven largely by the APOBEC and ADAR deaminases during the acute phase of infection in each individual: the virus varies largely at the RNA level, presumably to adjust its replicative efficacy to the biochemical and genetic background in which it finds itself. There is a surprising high level of sequence conservation in the functional status of the AA sequences in the mature proteins in the CDS mutations. As explained, the discovery of the generation of these haplotypes during the innate immune response to the virus, allows rational ordering of the data on COVID-19.

A reviewer has asked an important question, to paraphrase : "How likely is it that these apparent mild alterations in strain haplotypes are more adaptive, and could they contribute to conditions of P-to-P transmissions?" We have largely addressed this already - in epicenters during explosive outbreaks the "dominant" haplotype seems to be unmutated (or lightly mutated with largely synonymous changes). We have mentioned this may reflect in part the possible P-to-P sharing amongst vulnerable elderly co-morbids in a localised outbreak situation (nursing facility) of the *same unmutated sequence* (as they are expected to all have compromised or poor innate immune defences). As well as the examples we have seen already in Wuhan/China (Figure 2), the Californian cruise ship (Figure 3) and at the height of the NYC epidemic itself (Figure 10) we have also found a striking example of this phenomenon among a group of COVID-19 sequences collected on the same day, March 13 in Washington State and released as one upload batch on March 31 into NCBI Virus. These 21 sequences are MT262896-MT262916 inclusive. They are all of the Sa haplotype (Table 2). Nine have a synonymous T>C change at MC3 in a Tyrosine encoding UAU codon at p.11320. This could be a newly identified riboswitch generating a new P-to-P spreading haplotype on the Sa haplotype background? The only other changes are in MT262899 with a synonymous T>Cat p.13845; and in MT262915 a putative "benign" Non-Polar-to-Polar nonsynonymous G>T change at a WG site implying an 8oxoG modification. Otherwise, all other sequences are unmuted in relation to the Hu-1 reference. This is the most extreme example of this phenomenon uncovered and supports the implied speculation in the reviewer's question and thus adds to our understanding of how COVID-19 may spread in defined situations.

Our caveats are laid out - we lack clinical and patient data, nor do we have any evidence for dose at time of infections, particularly in the explosive epicentres of Wuhan and New York. We also do not have temporal

COVID-19 sequence data for individual patients to identify virus sequence changes in a single host during the acute phase of infection. We can only make inferences about such matters. The striking difference between the diversity of haplotypes, and extent of SNV patterns, between COVID-19 sequence collections from patients in Wuhan/China versus New York is striking. In our view such patterns are consistent with the prediction of the deaminase-driven riboswitch RNA haplotype model that we have used to order the data on COVID-19. i.e. the incoming virus adapts by locking in an RNA haplotype suitable for rapid replication in that host cell under selection. We would expect the ethnic genetic diversity in New York City to far exceed almost pure Chinese genotypes in Wuhan and throughout China.

In our opinion, the implications for vaccine design should incorporate boosting innate immunity, for example by BCG vaccination as for the lung infection tuberculosis. Since BCG is a widely accepted non-specific activator of innate immunitywe might expect it to contribute to elevated APOBEC and ADAR expression and thus mutagenic attacks on the virus genome. Such vaccinations logically imply that innate immune responses, and boosting mutagenic APOBEC and ADAR levels could be an important part of vaccine design.

A legitimate question, posed by a reviewer has been whether we can systematically independently test the hypothesis or compare an array of evidence supporting or against the plausibility of the haplotype hypothesis. We therefore ask can we find situations where the "world set" of the main COVID-19 haplotypes in the Northern Hemisphere may be sampled by travellers coming into Australia? This is important as all outbreaks in Australia have been by travellers returning home from the Northern Hemisphere. Over the period of the pandemic Victoria, Australia has been the destination for Australians returning home as 'COVID-19 refugees" from Northern Hemisphere locations - Europe, China, USA. On arrival they have all been quarantined for two weeks to limit spread of the disease to local citizens. We ask: Are the main set of haplotypes as tabulated here (Table 2) evident among these incoming travellers from exposure to the Northern Hemisphere pandemic? First there is no real dominant haplotype among these COVID-19 +ve travellers to Victoria. The distribution of major haplotypes (Table 2) in these incoming travellers for the haplotypes L, Ln, L241a, L241c , S and Sa were 13, 19, 6, 13, 15, and 11 a reflection of the main haplotypes in circulation in China, USA and Europe in this period (collections January 24- March 15 2020, MT450919-MT450995).

