bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Consistent and High-Frequency Identification of an Intra-Sample Genetic Variant of SARS-CoV-2 with Elevated Fusogenic Properties

Lynda Rocheleau[1], Geneviève Laroche[1], Kathy Fu[1,2,3], Marceline Côté[1,2,3], Patrick M Giguère[1], Marc-André Langlois[1,2,*] and Martin Pelchat[1,2,*]

[1] Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8M5, Canada.

[2] uOttawa Center for Infection, Immunity and Inflammation (CI3), Ottawa, Canada.

[3] Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Canada.

[*] Corresponding authors: Phone: (613) 562-5800 ext. 8846

E-mail: mpelchat@uottawa.ca

Phone: (613) 562-5800 ext. 7110

E-mail: langlois@uottawa.ca

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has a genome comprised of a ~30K nucleotides non-segmented, positive single-stranded RNA. Although its RNA-dependent RNA polymerase exhibits exonuclease proofreading activity, viral sequence diversity can be induced by replication errors and host factors. These variations can be observed in the population of viral sequences isolated from infected host cells and are not necessarily reflected in the genome of transmitted founder viruses. We profiled intra-sample genetic diversity of SARS-CoV-2 variants using 15,289 high-throughput sequencing datasets from infected individuals and infected cell lines. Most of the genetic variations observed, including C->U and G->U, were consistent with errors due to heat-induced DNA damage during sample processing, and/or sequencing protocols. Despite high mutational background, we confidently identified intra-variable positions recurrent in the samples analyzed, including several positions at the end of the gene encoding the viral S protein. Notably, most of the samples possesses a C->A missense mutation resulting in the S protein lacking the last 20 amino acids (SΔ20). Here we demonstrate that SΔ20 exhibits increased cell-to-cell fusion and syncytia formations. Our findings are suggestive of the consistent emergence of high-frequency viral quasispecies that are not horizontally transmitted but involved in intra-host infection and spread.

Author summary

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its associated disease, COVID-19, has caused significant worldwide mortality and unprecedented economic burden. Here we studied the intra-host genetic diversity of SARS-CoV-2 genomes and identified a high-

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

frequency and recurrent non-sense mutation yielding a truncated form of the viral spike protein, in both human COVID-19 samples and in cell culture experiments. Through the use of a functional assay, we observed that this truncated spike protein displays an elevated fusogenic potential and forms syncytia. Given the high frequency at which this mutation independently arises across various samples, it can be hypothesized that this deletion mutation provides a selective advantage to viral replication and may also have a role in pathogenesis in humans.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Observed for the first time in 2019, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its associated disease, COVID-19, has caused significant worldwide mortality and unprecedented economic burden. SARS-CoV-2 is an enveloped virus with a genome comprised of a ~30K nucleotides non-segmented, positive-sense single-stranded RNA (vRNA) [1,2]. The virus is composed of four main structural proteins, encoded at the last 3' terminal third of the viral genome: the spike glycoprotein (S), membrane (M), envelope (E) and the nucleocapsid (N) [3–5]. Attachment to the host receptor angiotensin-converting enzyme 2 (ACE2) is mediated by the S protein expressed on the surface of the virion [6]. Following its association, the S protein is cleaved into two separate polypeptides (S1 and S2), which triggers the fusion of the viral particle with the cellular membrane [7,8]. Once inside a cell, its RNA-dependent RNA polymerase (RdRp), which is encoded in the first open reading frame of the viral genome [9], carries out transcription and replication of the vRNA genome. In addition, mRNAs coding for the structural proteins (*e.g.*, S, M, E and N) are expressed by subgenomic RNAs [9]. Once translated, the S, M and E proteins localize and accumulate in the endoplasmic reticulum–Golgi intermediate compartment (ERGIC) [10]. The S protein possesses an endoplasmic reticulum retrieval signal (ERRS) at its carboxy terminus, which is required for this localization [11]. At this location, the N protein associates with the viral genome and assembles into virions, which are transported along the endosomal network and released by exocytosis [9]. It was also observed for several coronaviruses that the S protein can localize to the cell surface and mediate cell fusion between adjacent cells, producing multinucleated cells or syncytia [8,12,13].

Genomic sequencing of SARS-CoV-2 vRNA from infected populations has demonstrated genetic heterogeneity [14–20]. Several recurrent mutations have been identified in consensus

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequences, and the geographical distribution of clades was established. Because they induce an abundance of missense rather than synonymous or non-sense mutations, it was suggested that regions of the SARS-CoV-2 genome were actively evolving and might contribute to pandemic spreading [20]. It was observed that variations are mainly comprised of transition mutations (purine->purine or pyrimidine->pyrimidine) with a prevalence of C->U transitions and might occur within a sequence context reminiscent of APOBEC-mediated deamination (*i.e.,* [AU]C[AU]; [21,22]). Consequently, it was proposed that host editing enzymes might be involved in coronavirus genome editing [23,24].

Consensus mutations are only part of the genetic landscape in regard to RNA viruses. Replication of RNA viruses typically produces quasispecies in which the viral RNA genomes do not exist as single sequence entity but as a population of genetic variants [25]. These mutations are most frequently caused by the error-prone nature of each of their respective viral RdRps and by host RNA editing enzymes, such as APOBECs and ADARs [26]. However, the RdRp complex of large RNA viruses, such as coronaviruses, sometimes possess exonuclease proofreading activity, and consequently have lower error rates [25,27]. Quasispecies may sometimes exhibit diminished replicative fitness or deleterious mutations and exert different roles that are not directly linked to viral genomic propagation [28]. Mutations that form the intra-host genetic spectrum have been shown to help viruses evade cytotoxic T cell recognition and neutralizing antibodies and also render viruses more resistant to antiviral drugs [28]. These mutations can also be involved in modulating the virulence and transmissibility of the quasispecies [28].

In this study, we focussed on assessing intra-genetic variations of SARS-CoV-2. We analyzed high-throughput sequencing datasets to profile the sequence diversity of SARS-CoV-2 variants within distinct sample populations. We observed high genetic intra-variability of the viral

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genome. By comparing variation profiles between samples from different donors and cell lines, we identified highly conserved subspecies that independently and recurrently arose in different datasets and, therefore, in different individuals. We further analyzed the dominant variant SΔ20 in a functional assay and demonstrate that this truncated spike protein enhances syncytium formation. Here we provide evidence for the existence of a consistently emerging variant identified across geographical regions that may influence intra-host SARS-CoV-2 infectivity and pathogenicity.

Results

High intra-genetic variability of the SARS-CoV-2 genome in infected individuals.

To assess the extent of SARS-CoV-2 sequence intra-genetic variability, we analyzed 15,224 publicly available high-throughput sequencing datasets from infected individuals. The raw sequencing reads were mapped to the SARS-CoV-2 isolate Wuhan-Hu-1 reference genome, and the composition of each nucleotide at each position on the viral genome was generated. Consensus sequences were produced for each dataset and the nucleotide composition for each position were compared to respective consensuses. To reduce the number of variations due to amplification bias and sequencing errors, duplicated reads were combined, and only positions mapped with a sequencing depth of 50 reads and having at least 5 reads with variations compared to the sample consensus were considered. Overall, we identified 301,742 variations from 11,362 samples located on 26,113 positions of the 29,903 nt SARS-CoV-2 genome. We observed an average of 26.6+/-132.0 variable nucleotides per sample (ranging from 1 to 5295 variations/sample; Fig 1A).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Analysis of the type of intra-genetic variations present in SARS-CoV-2 samples from infected individuals.

The analysis of the type of nucleotide changes within samples revealed that 52.2% were transitions (either purine->purine or pyrimidine->pyrimidine) and 47.8% were transversions (purine->pyrimidine or pyrimidine->purine). Notably, the highest nucleotide variations corresponded to C->U transitions (43.5%) followed by G->U (28.1%) transversion (Fig 1B), both types encompassing 71.6% of all variations. Since editing by host enzymes depends on the sequence context, we extracted two nucleotides upstream and downstream from each genomic position corresponding to variations and generated sequence logos. Our results indicated a high number of As and Us around all variation types and sites (62.1+/-3.4%; Fig 1B). Because SARS-CoV-2 is composed of 62% A/U, this suggests the observed number of As and Us around variation sites are mainly due to the A/U content of the viral genome, that no motifs are enriched around these sites and that these intra-genetic variations are likely not originating from host editing enzymes.

Identification of recurrent genetic variants of SARS-CoV-2 in samples from infected individuals.

To identify biologically relevant intra-genetic variations, we examined the variable positions that are recurrent in the samples analyzed. The variable positions were tabulated for each sample and then recurrent intra-genetic variations were calculated as percentages of samples containing variation at each position. Most variations are distributed homogeneously on the viral genome and most are poorly shared amongst samples (Fig 1C and 1D). However, our analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

reveals 15 recurrent intra-variations shared by at least 5% of the samples analyzed (Fig 1C, above blue line; Table 1). Amongst these, four transversions (at nt 25,324, 25,334, 25,336 and 25,337) located at the 3' end of the S gene are the most recurrent variations (inset of Fig 1C and Table 1). Three of these transversions (at nt 25,334, 25,336 and 25,337) correspond to missense mutations: E1258D (46.4%), E1258Q (27.6%) and D1259H (20.1%). Interestingly the most observed variation (at nt 25,324) is shared by 58.7% of the samples (6,668 of the 11,362 samples) and corresponds to a C->A transversion producing a nonsense mutation at amino acid 1,254 of the S protein (Fig 1C and 1D, red line; Fig 2B, red rectangle). The resulting S protein lacks the last 20 amino acids (SΔ20), which includes the ERRS motif at its carboxy terminus (Fig 2B, white letters on a black background). Amongst the sample with this intra-genetic variation, this C->A transversion represents from 2.9% to 42.4% of the subspecies identified (mean or 8.2+/-2.9%; Fig 2C and Table 1).

Table 1: Recurrent SARS-CoV-2 genome intra-variations shared by at least 5% infected individuals. Frequency distributions were calculated using data similar to Fig 2C. The variations are sorted by their recurrence, with the most shared variation at the top of the table.

