Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic

Leszek J. Klimczak[1], Thomas A. Randall[1], Natalie Saini[2], Jian-Liang Li[1], and Dmitry A. Gordenin[2] *

1. Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, NIH, Durham, NC 27709

2. Mechanisms of Genome Dynamics Group, National Institute of Environmental Health Sciences, NIH, Durham, NC 27709

* Corresponding author

E-mail: gordenin@niehs.nih.gov

Abstract

Genomes of tens of thousands of SARS-CoV2 isolates have been sequenced across the world and the total number of changes (predominantly single base substitutions) in these isolates exceeds ten thousand. We compared the mutational spectrum in the new SARS-CoV-2 mutation dataset with the previously published mutation spectrum in hypermutated genomes of rubella - another positive single stranded (ss) RNA virus. Each of the rubella isolates arose by accumulation of hundreds of mutations during propagation in a single subject, while SARS-CoV-2 mutation spectrum represents a collection events in multiple virus isolates from individuals across the world. We found a clear similarity between the spectra of single base substitutions in rubella and in SARS-CoV-2, with C to U as well as A to G and U to C being the most prominent in plus strand genomic RNA of each virus. Of those, U to C changes universally showed preference for loops versus stems in predicted RNA secondary structure. Similarly, to what was previously reported for rubella, C to U changes showed enrichment in the uCn motif, which suggested a subclass of APOBEC cytidine deaminase being a source of these substitutions. We also found enrichment of several other trinucleotide-centered mutation motifs only in SARS-CoV-2 - likely indicative of a mutation process characteristic to this virus. Altogether, the results of this analysis suggest that the mutation mechanisms that lead to hypermutation of the rubella vaccine virus in a rare pathological condition may also operate in the background of the SARS-CoV-2 viruses currently propagating in the human population.

Introduction

RNA viruses can show a high mutation rate [1], which often results in fast emergence of viral quasispecies - populations of viruses differing in several genomic positions from the original virus [2]. Errors made by the RNA-dependent RNA-polymerase (RdRP) viral replicase are a source of mutations, however in coronaviruses some of these errors can be corrected by a proofreading RNA exonuclease ExoN [3, 4]. Another source of mutations comes from RNA base editing by two classes of endogenous enzymes: adenine deaminases (ADAR) and cytosine deaminases APOBECs (APOlipoprotein B mRNA Editing Complex like polypeptides) which have a broad range of functions spanning from site-specific editing of cellular mRNAs to inhibiting viral and retrotransposon proliferation [5-7].

ADARs (ADAR1 and ADAR2) are double-strand (ds) RNA-specific enzymes converting adenine to inosine (A to I). Since inosine pairs with cytosine, this will result in A to G changes after the next round of replication. The preference of ADARs for certain deamination motifs - reflecting a combination of immediate nucleotide context and the anticipated dsRNA formed by folding - was assessed for *in vitro* editing of several RNA substrates. Based on these data, software was developed aimed to assign predictive ADAR deamination scores to any A position in a given RNA molecule [8]. The ADAR editing sites that were deduced in RNAs of cultured stimulated immune cells [9] agreed with the preferences defined in the in vitro study. It remains to be established whether these preferences would hold for a wide variety of RNA substrates in conditions of controlled in vivo expression of either ADAR1 or ADAR2.

Unlike ADARs, the structure of APOBEC enzymes allow deamination only in single-strand (ss) RNA or in ssDNA. At least two APOBECs, APOBEC1 and APOBEC3A are capable of deaminating cytosine to uracil in RNA, however an RNA editing capacity of other APOBECs cannot be excluded [10-12]. Cytosine deamination in RNA creates the normal RNA base –

uracil, which can be then accurately copied in subsequent rounds of RNA replication. DNA deamination motifs or mutation signatures (i.e., the immediate nucleotide contexts around deaminated Cs) of several human APOBECs were first defined in model microbial systems and then found in genomic DNAs of human cancers, where they can cause hypermutation clusters [13-18]. The preferred DNA deamination motif of APOBEC3G (A3G) is nCc to nTc (n=any base; the mutated nucleotide and the resulting nucleotide are capitalized). Other APOBECs show preference for the tCn to tTn deamination motif or to a more stringently defined trinucleotide. For example, both APOBEC3A (A3A) and APOBEC3B (A3B) prefer tCa as a target, both in the yeast model and in human cancers [14].