In follow up studies we plan to definitively identify the APOBEC and ADAR variants and isoforms responsible for the mutagenesis of the COVID-19 genome during the innate immune response phase in infected COVID-19 patients as demonstrated for HCV and ZIKV (Lindley and Steele 2018). The present study also has wider implications for the actual origins of COVID-19 pandemic beginning in Wuhan, China. Those analyses will be

pursued in other publications focusing on the implications of these data for the origin and global spread of this suddenly emergent pandemic disease.

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., et al (2013) Signatures of mutational processes in human cancer Nature 500, 415-421 https://www.ncbi.nlm.nih.gov/pubmed/23945592

Andersen, K. (2020) Clock and TMRCA based on 27 genomes . Novel 2019 coronavirus http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347

Andino, R., and Domingo, E. (2015). Viral quasispecies. Virology 479-480, 46-51. DOI: 10.1016/j.virol.2015.03. 022

Buhr, F., Jha, S., Thommen, M., Mittlestael, J., Kutz, F., Schwalbe, H., Rodina, M.V., and Komar, A.A. (2016). Synonymous codons direct cotranslational folding towards different protein conformations. Mol. Cell 61, 341-351. http://dx.doi.org/10.1016/j.molcel.2016.01.008

Dorp, L.V., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., et al. (2020a) Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 83, 104351 https://doi.org/10.1016/j.meegid.2020.104351

Dorp, L.V., Richard, D., Tan, C.C.S., Shaw, L.P., Acman, M., and Balloux, F. (2020b) No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. BioRxiv reprint doi: https://doi.org/10.1101/2020.05.21.108506 posted May 21 , 2020

Eigen, M., and Schuster, P. (1979). The Hypercycle: A Principle of Natural Self-Organization. Springer, Berlin.

Eifler, T., Pokharel, S., and Beal, P.A. (2013). RNA-Seq analysis identifies a novel set of editing substrates for human ADAR2 present in Saccharomyces cerevisiae. Biochemistry 52, 7857-7869. DOI: 10.1021/bi4006539

Franklin, A., Steele, E.J., and Lindley, R.A. (2020) A proposed reverse transcription mechanism for (CAG)n and similar expandable repeats that cause neurological and other diseases. Heliyon 6, e03258

Gilbert, S.D. and Lafontaine, R.T. (2006) Riboswitches: Fold and Function. Chemistry & Biology, 13 , 857-868. https://doi.org/10.1016/j.chembiol.2006.08.002

Li, H., Stoddard, M.B., Wang, S., Blair, L.M., Giorgi, E.E., Parrish, E.H., Learn, G.H., Hraber, P., Goepfert, P.A., Saag, M.S., et al. (2012). Elucidation of hepatitis C virus transmission and early diversification by single genome sequencing. PLoS Pathog 8(8), e1002880. doi: 10.1371/journal.ppat.1002880

Lindley, R.A. (2013). The importance of codon context for understanding the Ig-like somatic hypermutation strand- biased patterns in TP53 mutations in breast cancer. Cancer Genetics 5, 2619-2640. DOI: 10.1016/j.cancergen. 2013.05.016

Lindley, R.A., and Hall, N.E 2018. (2018) APOBEC and ADAR deaminases may cause many single nucleotide polymorphisms curated in the OMIM database. Mutat Res Fund Mol Mech Mutagen 810, 33-38. https://doi.org/10.1016/j.mrfmmm.2018.03.008

Lindley, R.A., and Steele, E.J. (2018) ADAR and APOBEC editing signatures in viral RNA during acute-phase Innate Immune responses of the host-parasite relationship to Flaviviruses. Research Reports 2:e1- e22. doi:10.9777/rr.2018.10325.