Position (nt)	Proportion of Samples (%)	Type of Variation	Gene	Amino acid	Consensus codon	Variant codon	Consensus amino acid	Variant amino acid	Context (-2 to +2)	Frequency distribution (% of population)			
										Mean	Stand. Dev.	Min	Max
25324	58.69	C->A	S	1254	UG**C**	UG**A**	CYS (C)	STOP	UG**C**AA	8.19	2.89	2.86	42.37
25336	46.37	A->C	S	1258	GA**A**	GA**C**	GLU (E)	ASP (D)	GA**A**GA	6.38	2.10	2.42	29.09
25334	27.57	G->C	S	1258	**G**AA	**C**AA	GLU (E)	GLN (Q)	AU**G**AA	4.76	1.63	2.03	22.81
25337	20.11	G->C	S	1259	**G**AC	**C**AC	ASP (D)	HIS (H)	AA**G**AC	4.68	2.12	2.07	28.57
29187	10.95	C->U	N	305	G**C**A	G**U**A	ALA (A)	VAL (V)	UG**C**AC	3.35	2.53	1.81	46.91
29188	10.68	A->G	N	305	GC**A**	GC**G**	ALA (A)	ALA (A)	GC**A**CA	3.32	2.56	1.79	46.91
18591	10.21	C->G	ORF1ab	6108	GU**C**	GU**G**	VAL (V)	VAL (V)	GU**C**UU	3.78	0.85	2.54	7.96
11874	10.02	U->C	ORF1ab	3870	G**U**A	G**C**A	VAL (V)	ALA (A)	AG**U**AG	4.21	2.27	2.08	38.55
15965	8.12	G->U	ORF1ab	5233	U**G**U	U**U**U	CYS (C)	PHE (F)	CU**G**UU	3.01	2.48	1.88	44.19

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

29039	7.95	A->U	N	256	A**A**G	A**U**G	LYS (K)	MET (M)	CU**A**AG	4.26	1.66	2.06	21.74
6696	7.19	C->U	ORF1ab	2144	C**C**U	C**U**U	PRO (P)	LEU (L)	GC**C**UU	3.59	3.31	1.92	48.85
28253	6.51	C->U	ORF8	120	UU**C**	UU**U**	PHE (F)	PHE (F)	UU**C**AU	8.58	7.98	1.86	48.42
635	5.18	C->U	ORF1ab	124	**C**GU	**U**GU	ARG (R)	CYS (C)	UU**C**GU	8.72	6.50	1.92	48.00
9502	5.17	C->U	ORF1ab	3079	GC**C**	GC**U**	ALA (A)	ALA (A)	GC**C**UU	3.98	3.29	1.99	49.40
25323	5.14	G->C	S	1254	U**G**C	U**C**C	CYS (C)	SER (S)	CU**G**CA	4.27	1.84	2.12	16.95

155

156

157 **Analysis of intra-genetic variations present in SARS-CoV-2 samples from infected cells.**

158 To further investigate variations in a more controlled system, and to determine whether
159 host proteins are involved in SARS-CoV-2 genome editing, we used 65 high-throughput
160 sequencing datasets generated in a recent transcription profiling study of several cell lines infected
161 with SARS-CoV-2 [29]. Firstly, we mapped raw sequencing reads to the human genome to assess
162 host modifying enzyme expression. For all cell lines, normalized counts for mRNAs corresponding
163 to most host modifying enzymes were very low or non-detected (Fig 3), suggesting that these cell
164 lines poorly expressed these host editing proteins. As above, the raw sequencing reads from
165 infected cells were mapped to the SARS-CoV-2 genome sequence, the composition of each
166 nucleotide at each position on the viral genome were generated, and nucleotide variations when
167 compared to respective consensus sequences were calculated. Because the sequencing depths of
168 the samples were low, we considered positions mapped by at least 20 reads and having at least 2
169 reads with variation compared to the sample consensus. In the samples derived from infected cells,
170 we observed 29.7% and 70.3% of transitions and transversions, respectively. Similar to
171 observations in samples from infected individuals, the highest nucleotide variations corresponded
172 to G->U transversions (26.1%) and C->U (21.6%) transitions (Fig 4B). We then analyzed
173 nucleotide compositions two nucleotides upstream and downstream of the intra-genetic variations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

As above, a high number of A/U (57.8+/-7.7%) were present around sites showing variation (Fig 4B), consistent with the 62% A/U composition of the SARS-CoV-2 genome, indicating no enrichment of sequence motifs around these sites, except for the expected high number of As and Us.

We then examined the intra-genetic variable positions that are recurrent amongst the cell lines analyzed. We identified 29 positions within the viral populations showing intra-genetic variation enrichment in at least 10% of the cell cultures and most of them are located on structural genes, which are encoded at the last 3' terminal third of the viral genome (Fig 4C and 4D). Similar to our observation from the samples from infected individuals, a cluster of recurrent variations is located at the 3'end of the S gene, including the C->A transversion at position 25,324 shared in 58.9% of the cell lines analyzed (Fig 4C and 4D, red line; Table 2). Overall, our results indicate consistent results between intra-genetic variations observed in infected cell lines and in samples from infected individuals, including the presence the viral subspecies resulting in a S protein truncated of its last 20 amino acids (SΔ20).

Table 2: Recurrent SARS-CoV-2 genome intra-variations shared by at least 10% infected cell cultures. Frequency distributions were calculated using data similar to Fig 2C. The variations are sorted by their recurrence, with the most shared variation at the top of the table.

Position (nt)	Proportion of Samples (%)	Type of Variation	Gene	Amino acid	Consensus codon	Variant codon	Consensus amino acid	Variant amino acid	Context (-2 to +2)	Frequency distribution (% of population)			
										Mean	Stand. Dev.	Min	Max
28853	82.14	U>A	N	194	**U**CA	**A**CA	SER (S)	THR (T)	GU**U**CA	8.16	7.99	1.52	35.42
25336	58.93	A>C	S	1258	GA**A**	GA**C**	GLU (E)	ASP (D)	GA**A**GA	21.77	5.81	12.00	42.22
25324	58.93	C>A	S	1254	UG**C**	UG**A**	CYS (C)	STOP	UG**C**AA	25.21	7.37	12.00	42.37
23525	58.93	C>U	S	655	**C**AU	**U**AU	HIS (H)	TYR (Y)	AA**C**AU	8.36	3.35	3.49	16.67
25337	58.93	G>C	S	1259	**G**AC	**C**AC	ASP	HIS	AA**G**AC	20.43	4.65	12.86	35.56

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

							(D)	(H)					
25334	57.14	G>C	S	1258	**G**AA	**C**AA	GLU (E)	GLN (Q)	AU**G**AA	12.98	6.44	3.08	22.81
25381	55.36	A>C	S	1273	AC**A**	AC**C**	THR (T)	THR (T)	AC**A**UA	26.73	5.21	8.33	37.50
22343	55.36	G>C	S	261	**G**GU	**C**GU	GLY (G)	ARG (R)	CU**G**GU	6.51	2.84	2.27	13.79
25323	55.36	G>C	S	1254	U**G**C	U**C**C	CYS (C)	SER (S)	CU**G**CA	9.03	4.13	2.82	17.24
25331	55.36	G>U	S	1257	**G**AU	**U**AU	ASP (D)	TYR (Y)	UU**G**AU	6.35	2.94	2.60	13.33
27883	53.57	C>U	ORF7b	43	G**C**C	G**U**C	ALA (A)	VAL (V)	CG**C**CU	6.74	2.43	2.40	11.19
27882	53.57	G>C	ORF7b	43	**G**CC	**C**CC	ALA (A)	PRO (P)	AC**G**CC	6.88	2.52	2.40	11.67
25296	51.79	A>C	S	1245	A**A**G	A**C**G	LYS (K)	THR (T)	CA**A**GG	7.16	2.38	2.94	12.96
23606	51.79	C>U	S	682	**C**GG	**U**GG	ARG (R)	TRP (W)	CU**C**GG	31.65	12.73	3.95	48.15
25327	48.21	A>U	S	1255	AA**A**	AA**U**	LYS (K)	ASN (N)	AA**A**UU	5.31	2.23	2.60	9.43
23616	48.21	G>A	S	685	C**G**U	C**A**U	ARG (R)	HIS (H)	AC**G**UA	21.11	10.20	2.38	38.71
23616	44.64	G>C	S	685	C**G**U	C**C**U	ARG (R)	PRO (P)	AC**G**UA	21.11	10.20	2.38	38.71
21550	41.07	A>C	ORF1ab	7095	**A**AC	**C**AC	ASN (N)	HIS (H)	AC**A**AC	39.31	9.06	18.75	50.00
21551	41.07	A>U	ORF1ab	7095	A**A**C	A**U**C	ASN (N)	ILE (I)	CA**A**CU	38.79	9.19	18.75	50.00
25286	39.29	A>U	S	1242	**A**GU	**U**GU	SER (S)	CYS (C)	GU**A**GU	4.05	1.27	2.63	7.84
25314	39.29	G>U	S	1251	G**G**A	G**U**A	GLY (G)	VAL (V)	UG**G**AU	4.15	1.54	2.56	7.14
27134	32.14	U>C	M	204	UA**U**	UA**C**	TYR (Y)	TYR (Y)	UA**U**AA	3.17	1.16	1.87	5.75
22206	30.36	A>G	S	215	G**A**U	G**G**U	ASP (D)	GLY (G)	UG**A**UC	4.30	1.64	2.44	9.21
25316	30.36	U>C	S	1252	**U**CC	**C**CC	SER (S)	PRO (P)	GA**U**CC	4.85	1.89	2.67	9.38
26542	28.57	C>U	M	7	A**C**U	A**U**U	THR (T)	ILE (I)	UA**C**UA	11.69	15.28	1.96	47.01
25296	26.79	A>U	S	1245	A**A**G	A**U**G	LYS (K)	MET (M)	CA**A**GG	7.16	2.38	2.94	12.96
25277	25.00	A>U	S	1239	**A**GU	**U**GU	SER (S)	CYS (C)	CC**A**GU	3.50	0.69	2.67	5.06
17641	25.00	G>A	ORF1ab	5792	**G**CU	**A**CU	ALA (A)	THR (T)	CA**G**CU	4.28	1.83	2.56	9.09
25331	25.00	G>C	S	1257	**G**AU	**C**AU	ASP (D)	HIS (H)	UU**G**AU	6.35	2.94	2.60	13.33
25334	25.00	G>U	S	1258	**G**AA	**U**AA	GLU (E)	STOP	AU**G**AA	12.98	6.44	3.08	22.81
25323	23.21	G>U	S	1254	U**G**C	U**U**C	CYS (C)	PHE (F)	CU**G**CA	9.03	4.13	2.82	17.24
25316	19.64	U>G	S	1252	**U**CC	**G**CC	SER (S)	ALA (A)	GA**U**CC	4.85	1.89	2.67	9.38
25312	17.86	U>G	S	1250	UG**U**	UG**G**	CYS (C)	TRP (W)	UG**U**GG	3.50	0.72	2.56	4.76
20956	14.29	C>U	ORF1ab	6897	**C**UU	**U**UU	LEU (L)	PHE (F)	AU**C**UU	14.13	14.84	2.38	35.48
21550	12.50	A>C	ORF1ab	7095	**A**AC	**C**AC	ASN (N)	HIS (H)	AC**A**AC	39.31	9.06	18.75	50.00
21551	12.50	A>U	ORF1ab	7095	A**A**C	A**U**C	ASN (N)	ILE (I)	CA**A**CU	38.79	9.19	18.75	50.00
25273	10.71	G>C	S	1237	AU**G**	AU**C**	MET (M)	ILE (I)	AU**G**AC	2.95	0.44	2.53	3.77

SARS-CoV-2 SΔ20 increases cell-to-cell fusion and the size of syncytia.