An indication of frequent RNA editing was recently found in the isolates of the hypermutated plus-strand ssRNA rubella vaccine virus from cutaneous granulomas of children with primary immunodeficiencies [19]. Altogether, genomes of six independent isolates of the hypermutated vaccine-derived viruses contained 993 mutations. Most changes were C to U in the genomic plus-strand RNA. These C to U changes showed high enrichment of a uCa to uUa RNA editing motif – a match to the characteristic A3A or A3B mutagenic motif in DNA. While the similarity between the C to U RNA editing motif in rubella and the DNA editing motifs strongly suggested the nature of the editing enzyme, signature motifs of APOBEC cytosine deamination in RNA are yet to be confirmed in a direct study involving expression of an APOBEC enzyme and collection of in vivo-editing spectrum data. The second most prevalent type of editing event was A to G change in the rubella plus or minus-strand, revealed as either U to C or G to A changes in the reported plus-strand sequence, respectively. These changes would be expected to result from ADAR editing. Minus-strand RNA in rubella as well as in Coronaviridae would often exist within completely- or partially double-stranded RNA [20, 21], which would be the right substrate for an ADAR. This strand is a template for the multiple rounds of transcription generating many plus-strand partial- or full-size genomic RNAs. Thus, an A to G editing event in the minus-strand of

dsRNA at the beginning of replication cycle would be carried as a U to C change to multiple rubella genomes. A to G editing events in the plus-strand of the dsRNA intermediate may directly contribute to the mutation spectrum in plus-strand viral genomes or propagate the mutation via the subsequent rounds of replication within the same cell.

In summary, the previous analysis of the mutation spectrum and mutational signatures of hypermutated rubella genomes provided a strong indication of hyperediting by APOBEC cytidine deaminases as well as suggested editing by ADAR adenine deaminases [19]. Both rubella and Coroniviridae are positive ssRNA viruses which produce many copies of the genomic positive RNA strand and also have dsRNA intermediates in their replication cycles [20-22], which can serve as substrates for APOBECs and for ADARs, respectively. Indeed, recent analyses suggested APOBEC and ADAR editing in SARS-CoV-2 based on an excess of C to U changes and A to G in sequencing reads from lavages of two COVID-19 patients or in genome alignments [23]. Based on the similarity between the preferred RNA editing motifs in rubella and the APOBEC DNA hypermutation motifs, we sought to determine whether similar mutational signature motifs can be detected in a collection of 32,341 whole genome sequences of multiple SARS-CoV-2 isolates that have been sequenced during the current COVID-19 pandemic. We present here the evidence indicating a similarity between the RNA editing spectra and mutational signatures between the hypermutated rubella isolates and the load of editing changes accumulated in this collection of SARS-CoV-2 genomes. We also found several new trinucleotide-centered mutational motifs unique to SARS-CoV-2.

Materials and Methods

SARS-CoV-2 genomes

SARS-CoV-2 genome sequences in the FASTA format were downloaded from https://www.epicov.org/epi3/frontend# at 13:10 EST on 2020/06/24 after applying the following download filters: (i) “complete sequence”; (ii) “high coverage”; (iii) “human”; (iv) “hCOV-19/...”. The downloaded 32,341 FASTA entries were edited to remove spaces from FASTA headers (fatal defects for many tools) and reformatted to a consistent line length of 80 characters. Several samples with non-standard FASTA problems (many of them contain hyphens) that cannot be reasonably fixed and failed at the stage of alignment with the reference, therefore only 32,115 isolates were included into mutation calling.

Mutation calls in SARS-CoV-2 genomes

Mutations in individual isolates were identified using MUMmer 3.23 ([24] and http://mummer.sourceforge.net/) by making pairwise alignments with the original Wuhan isolate (GenBank entry NC_045512.2) using the command:

```
nucmer NC_045512.2.fasta query.fasta
```

The SNP variants output was generated using the command:

```
show-snps -T -Clr out.delta
```

and concatenating the individual results into a single tab-delimited text file.