Lindley, R.A., Humbert, P., Larmer, C., Akmeemana, E.H., and Pendlebury, C.R.R. (2016). Association between targeted somatic mutation (TSM) signatures and HGS-OvCa progression. Cancer Med. 5, 2629-2640. DOI: 10.1002/cam4.825

Pirakitikulr, N., Kohlway, A., Lindenbach, B.D., and Pyle, A.M. (2016). The coding region of the HCV genome contains a network of regulatory RNA structures. Mol. Cell 61, 1-10. http://dx.doi.org/10.1016/j.molcel.2016.01.024

Sharma, S., Patnaik, S.K., Taggart, R.T., Kannisto, E.D., Enriquez, S.M., Gollnick, P., and Baysal, B.E. (2015). APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nat. Commun. 6. 6881. doi: 10.1038/ncomms7881 [SEP]

Sharma, S., Santosh, K.. Patnaik, S.K., Kemera, Z., and Baysal, B.E. (2016a). Transient overexpression of exogenous APOBEC3A causes C-to-U RNA editing of thousands of genes. RNA Biology 5, 1-8 doi: 10.1080/15476286.2016.1184387

Sharma, S., Patnaik, S.K., Taggart, R.T. and Basal, B.E. (2016b) The double-domain cytidine deaminase APOBEC3G is a cellular site- specific RNA editing enzyme. Sci. Rep. 6, 39100; doi: 10.1038/srep39100 (2016).

Schneider, W.M., Chevillotte, M.D. and Rice, C.M. (2014) Interferon-stimulated genes: acomplex web of host defenses. Annu. Rev. Immunol. 232, 513–545 doi:10.1146/annurev-immunol-032713-120231

Schoggins, J.W. and Rice, C.M. (2011) Interferon-stimulated genes and their antiviral effector functions Curr. Opin. Virol. 1, 519 - 525. doi: 10.1016/j.coviro.2011.10.008

Shi ,Y., Wang, Y., Shao, C., Huang, J., Gan, J., Huang, X., et al. (2020). COVID-19 Infection: The Perspectives on Immune Responses Cell Death Differ.27, 1451-1454. doi: 10.1038/s41418-020-0530-3.

Steele, E.J. and Lindley, R.A. (2017) ADAR deaminase A-to-I editing of DNA and RNA moieties of RNA:DNA Hybrids has implications for the mechanism of Ig somatic hypermutation. DNA Repair 55, 1 - 6. doi: 10.1016/j.dnarep.2017.04.004

Stoddard, M.B., Li, H., Wang, S., Saeed, M., Andrus, L., Ding, W., Jiang, X., Learn, G.H., von Schaewen, M., Wen, J., Goepfert, P.A., Hahn, B.H., Ploss, A., Rice, C.M., and Shaw G.M. (2015). Identification, molecular cloning, and analysis of full-length hepatitis C virus transmitted/founder genotypes 1, 3, and 4. mBio 6(2), e02518-14. doi: 10.1128/mBio.02518-14

Tan, Z., Zhang, W., Shi, Y. and Wang, F. (2015) RNA Folding: Structure Prediction, Folding Kinetics and Ion Electrostatics Adv Exp Med Biol. 2015;827:143-83. doi: 10.1007/978-94-017-9245-5_11. PMID: 25387965

Tang, X., Wu, C., Li, X., Song, Y., et al. (2020) On the origin and continuing evolution of SARS-CoV-2. National Science Review (2020). nwaa036, https://doi.org/10.1093/nsr/nwaa036

Thimme, R., Binder, M., and Bartenschlager, R. (2012). Failure of innate and adaptive immune responses in controlling hepatitis C virus infection. FEMS Microbiol. Rev. 36, 663– 683. DOI: 10.1111/j.1574-6976.2011.00319.x

Widom, J.R. Nedialkov, Y.A., Rai, V., Hayes, R.L., Brooks,C.L., 3rd, Artsimovitch ,I. and Walter, N.G. (2018) Ligand modulates cross-coupling between riboswitch folding and transcriptional pausing .Mol Cell. 72, 541-552.e6. doi: 10.1016/j.molcel.2018.08.046. PMID: 30388413

Yang, D. and Leibowitz, J.L. (2015) The structure and functions of coronavirus genomic 3′ and 5′ ends Virus Research 206 :120–133. https://doi.org/10.1016/j.virusres.2015.02.025