For several coronaviruses, the S protein mediates syncytia formation [8,12,13]. To investigate the effect of SΔ20 on cell-to-cell fusion, HEK293T cells stably expressing human

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ACE2 were co-transfected with a plasmid expressing GFP and plasmids expressing or not wild-type S or SΔ20 under a cytomegalovirus (CMV) major immediate early promoter. As expected, in the absence of the S protein (*i.e.*, pCAGGS alone) syncytia formation was not observed (Fig 5A). Consistent with previous findings [8], we observed the presence of large cytoplasmic masses containing GFP in the presence of the wild-type S protein, indicating induction of cell-to-cell fusion (Fig 5A and 5B). Our results not only indicate that SΔ20 also induces fusion, but that the cytoplasmic masses are larger than the wild-type S protein (Fig 5A and 5B). To complement this approach, we quantified cell fusion using a bi-molecular fluorescence complementation (BiFc) assay composed of target cells containing GCN4 leucine zipper-Venus1 (ZipV1) with or without myc-ACE2, and effector cells containing GCN4 leucine zipper-Venus2 (ZipV2) with or without S/ SΔ20 (Fig 5C). Two hours after incubating the two cell populations, we observed an increase in fluorescence signal in the presence of wild-type S and ACE2, indicating fluorescence complementation and thus cell-to-cell fusion (Fig 5D). Consistent with our microscopy observations, fluorescence complementation was increased in the presence of SΔ20. Taken together, our results indicate that SΔ20 displays increased syncytia formation in HEK293T cells expressing ACE2 as compared to the wild-type S protein.

Discussion

Previous analyses of SARS-CoV-2 nucleotide variations indicated the prevalence of C->U transitions suggesting that the viral genome was actively evolving and those host editing enzymes, such as APOBECs and ADARs, might be involved in this process [23,24]. Although instructive on the role of host involvement in SARS-CoV-2 genome evolution, these studies were performed

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on consensus sequences (*i.e.,* one per sample) and explore only part of the genetic landscape of this RNA virus. Here, we used a large number of high-throughput sequencing datasets to profile the intra-sample sequence diversity of SARS-CoV-2 variants, both in infected individuals and infected cell lines. We observed extensive genetic variability of the viral genome, including a high number of transversions, and identified several positions with recurrent intra-variability in the samples analyzed. Notably, most of the samples possessed a C->A missense mutation resulting in the S protein lacking the last 20 amino acids (SΔ20) that increases cell-to-cell fusion and syncytia formations.

Most intra-sample variations are distributed homogeneously across the viral genome, are not conserved or recurrent amongst samples, and a large number of them are C->U or G->U mutations. Previous analyses of SARS-CoV-2 sequence variations proposed that host editing enzymes might be involved in coronavirus transition editing based on results showing that C->U transitions occur within a sequence context reminiscent of APOBEC1-mediated deamination (*i.e.*, [AU]C[AU]) [21–24]. Here, we investigated nucleotide compositions at each variation site and observed a high number of As and Us around all variation types and sites. However, since the SARS-CoV-2 genome is 62% A/U-rich and similar percentages of As and Us were observed around all variations, we concluded that no motifs are enriched around these variations in the viral subspecies analyzed. Consequently, our results cannot support that host editing enzymes are a major source of these intra-sample variations.

Although it is possible that host RNA-editing enzyme are responsible for the occurrence of some variations, C->U transitions and G->U transversions are also generally associated with nucleotide deamination and oxidation, respectively [30–37]. It is common practice to thermally inactivate SARS-CoV-2 samples before performing RNA extractions followed by RT-PCR and

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequencing [38]. Heating samples can form free radicals, such as 8-hydroxy-20-deoxyguanine (8-Oxo-dG), that could cause high levels of C->A and G ->U mutations and promote the hydrolytic deamination of C->U [30–33,35,37,39,40]. It was previously reported that these types of mutations occur at low frequency, that they are mostly detected when sequencing is performed on only one DNA strand and that they are highly variable across independent experiments [32,34]. Consequently, most transversions observed in our analysis are likely due to heat-induced damage, RNA extraction, storage, shearing and/or RT-PCR amplification errors. However, we identified several positions with intra-sample variability recurrent in several independent samples, both from infected individuals and infected cells. They were detected at moderate to high frequencies, ranging from 2.5% to 39.3% per sample (Table 1 and 2), and most were derived from pair-end sequencing (90.7%) in which the two strands of a DNA duplex were considered. Thus, it is likely that these variations are genuine and represent hot spots for SARS-CoV-2 genome intra-sample variability.

Amongst the variable positions identified in infected cells, most of them are located at the last 3' terminal third of the viral genome. These cells were infected with a high number of viruses (*i.e.,* high multiplicity of infection; MOI) for 24h [29]. The presence of several variations at positions in the region coding for the main structural proteins likely reflects that this is a region with increased transcriptional activity due to the requirement of producing their encoded mRNAs from sub-genomic negative-sense RNAs [9].

Interestingly, a cluster of variations located at the 3'end of the S gene was observed for the two datasets analyzed. They correspond to four transversions located at the 3'end of the S gene and are shared by a large proportion of the samples. Three of these correspond to missense mutations changing the charged side chains of two amino acids (E1258D, E1258Q and D1259H).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Notably, most of the samples possess a variability at position 25,324 producing a nonsense mutation at amino acid 1,254 of the S protein. The resulting protein lacks the last 20 amino acids (SΔ20) and thus does not include the ERRS motif at its carboxy terminus. For SARS-CoV-1, the ERRS domain localizes the S protein to the ERGIC and facilitates its incorporation into virions [11]. Deletion of this motif might cause the S protein of SARS-CoV-2 to accumulate to the plasma membrane and increase the formation of large multinucleated cells known as syncytia. Consistent with these observations, our results indicate an increase in syncytia formation frequency and size with SΔ20 as compared to the complete S protein. Although the biological function of SΔ20 is unknown, it is possible that the formation of syncytia facilitates the spread of the virus directly to neighboring cells. SΔ20 might also increase virus replication, as similar mutants (SΔ18, SΔ19 and SΔ21) were recently reported to increase both infectivity and replication of vesicular stomatitis virus (VSV) and human immunodeficiency virus (HIV) pseudotyped with SARS-CoV-2 S protein in cultured cells [41–44]. It is also tempting to suggest a link between SARS-CoV-2 pathogenesis and the presence of SΔ20, since severe cases of the disease were recently linked to considerable lung damage and the occurrence of syncytia [45,46].

Our findings indicate the presence of consistent intra-sample genetic variants of SARS-CoV-2, including a recurrent sub-population of SΔ20 variants with elevated fusogenic properties. Further investigation is required to better define the extent of SARS-CoV-2 variability in infected hosts and to assess the role of these subspecies in the life cycle of this virus. More importantly, further studies on the presence of SΔ20 and its link with viral pathogenicity could lead to better diagnostic strategies and designer treatments for COVID-19.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Analysis of intra-variability within SARS-CoV-2 samples.

15,289 publicly available high-throughput sequencing datasets were downloaded from the NCBI Sequence Read Archive (up to July 10, 2020). They comprise of 15,224 and 65 datasets from infected individuals and infected cell lines, respectively. The datasets from infected cells were generated by Blanco-Melo et al. [29]. Duplicated reads were combined to reduce amplification bias and mapped to the SARS-CoV-2 isolate Wuhan-Hu-1 reference genome (NC_045512v2) using hisat2 (v.2.1.0)[47]. For each dataset, the consensus sequences and the frequency of nucleotides at each position were extracted from files generated by bcftools (v.1.10.2) of the samtools package (v.1.1) with an in-house Perl script [48,49]. All further calculations were performed in R. To reduce the number of variations due to sequencing errors and/or protocol differences, only positions mapped with a sequencing depth of 50 reads and having at least 5 reads with variations compared to the sample consensus were considered. Sequence logos were generated with the ggseqlogo package (v.0.1) [50].

Differential expression analysis of transcript coding for APOBECs and ADARs.

High-throughput sequencing datasets generated in a recent transcription profiling study of several cell lines infected with SARS-CoV-2 were downloaded from SRA [29]. Duplicated reads were combined and mapped to the human reference genome (Homo_sapiens.GRCh38.83) using hisat2 (v.2.1.0)[47]. Transcript abundance was performed using HTSeq 0.12.4 [51] and normalized into Transcripts Per Million (TPM) in R.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

309 **Cell culture and plasmids.**

310 Human embryonic kidney 293T (HEK293T) were obtained from the American Type
311 Culture Collection (ATCC CRL-11268) and maintained in Dulbecco's Modified Eagle's Medium
312 (DMEM) supplemented with 5% fetal bovine serum (Fisher Scientific), 5% bovine calf serum
313 (Fisher Scientific) and 1x antibiotic-antimycotic (Fisher Scientific). HEK293T stably expressing
314 human ACE2 (kind gifts of Dr. Hyeryun Choe, The Scripps Research Institute) were cultured and
315 maintained in DMEM (Wisent) supplemented with 10% fetal bovine serum (Sigma), 1 U/mL
316 penicillin, 1 µg/mL streptomycin, and 3 µg/mL glutamine (Corning). All cells were cultured at 37
317 °C in a humidified atmosphere containing 5% CO2. The original bi-molecular fluorescence
318 complementation (BiFc) constructs GCN4 leucine zipper-Venus1 (ZipV1) and GCN4 leucine
319 zipper-Venus2 (ZipV2) were sourced from Stephen W. Michnick (reviewed in [52]). pCEP4(myc-
320 ACE2) was a gift from Erik Procko (Addgene plasmid #141185). pCAGGS expressing the SARS-
321 CoV-2 S protein (pCAGGS(S)) was provided by Dr. Florian Krammer (Mount Sinai).
322 pCAGGS(SΔ20) was generated using overlapping PCR to introduce a termination codon at residue
323 1254.