For compatibility with other mutation analysis tools, the variant tables were created using the Mutation Annotation Format (MAF): https://software.broadinstitute.org/software/igv/MutationAnnotationFormat but any suitable mutation representation format can be used instead. Functional annotation of the mutations was performed using the standard protocol of ANNOVAR ([25] and https://doc-openbio.readthedocs.io/projects/annovar/en/latest/) based on the genome annotations in GenBank entry NC_045512.2

Out of 251,481 mutations initially called in 32,115 isolates, 251,273 were retained after removing redundant DNA symbols (anything but A,C,G,T) as well as mutation calls separated by less than 20 nt from either end of the reference, of which 243,454 were SNVs in 32070 isolates. Those mutations, redundantly spread in multiple isolates, were collapsed into a non-duplicated MAF designated as NoDups (up to three substitution types at each individual base position in the genome) of 13,736 mutations, 12,156 of which were SNVs. The NoDups filtered MAF was further subdivided into two MAFs: (i) NoDupsNonFunc MAF containing only 4,740 base substitutions that either caused a synonymous change in protein or were located in non-coding regions and therefore were annotated as non-functional; (ii) NoDupsFunc MAF containing only 7,416 base substitutions causing either aminoacid change or protein-truncation and therefore annotated as functional.

Rubella genome and mutation data

The set of 993 base substitutions identified in six hypermutated isolates of rubella RA27/3 vaccine strain listed in MAF format were obtained from a previous study [19]. RA27/3 strain reference sequence GenBank entry FJ211588 was used for RNA-fold and nucleotide context annotations. Rubella mutation calls were compared with de-duplicated sets of SARS-CoV-2 mutation calls from 32,115 isolates contained in three versions of filtered MAFs (see “Design of the analysis” in Results).

Comparison of SARS-CoV-2 and rubella base substitution spectra

The first indication of certain mutagenic mechanisms prevailing in generation of mutation load is a non-uniform distribution of base substitutions. Base substitution counts in each virus depend on both the relative probability of a given base substitution within the group of three possible substitutions of a given base and on the prevalence of each of four bases in a viral genome. Thus, in order to correct for the latter, we calculated densities of each of twelve possible base

149 substitutions in each SARS-CoV-2 and rubella MAFs, dividing a base substitution count by the
150 number of the mutated base in the reference sequence. We then assessed similarity of base
151 substitution densities distributions between rubella and each of SARS-CoV-2 filtered MAF using
152 non-parametric Spearman correlation with the null hypothesis that, there is no positive
153 correlation between spectra in rubella and SARS-CoV-2.
154
155 **Statistical evaluation of mutagenesis in trinucleotide-centered mutation motifs**
156 Calculating enrichment and statistical evaluation of mutagenesis in a small number of
157 trinucleotide-centered mutation motifs identified from mechanistic knowledge turned productive
158 in our prior assessments of mutagenesis associated with established mechanisms and known
159 preference to certain trinucleotide motifs [14, 19, 26, 27]. In this study we extended statistical
160 evaluation to all 192 possible trinucleotide centered motifs.
161 Trinucleotide and single-nucleotide frequencies in the genomic background were calculated
162 using two alternative methods:
163 (i) context-based – counts in the 41 nt windows centered around each mutation location;
164 (ii) reference-based – counts in the whole reference genome.
165 In both cases, Jellyfish ([28] and https://www.cbcb.umd.edu/software/jellyfish/) was used to
166 calculate the counts of tri- and mononucleotides (k-mers with k equal 3 or 1, respectively) in the
167 appropriate FASTA sequences (multiple FASTA entries for context, single entry for the
168 reference). Each of the three substitution types in each of the 64 trinucleotides (total of 192)
169 centered around the mutated base were counted with a set of 192 counters based on string-
170 indexed arrays implemented as simple commands in Awk.
171 Counts of single nucleotide mutations, mutated trinucleotide motifs as well as trinucleotide and
172 single-nucleotide frequencies in the genomic background were used to calculate enrichment
173 with mutagenesis in each of 192 motifs over the presence expected for random mutagenesis as
174 follows:

Enrichment (E) of xYz to xMz mutations calculated as

E(xYz to xMz)=((xYz to xMz_counts)/(Y to M_counts))/(xyz_counts/y_counts),

where

Y and M are the original nucleotide and the nucleotide resulting from mutation, respectively,

y is the nucleotide in the context identical to Y in mutation motif

x and z are 5' and 3' flanking nucleotides in a motif, respectively

Statistical evaluation of Enrichment values was performed by two-tailed Fisher's exact test p-value comparing two ratios:

((xYz:M_counts)/(Y:M_counts-xYz:M_counts)) *vs* (xyz_counts/y_counts-xyz_counts)

P-values were then corrected for multiple hypotheses testing by Benjamini-Hochberg FDR including all 192 motifs. Only values passing FDR=0.05 were considered statistically significant.

A minimum estimate of the number of mutations in a sample caused by xYz to xMz specific mutagenesis in excess of what would be expected by random mutagenesis was calculated as follows:

xYz:M_MutLoad_MinEstimate=[xYz:M_counts]*[(xYz:M_enrich - 1)/xYz:M_enrich].

Calculated values were rounded to the nearest whole number. xYz:M_MutLoad_MinEstimate was calculated only for samples passing FDR=0.05, signifying a statistical over-representation of motif-specific mutagenesis. Samples with FDR>0.05 received a value of 0.

Statistical evaluation of preference to loop or stem locations in predicted RNA secondary structure

The RNAfold function of the ViennaRNA Package 2.0 [29] was used to determine the secondary structure of the complete FASTA sequences of the reference genomes for the SARS-CoV-2

virus (NC_045512.2) and the RA27/3 rubella vaccine virus (FJ211588). A sample command for generating the secondary structure of SARS-CoV-2 genome shown below:

```
RNAfold -d2 --noLP < nc_045512.2.ref.fasta > nc_045512.2.ref.RNAfold.out
```

The output for each analysis (.out) in dot-bracket notation was input into BBEdit (https://www.barebones.com/products/bbedit/) and all characters in both sequence and notation rows were made space delimited. Each of these rows were pasted into Excel and turned into space delimited cells. Sequence and notation were separately copied and pasted using the "Transform" function into a new Excel spreadsheet. A column with the nucleotide position was added and the file saved as a tab delimited text file *RNAfold.txt. For each resulting file the first column was the nucleotide position, the second column is the nucleotide, and the third column was the annotation of that nucleotide in dot-bracket notation. The *RNAfold.txt files were used to add a stem-loop annotation column "RNAfold" to all MAF files using the vlookup function in Excel and saved as a tab delimited text file.

For searching for motifs and trinucleotides, *RNAfold.txt files were used to create a searchable text files as follows. Columns two and three of each file were copied into a two new text files. On command line, the two columns in each file were merged using

```
awk '{$(NF-1)=$(NF-1)""$NF;$NF=""}1' OFS="\t"
```

The output file from this was opened in BBEdit, the line breaks were removed, resulting in a file containing nucleotides and annotation of those nucleotides in a single row as an interleaved and searchable format as CoV2_annot_final.txt and Rubella_annot_final.txt.

These files, displayed in BBEdit, were used to separately count all single nucleotides and all 64 trinucleotides classified as either stem or loop location based on the stem or loop annotation of the individual nucleotide position or of a central position in each trinucleotide.

Statistical evaluation of differences between loop vs stem single base substitution mutagenesis or trinucleotide motif associated mutagenesis was by comparing mutation densities in loop vs stem:

mutLoop/refLoop - density of a substitution type or a trinucleotide motif mutation type in loops where

mutLoop and refLoop are counts in loops of a given type of events mutations or nucleotides in reference, respectively,

and

mutStem/refStem - density of a substitution type or a trinucleotide motif mutation type in stems where

mutStem and refStem are counts in stems of a given type of events mutations or nucleotides in reference, respectively.

Statistical evaluation of loop vs stem mutagenesis was performed by two-tailed Fisher's exact test comparing ratios (mutLoop/(refLoop-mutLoop)) and (mutStem/(refStem-mutStem)) for either base substitutions or for trinucleotide motifs. Fisher's exact test p-value was corrected by Benjamini-Hochberg for the set of 12 possible base substitutions or for 16 possible tri-nucleotide centered around a given base substitution.