324

325 **Syncytium Formation Assay.**

326 HEK293T expressing ACE2 cells were seeded in 24-well plates in complete media to
327 obtain an 85% confluence the following day. Cells were then placed in serum-free DMEM and
328 transiently co-transfected using JetPRIME (Polyplus Transfection, France) with plasmids
329 encoding GFP (MLV-GFP, a kind gift of Dr. James Cunningham, Brigham and Women's
330 Hospital), pCAGGS(S) or pCAGGS(SΔ20), and empty pCAGGS at a 0.15:0.2:0.65 ratio. 18 hours

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

post transfection, cells were imaged (ZOE Fluorescent Cell Imager, Bio-Rad) for syncytia formation using the Green channel to visualize fusion of GFP positive cells as performed previously [53]. GFP area was quantified on ImageJ [54].

Quantification of cell-to-cell fusion using bi-molecular fluorescence complementation (BiFc).

HEK293T cells were seeded in a 12-well microplate (500,000 cells/well) in complete media for 24h. Transient transfections were performed using JetPRIME (Polyplus transfection, France) according to the manufacturer's instructions. Target cells were transfected with ZipV1 (0.5µg) or pCEP4(myc-ACE2) (0.05µg) and ZipV1 (0.5µg). Effector cells population were transfected with ZipV2 (0.5µg) and pCAGGS(S) (0.125µg) or pCAGGS(SΔ20) (0.125µg). Total DNA was normalized using the empty pCAGGS vector DNA to 1µg. Following transfection, cells were incubated at 37 °C for 24 h. Then, cells were rinsed with PBS and detached with versene (PBS, 0.53mM EDTA) and counted. 35,000 cells/well of both populations were co-seeded in DMEM without serum in a 384-well black plate with optical clear bottom and incubated for 2 hours at 37 °C, 5% CO_2. BiFC signal was acquired using Biotek Synergy Neo2 plate reader (BioTek) using monochromator set to excitation/emission of 500 and 542 nm.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgments

K.F. is supported by an Ontario Graduate Scholarship. M.-A.L. holds a Canada Research Chair in Molecular Virology and Intrinsic Immunity. M.C. is a Canada Research Chair in Molecular Virology and Antiviral Therapeutics. This work was supported by a COVID-19 Rapid Research grant from the Canadian Institutes for Health Research (CIHR; OV1 170355) to M.-A.L and M.P., and a COVID-19 Rapid Research Grant (OV3 170632) to M.C. and P.M.G.

References

1. Hu B, Guo H, Zhou P, Shi Z-L. Characteristics of SARS-CoV-2 and COVID-19. Nature reviews Microbiology. 2020. doi:10.1038/s41579-020-00459-7

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020;382: 727–733. doi:10.1056/NEJMoa2001017

3. Fehr AR, Perlman S. Coronaviruses: An Overview of Their Replication and Pathogenesis. 2015. pp. 1–23. doi:10.1007/978-1-4939-2438-7_1

4. Hartenian E, Nandakumar D, Lari A, Ly M, Tucker JM, Glaunsinger BA. The molecular virology of coronaviruses. Journal of Biological Chemistry. 2020;295: 12910–12934. doi:10.1074/jbc.REV120.013930

5. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. Cells. 2020;9: 1267. doi:10.3390/cells9051267

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

6. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181: 271-280.e8. doi:10.1016/j.cell.2020.02.052

7. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181: 271-280.e8. doi:10.1016/j.cell.2020.02.052

8. Hoffmann M, Kleine-Weber H, Pöhlmann S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Molecular Cell. 2020;78: 779-784.e5. doi:10.1016/j.molcel.2020.04.022

9. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nature Reviews Microbiology. 2020. doi:10.1038/s41579-020-00468-6

10. McBride CE, Li J, Machamer CE. The Cytoplasmic Tail of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein Contains a Novel Endoplasmic Reticulum Retrieval Signal That Binds COPI and Promotes Interaction with Membrane Protein. Journal of Virology. 2007;81: 2418–2428. doi:10.1128/JVI.02146-06

11. Lontok E, Corse E, Machamer CE. Intracellular targeting signals contribute to localization of coronavirus spike proteins near the virus assembly site. Journal of virology. 2004;78: 5913–22. doi:10.1128/JVI.78.11.5913-5922.2004

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

12. Qian Z, Dominguez SR, Holmes K v. Role of the Spike Glycoprotein of Human Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in Virus Entry and Syncytia Formation. Ren X, editor. PLoS ONE. 2013;8: e76469. doi:10.1371/journal.pone.0076469

13. Matsuyama S, Nagata N, Shirato K, Kawase M, Takeda M, Taguchi F. Efficient Activation of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein by the Transmembrane Protease TMPRSS2. Journal of Virology. 2010;84: 12658–12664. doi:10.1128/JVI.01542-10

14. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. New England Journal of Medicine. 2020;382: 2302–2315. doi:10.1056/NEJMoa2006100

15. Kim J-S, Jang J-H, Kim J-M, Chung Y-S, Yoo C-K, Han M-G. Genome-Wide Identification and Characterization of Point Mutations in the SARS-CoV-2 Genome. Osong Public Health and Research Perspectives. 2020;11: 101–111. doi:10.24171/j.phrp.2020.11.3.05

16. Phan T. Genetic diversity and evolution of SARS-CoV-2. Infection, Genetics and Evolution. 2020;81: 104260. doi:10.1016/j.meegid.2020.104260

17. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. 2020;83: 104351. doi:10.1016/j.meegid.2020.104351

18. Vankadari N. Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution. Gene. 2020;752: 144792. doi:10.1016/j.gene.2020.144792

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

19. Mavian C, Marini S, Prosperi M, Salemi M. A Snapshot of SARS-CoV-2 Genome Availability up to April 2020 and its Implications: Data Analysis. JMIR Public Health and Surveillance. 2020;6: e19170. doi:10.2196/19170

20. Farkas C, Fuentes-Villalobos F, Garrido JL, Haigh J, Barría MI. Insights on early mutational events in SARS-CoV-2 virus reveal founder effects across geographical regions. PeerJ. 2020;8: e9255. doi:10.7717/peerj.9255

21. Rosenberg BR, Hamilton CE, Mwangi MM, Dewell S, Papavasiliou FN. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3′ UTRs. Nature Structural & Molecular Biology. 2011;18: 230–236. doi:10.1038/nsmb.1975

22. Lerner T, Papavasiliou F, Pecori R. RNA Editors, Cofactors, and mRNA Targets: An Overview of the C-to-U RNA Editing Machinery and Its Implication in Human Disease. Genes. 2018;10: 13. doi:10.3390/genes10010013

23. di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Science Advances. 2020;6: eabb5813. doi:10.1126/sciadv.abb5813

24. Simmonds P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere. 2020;5. doi:10.1128/mSphere.00408-20

25. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral Mutation Rates. Journal of Virology. 2010;84: 9733–9748. doi:10.1128/JVI.00694-10

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26. Drake JW, Holland JJ. Mutation rates among RNA viruses. Proceedings of the National Academy of Sciences of the United States of America. 1999;96: 13910–3. doi:10.1073/pnas.96.24.13910

27. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses. RNA Biology. 2011;8: 270–279. doi:10.4161/rna.8.2.15013

28. Domingo E, Perales C. Viral quasispecies. PLOS Genetics. 2019;15: e1008271. doi:10.1371/journal.pgen.1008271

29. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, Uhl S, Hoagland D, Møller R, et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. Cell. 2020;181: 1036-1045.e9. doi:10.1016/j.cell.2020.04.026

30. Kreutzer DA, Essigmann JM. Oxidized, deaminated cytosines are a source of C --> T transitions in vivo. Proceedings of the National Academy of Sciences of the United States of America. 1998;95: 3578–82. doi:10.1073/pnas.95.7.3578

31. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Research. 2013;41: e67–e67. doi:10.1093/nar/gks1443

32. Chen L, Liu P, Evans TC, Ettwiller LM. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science. 2017;355: 752–756. doi:10.1126/science.aai8690

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

454 33. Belhadj Slimen I, Najar T, Ghram A, Dabbebi H, ben Mrad M, Abdrabbah M. Reactive
455 oxygen species, heat stress and oxidative-induced mitochondrial damage. A review.
456 International Journal of Hyperthermia. 2014;30: 513–523.
457 doi:10.3109/02656736.2014.971446

458 34. Ahn EH, Lee SH. Detection of Low-Frequency Mutations and Identification of Heat-
459 Induced Artifactual Mutations Using Duplex Sequencing. International Journal of
460 Molecular Sciences. 2019;20: 199. doi:10.3390/ijms20010199

461 35. Arbeithuber B, Makova KD, Tiemann-Boege I. Artifactual mutations resulting from
462 DNA lesions limit detection levels in ultrasensitive sequencing applications. DNA
463 Research. 2016;23: 547–559. doi:10.1093/dnares/dsw038

464 36. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification
465 of rare mutations with massively parallel sequencing. Proceedings of the National
466 Academy of Sciences. 2011;108: 9530–9535. doi:10.1073/pnas.1105422108

467 37. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 8-Hydroxyguanine, an abundant
468 form of oxidative DNA damage, causes G----T and A----C substitutions. The Journal of
469 biological chemistry. 1992;267: 166–72. Available:
470 http://www.ncbi.nlm.nih.gov/pubmed/1730583

471 38. Mancini F, Barbanti F, Scaturro M, Errico G, Iacobino A, Bella A, et al. Laboratory
472 management for SARS-CoV-2 detection: a user-friendly combination of the heat
473 treatment approach and rt-Real-time PCR testing. Emerging microbes & infections.
474 2020;9: 1393–1396. doi:10.1080/22221751.2020.1775500

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

475 39. Bruskov VI, Malakhova L v, Masalimov ZK, Chernikov A v. Heat-induced formation of
476 reactive oxygen species and 8-oxoguanine, a biomarker of damage to DNA. Nucleic acids
477 research. 2002;30: 1354–63. doi:10.1093/nar/30.6.1354

478 40. Lewis CA, Crayle J, Zhou S, Swanstrom R, Wolfenden R. Cytosine deamination and the
479 precipitous decline of spontaneous mutation during Earth's history. Proceedings of the
480 National Academy of Sciences of the United States of America. 2016;113: 8194–9.
481 doi:10.1073/pnas.1607580113

482 41. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of
483 SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nature
484 Communications. 2020;11: 1620. doi:10.1038/s41467-020-15562-9

485 42. Dieterle ME, Haslwanter D, Bortz RH, Wirchnianski AS, Lasso G, Vergnolle O, et al. A
486 Replication-Competent Vesicular Stomatitis Virus for Studies of SARS-CoV-2 Spike-
487 Mediated Cell Entry and Its Inhibition. Cell Host & Microbe. 2020;28: 486-496.e6.
488 doi:10.1016/j.chom.2020.06.020

489 43. Schmidt F, Weisblum Y, Muecksch F, Hoffmann H-H, Michailidis E, Lorenzi JCC, et al.
490 Measuring SARS-CoV-2 neutralizing antibody activity using pseudotyped and chimeric
491 viruses. Journal of Experimental Medicine. 2020;217. doi:10.1084/jem.20201181

492 44. Case JB, Rothlauf PW, Chen RE, Liu Z, Zhao H, Kim AS, et al. Neutralizing Antibody
493 and Soluble ACE2 Inhibition of a Replication-Competent VSV-SARS-CoV-2 and a
494 Clinical Isolate of SARS-CoV-2. Cell host & microbe. 2020;28: 475-485.e5.
495 doi:10.1016/j.chom.2020.06.021

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

45. Buchrieser J, Dufloo J, Hubert M, Monel B, Planas D, Michael Rajah M, et al. Syncytia formation by SARS-CoV-2 infected cells. The EMBO journal. 2020; e106267. doi:10.15252/embj.2020106267

46. Bussani R, Schneider E, Zentilin L, Collesi C, Ali H, Braga L, et al. Persistence of viral RNA, pneumocyte syncytia and thrombosis are hallmarks of advanced COVID-19 pathology. EBioMedicine. 2020;61: 103104. doi:10.1016/j.ebiom.2020.103104

47. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 2019;37: 907–915. doi:10.1038/s41587-019-0201-4

48. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–2993. doi:10.1093/bioinformatics/btr509

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25: 2078–9. doi:10.1093/bioinformatics/btp352

50. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics (Oxford, England). 2017;33: 3645–3647. doi:10.1093/bioinformatics/btx469

51. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31: 166–169. doi:10.1093/bioinformatics/btu638

52. Michnick SW, Ear PH, Landry C, Malleshaiah MK, Messier V. A toolkit of protein-fragment complementation assays for studying and dissecting large-scale and dynamic

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

protein-protein interactions in living cells. Methods in enzymology. 2010;470: 335–68. doi:10.1016/S0076-6879(10)70014-8

53. Côté M, Zheng Y-M, Liu S-L. Receptor Binding and Low pH Coactivate Oncogenic Retrovirus Envelope-Mediated Fusion. Journal of Virology. 2009;83: 11447–11455. doi:10.1128/JVI.00748-09

54. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nature Methods. 2012;9: 671–675. doi:10.1038/nmeth.2089

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Legends

Fig 1: Intra-sample variability of the SARS-CoV-2 genome in infected individuals. (**A**) Number of intra-variations observed for each sample analyzed. The red dots represent the 11,362 samples analyzed and the blue violon shows the distribution of the data. (**B**) Type of variation and sequence context for each intra-sample variable position. Bars represent the percentage of each type. Sequence context is represented by logos comprised of the consensus nucleotides (center) with two nucleotides upstream and two downstream from each intra-sample variable position. (**C**) Recurrent intra-genetic variations are represented as percentages of samples containing variations at each position. The SARS-CoV-2 genome and its genes are represented by yellow boxes below the graph. The blue line indicates 5% shared variations and was used to extract the recurrent intra-sample variations listed in Table 1. The inset represents a magnification of the cluster identified at the end of the S gene. (**D**) One-dimension representation of the data shown in panel C for each type of variation individually. The location of the C->A variation at position 25,324 is indicated by a red line in (C) and (D).

Fig 2: Localization of the C->A missense mutation on the SARS-CoV-2 S protein. (**A**) Schematic representation of the functional domain of the SARS-CoV-2 S protein. (**B**) Localization of the C->A variation on the carboxy terminal domain (CTD) of the S protein. The mutation is colored and boxed in red. The carboxy terminal domain (CTD) and the ERRS are colored in yellow and black, respectfully in (A) and (B). (**C**) Distribution of the intra-sample proportion of the C->A transversion at position 25,324 in the 6,668 samples containing this subspecies. The inset

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

548 represents the distribution using red dots to represent the samples having this intra-genetic
549 variation and a blue violon to show the distribution of the data.

550

551 **Fig 3: Heatmap representation for the expression of genes coding for APOBEC and ADAR**
552 **family members.** Counts are represented as Transcripts Per Million (TPM). The blue scale also
553 correlates with TPM values.

554

555 **Fig 4: Intra-sample variability of the SARS-CoV-2 genome in infected cells.** (**A**) Number of
556 intra-variations observed for each sample analyzed. The red dots represent the 11,362 samples
557 analyzed and the blue violon shows the distribution of the data. (**B**) Type of variation and sequence
558 context for each intra-sample variable position. Bars represent the percentage of each type.
559 Sequence context is represented by logos comprised of the consensus nucleotides (center) with
560 two nucleotides upstream and two downstream from each intra-sample variable position. (**C**)
561 Recurrent intra-genetic variations are represented as percentages of samples containing variation
562 at each position. The SARS-CoV-2 genome and its genes are represented by yellow boxes below
563 the graph. The blue line indicates 10% shared variations and was used to extract the intra-sample
564 variation listed in Table 2. The inset represents a magnification of the cluster identified at the end
565 of the S gene. (**D**) One-dimension representation of the data shown in panel C for each type of
566 variation individually. The location of the C->A variation at position 25,324 is indicated by a red
567 line in (C) and (D).

568

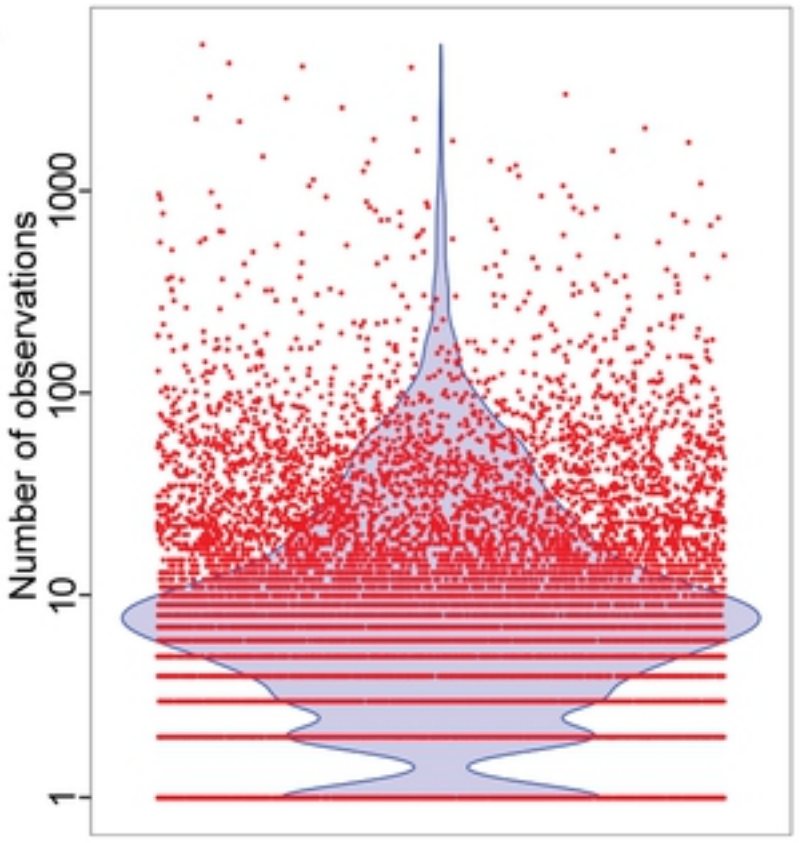
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

569 **Fig 5: Effect of SARS-CoV-2 S protein and SΔ20 on cell-to-cell fusion.** (**A**) Phase-contrast and
570 fluorescence microscopy of HEK293T/ACE2 cells expressing GFP with empty vector (pCAGGS),
571 or plasmid expressing wild-type S (pCAGGS(S WT)) or SΔ20 (pCAGGS(SΔ20)). (**B**)
572 Quantification of syncytia expressing GFP. The area of 129 and 125 syncytia from wild-type S
573 and SΔ20, respectfully, are represented by black dots. The p-value (pv) was calculated using
574 unpaired t-test. (**C**) Schematic representation of the bi-molecular fluorescence complementation
575 (BiFc) assay used to quantify cell fusion. (**D**) Quantification of cell fusion using the BiFc assay.
576 Bars represent averages ± standard deviations of five independent experiments and the p-value
577 (pv) was calculated using paired t-test. Schematic of BiFc created with BioRender.com.