Results

Design of the analysis

The overarching hypothesis of this study was that some of the processes generating RNA mutation load in population of SARS-CoV-2 genomes are similar (but not necessarily identical) to the processes that generated changes in genomic RNAs of hypermutated rubella viruses.

For that purpose, we obtained the viral genome FASTA files and processed them to obtain unique mutation calls and the mutation signatures as outlined in Fig 1.

Fig 1. Analysis workflow

MAF – mutation annotation format table

Each filtered MAF combines mutations from all samples into a single dataset.

32,341 FASTA files were downloaded from the GISAID Initiative [30] web site (https://www.gisaid.org/) on 06/22/2020, each containing a consensus whole genome-sequence of a SARS-CoV-2 virus isolated from a human subject and sequenced at high coverage. Based on the published analysis of the GISAID data for a subset of around 4000 of SARS-CoV-2 isolates across the world performed with the use of the Nextstrain package ([31] and https://nextstrain.org/ncov/global?l=clock), an average lineage of SARS-CoV-2 virus successfully transmitted from one subject to another would accumulate approximately 22 base substitutions per year (12-13 base substitutions for the period of December - June, 2020); a similar estimate was also obtained in [32]). The final FASTA sequence files of the individual isolates in GISAID represent a consensus derived from high coverage sequencing reads and contain information about the mutations present with high frequency in a sequenced viral isolate and therefore belong to viral particles capable of proliferation. We aligned each sequence against the sequence of presumably the earliest isolate of a SARS-CoV-2 genome (NC_045512. 2) and listed each change in a separate row of a mutation annotation file (MAF) Fig 1 and S1A Table. We annotated each of 251,273 mutation events in each isolate by surrounding +/- 20 nucleotides of genomic context around position of each mutation, by location in one or in several overlapping ORFs, by potential amino acid change or protein truncation effect, as well as by location of a change in self complementary area (predicted stem) or outside of such area (predicted loop) in plus-strand genomic RNA. Many independent isolates could

have originated from the already mutated virus spreading the same mutation(s) into genomes of multiple (up to thousands) downstream isolates (see column “times_refPos_mutated_inMAF” in S1A Table). Therefore, we also annotated each mutation in a sample by the number of different samples in which such a mutation was found. Since each genome of an individual isolate contained only few mutations and many of these mutations were identical in multiple isolates, we built our analysis to evaluate the overall spectrum of non-redundant mutation events that have accumulated in the human population through the current pandemic rather than the mutation spectra in individual isolates We de-duplicated the starting MAF and created three groups of mutations (S2A,B,C Table). The first group contained a pooled non-redundant set with no duplicates (NoDups) that listed each individual mutation only once regardless of how many isolates contained the same mutation (total 12,156 mutation events). While the individual isolates are not listed in this group any more, it contains a set of distinct events most closely representing the spectrum of unrelated mutation events rather than a complex downstream process of distributing mutant forms in human population. However, even in the NoDups list (S2C Table) the base substitutions in many positions could be under positive or negative selection, which could skew the spectrum of the observed changes from the mechanistic mutation spectrum that accurately reflects the underlying mutagenic processes. Therefore, we subdivided this group based on whether the changes yielded functional effects in the SARS-CoV-2 genome. Non-synonymous amino acid changes and changes introducing or removing stop codons were designated as functional (Func), while synonymous changes or changes outside ORFs were designated and non-functional (NonFunc). The content of NoDupsNonFunc group (S2A Table) would be the least affected by functional selection and thus, most accurately represent the impact of unconstrained mutational processes operating on the viral genome. While this group is smaller, it still contains a sufficient number of changes (4,740 mutation events) for detecting trends in the mutational patterns. The mutation spectrum of 7,416 NoDupsFunc events (S2B Table) was also analyzed. Each of the three SARS-CoV-2 mutation