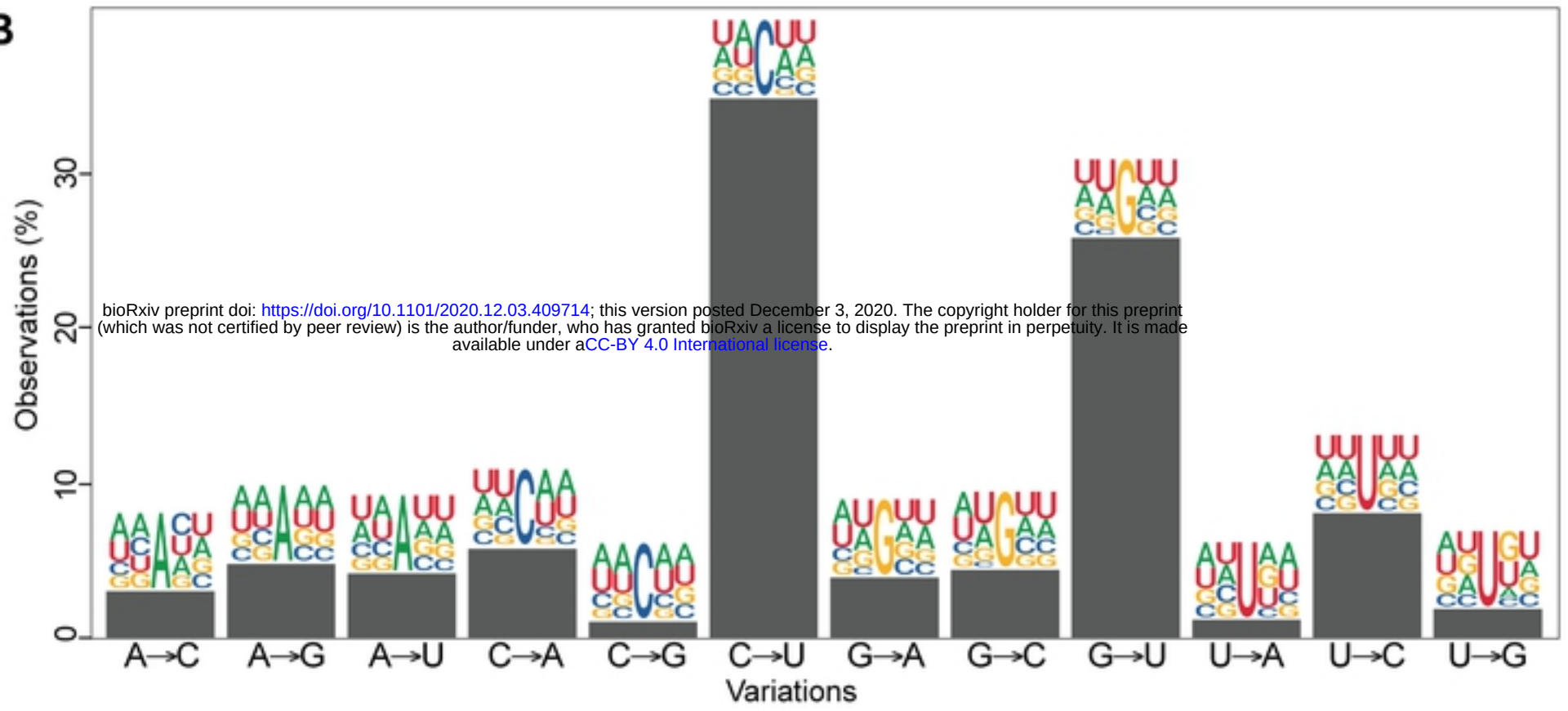
578

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

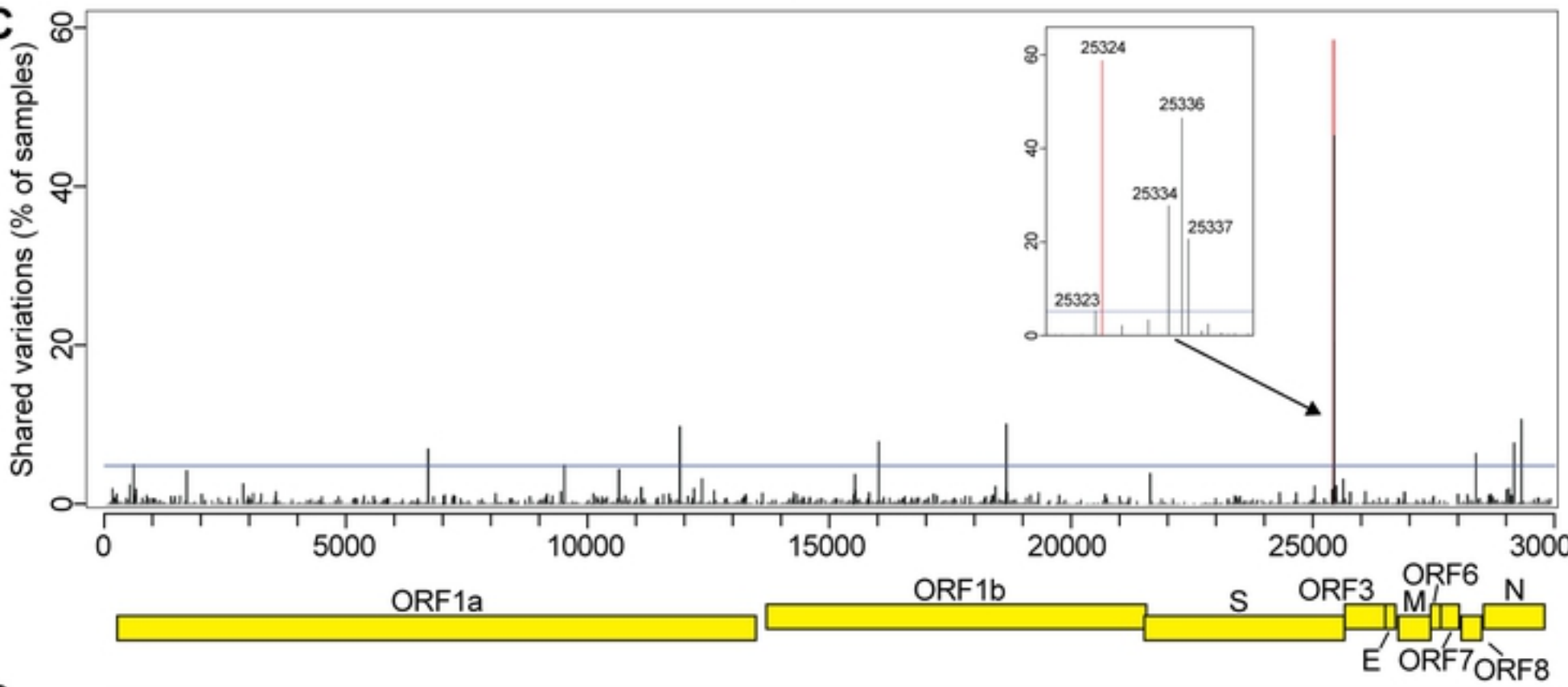
A



B



C



D

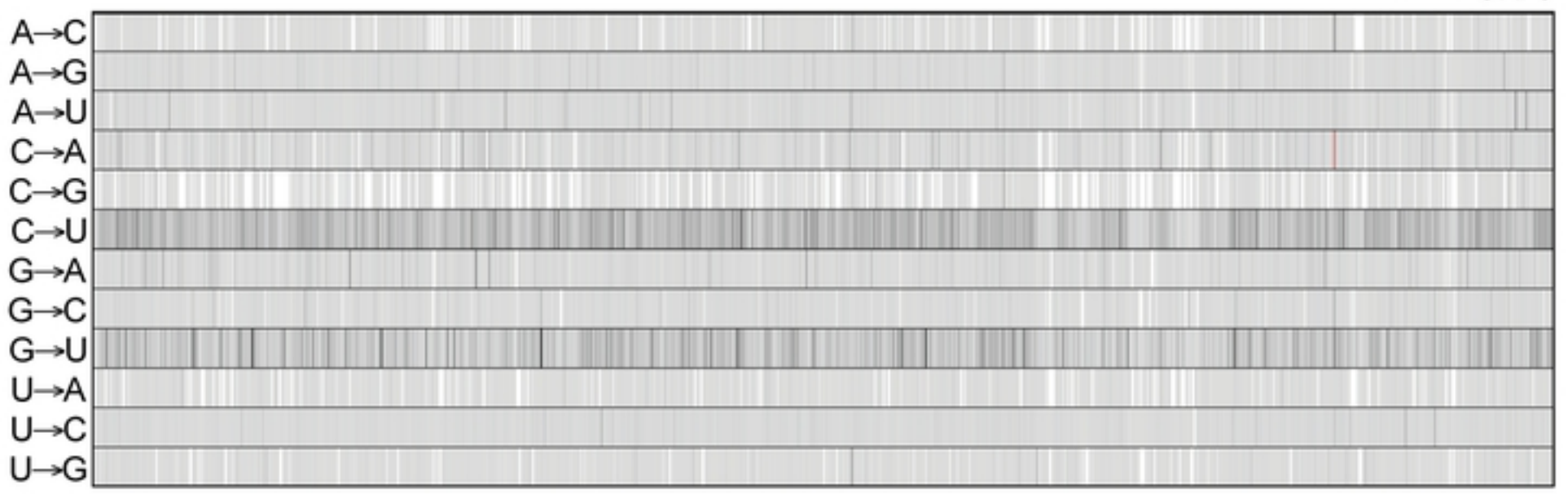


Fig 1

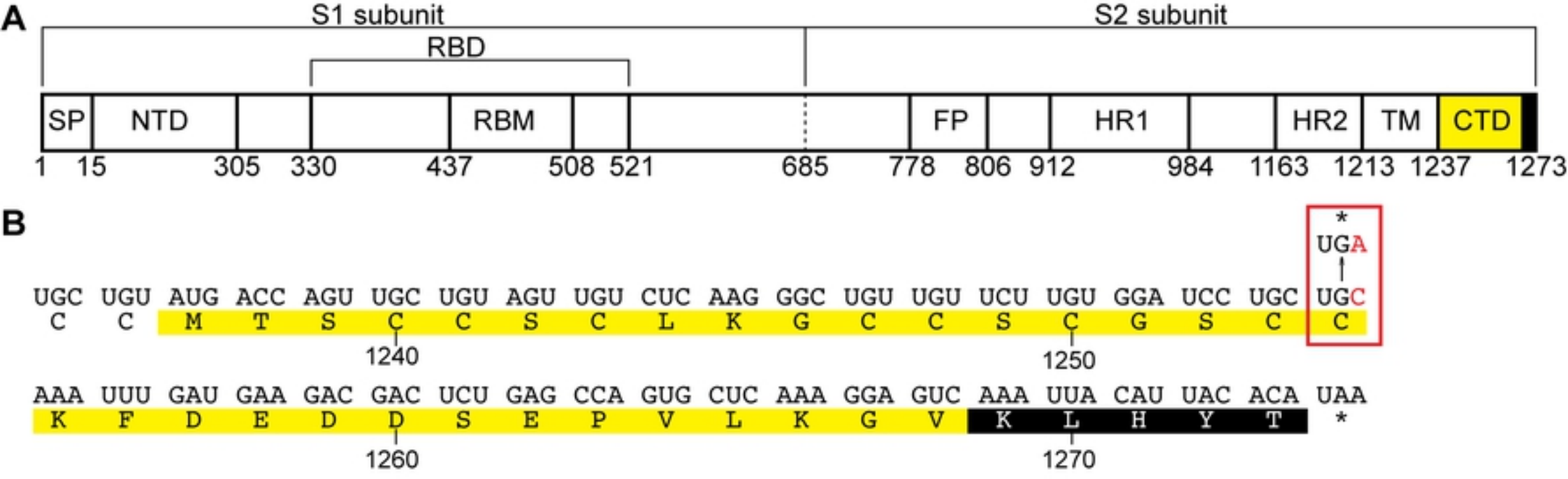
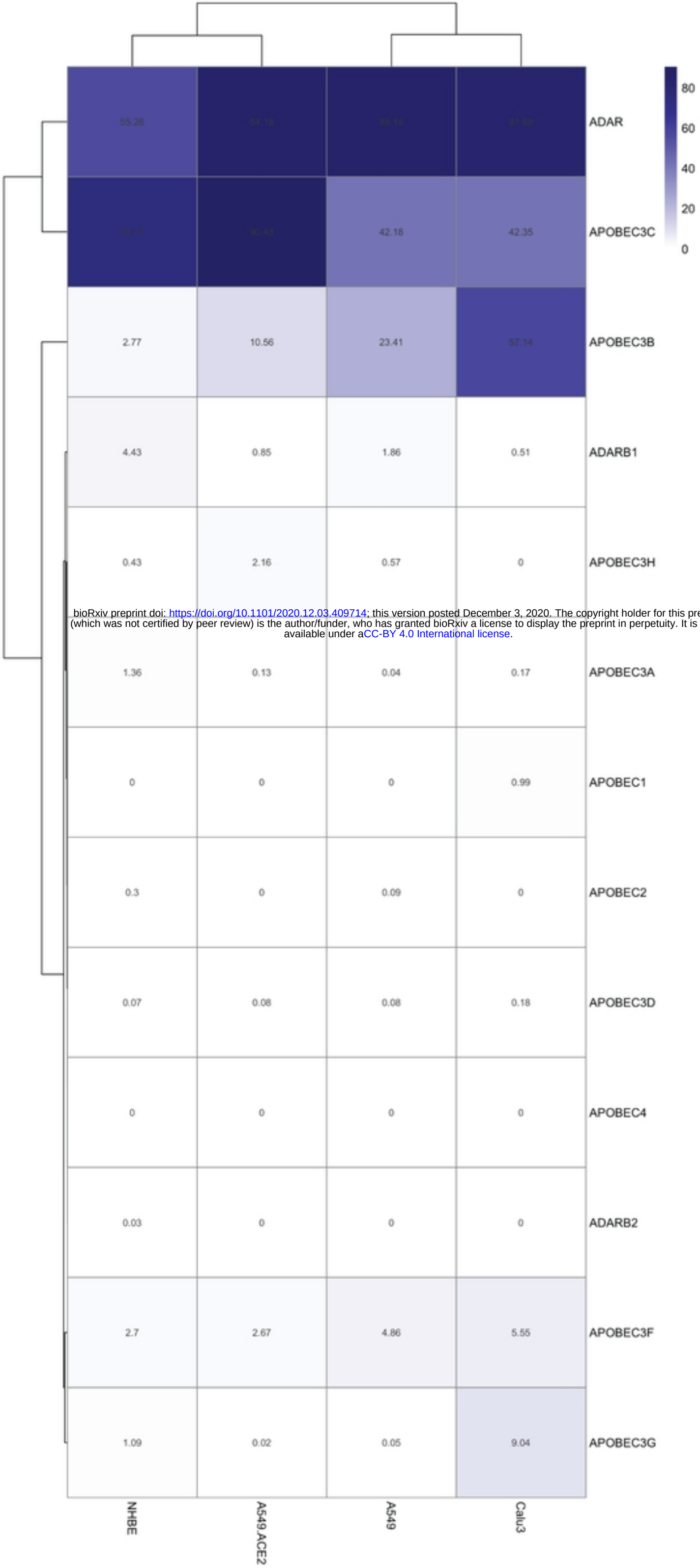

A
S1 subunit
S2 subunit
RBD
SP
NTD
RBM
FP
HR1
HR2
TM
CTD
1
15
305
330
437
508
521
685
778
806
912
984
1163
1213
1237
1273
B
UGC UGU AUG ACC AGU UGC UGU AGU UGU CUC AAG GGC UGU UGU UCU UGU GGA UCC UGC UGC
C C M T S C C S C L K G C C S C G S C C
1240
1250
*
UGA
AAA UUU GAU GAA GAC GAC UCU GAG CCA GUG CUC AAA GGA GUC AAA UUA CAU UAC ACA UAA
K F D E D D S E P V L K G V K L H Y T *
1260
1270
C
Number of samples
0
200
400
10
20
30
40
C→A variation at nt 25,324 (% / sample)
C→A (% / sample)
3
10
30

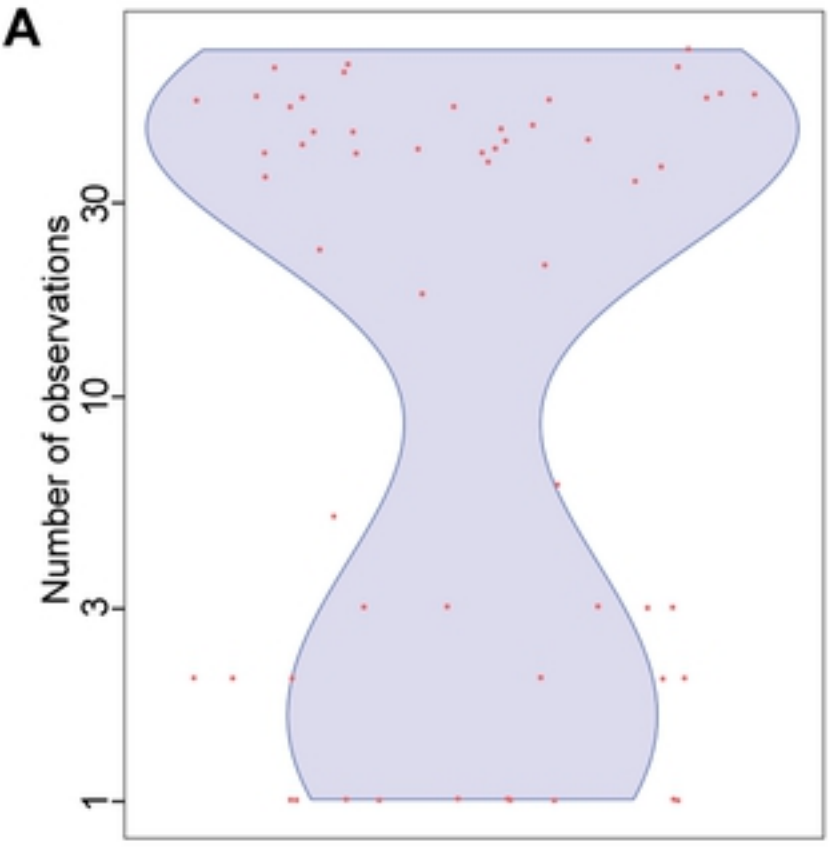
Fig 2



bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this pre
(which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is
available under aCC-BY 4.0 International license.

Fig 3

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



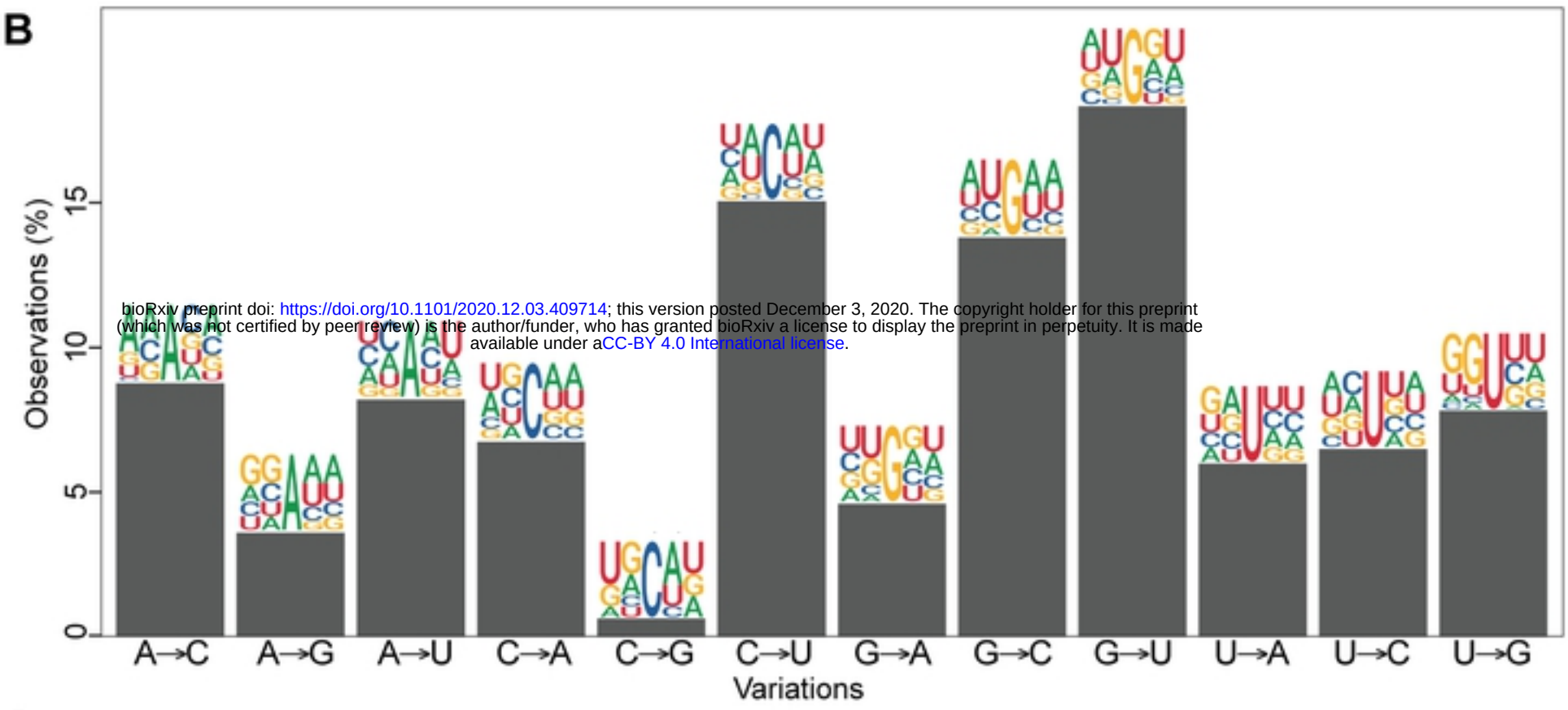
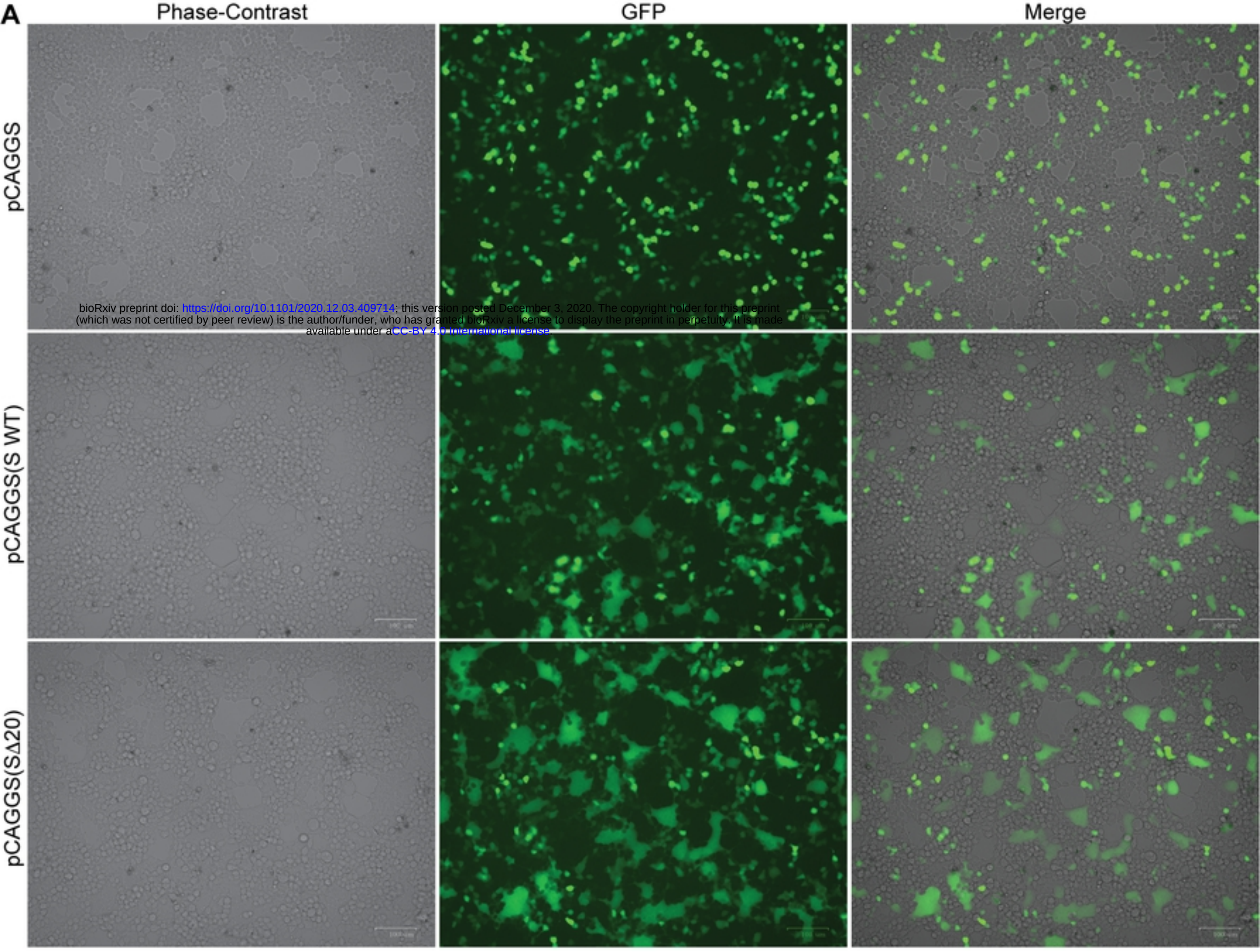
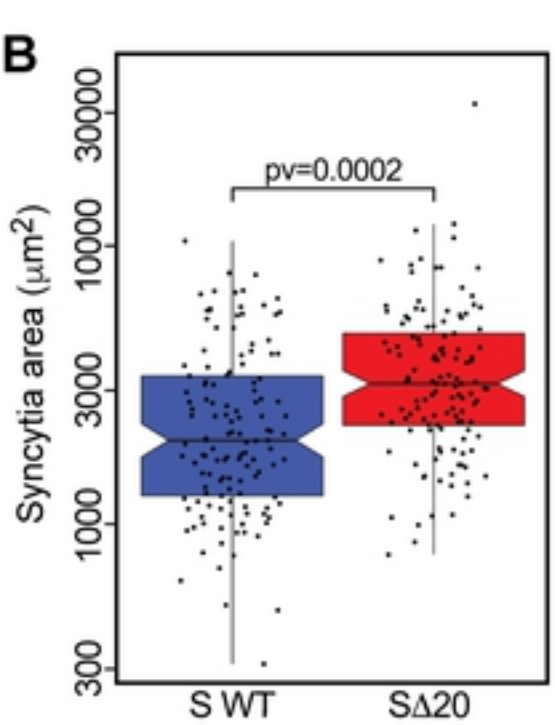
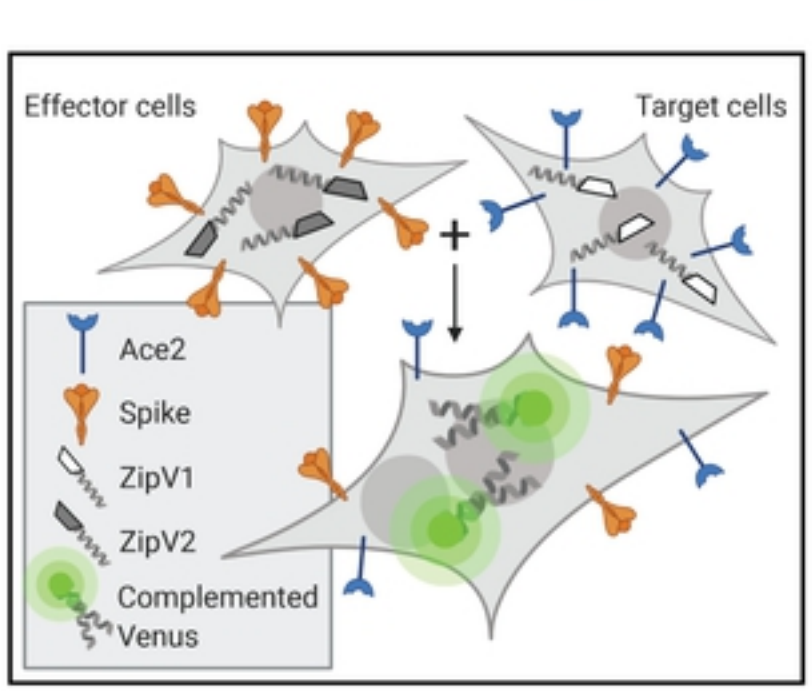


Fig 4

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.03.409714; this version posted December 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.







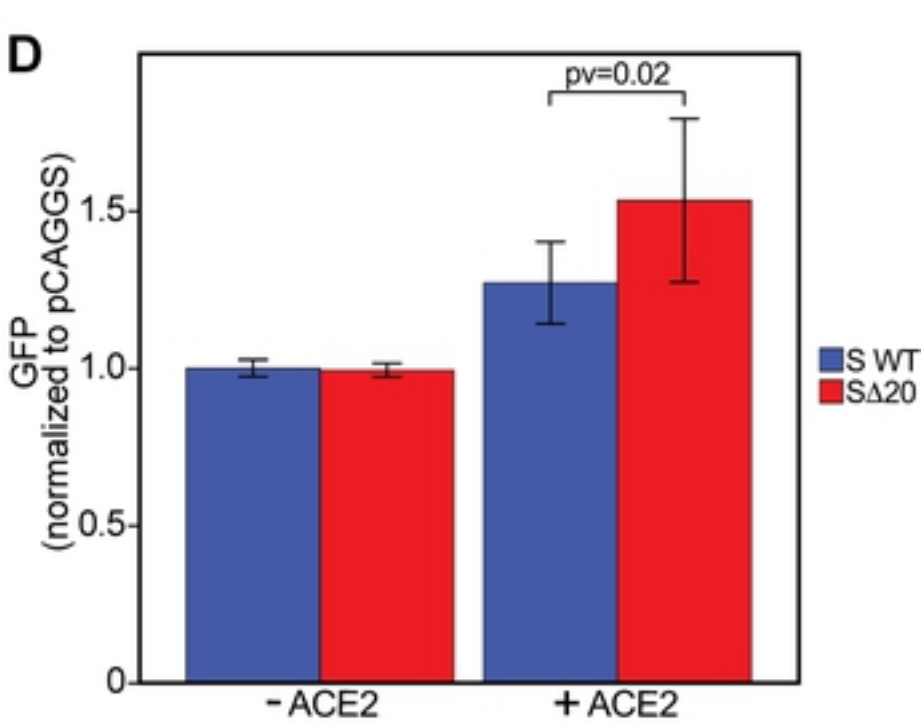


Fig 5