302 spectra was compared to the combined mutation spectrum (993 base substitutions) from six
303 independent isolates originated from the hypermutated rubella-vaccine virus [19] (S3A Table).
304 Unlike many SARS-CoV-2 isolates, where individual mutated event could be carried from one
305 isolate to another, each rubella isolate contained mutations that had occurred independently
306 from the vaccine virus in each subject. Thus, the total of the mutational events in six rubella
307 isolates was, at least in part, representative of the mutation spectrum. However, mutation
308 spectra in each rubella isolate may represent an unknown level of selection. Indeed, a number
309 of mutations was observed in more than one rubella isolate (see column
310 "times_refPos_mutated_inMAF" in S3A Table). Some level of selection was also indicated by
311 analysis of synonymous and nonsynonymous substitutions in each codon [19]. Therefore, we
312 made separate comparisons of the rubella mutation spectra with each of the three SARS-CoV-2
313 non-redundant MAFs (Fig 1): the non-duplicated mutation events (NoDups), and its two subsets
314 – the non-duplicated mutation events with potential of functional significance (NoDupsFunc) and
315 the non-duplicated non-functional mutation events (NoDupsNonFunc).
316
317 All mutations are reported based on the plus (genomic) strand of the virus. We started from
318 conventional comparisons of all possible single base substitutions and the mutation preference
319 for potential loop or stem parts of ssRNA secondary structure. Unlike in our previous analysis
320 of the mutation spectrum and signatures in the genomes of hypermutated rubella isolates,
321 where we followed only a limited set of motifs based on specific hypotheses, we used here an
322 "agnostic" approach analyzing all 192 possible trinucleotide-centered mutation motifs for
323 enrichment in the viral genomes. We also used existing software to calculate ADAR editing
324 scores [8]. Overall, our methodology allows to detect the mutational signatures that
325 predominate in the viral genomes. Comparisons with the hypermutated rubella genomes further
326 demonstrated the similarities in the mutational processes operating on both viral genomes.
327

Similarity of base substitution spectra between hypermutated rubella genomes and SARS-CoV-2.

We compared the distribution of densities of the 12 possible single base substitutions (counts of each base substitution normalized by the presence of the unmutated base in the genome). While density distributions reflect the contributions of different mutagenic processes in each dataset, density values for specific base substitutions cannot be compared directly between two viruses because they were obtained from vastly different genome numbers. Importantly, there was a statistically significant similarity between the distributions of base substitution densities in rubella and in each of three filtered SARS-CoV-2 MAFs as well as a similarity in several prevailing types of base substitutions (Fig 2 and S4 Table).

Fig 2. Comparison of base substitution spectra between rubella and SARS-CoV-2 datasets from filtered MAFs

The spectrum from each SARS-CoV-2 filtered MAF was compared with rubella spectrum. Bars represent densities of base substitutions in each dataset calculated by dividing counts of each base substitution by counts of the substituted base in the reference sequence. Connecting lines visualize overall parallelism between rubella and each filtered MAF. Insert boxes show Spearman r, its 95% CI, and one-tailed p-value for hypothesis about positive correlation between rubella and SARS-CoV-2 spectra. Source data are in S4 Table.

In both viruses, there was a very high frequency of the C to U changes, consistent with the hypothesis of cytidine deamination in the plus-strand (genomic) RNA. C to U changes in the minus-strand, which would be reported as G to A in the plus-strand, were less abundant in both viruses. Another class of highly abundant changes in both viruses were U to C changes in the plus RNA strand which could originate from A to G changes caused by ADAR adenine

353 deaminase in the minus-strand. The corresponding A to G changes in the plus-strand were less
354 abundant in rubella but were comparable with the C to U changes in SARS-CoV-2.
355
356 A prior study of hypermutated rubella genomes found small, but statistically significant increase
357 in ADAR scores (calculated as described by [8]) in U to C and A to G ADAR-like base
358 substitutions compared to two other types of substitutions in U or A nucleotides [19]. However,
359 no statistically significant increase in ADAR scores was found for the U to C and A to G changes
360 in the SARS-CoV-2 dataset analyzed in a similar way (S1 Fig and S1 Data). Since the ADAR
361 score tool was developed based on in vitro deamination of a perfectly paired dsRNA substrate,
362 there could be a difference in sequence preferences between this substrate and the actual
363 substrate of in vivo editing of SARS-CoV-2 RNA. Alternatively, abundant U to C and A to G
364 ADAR-like editing could be due to mechanisms not involving ADARs.
365
366 The only apparent discrepancy between the two viruses was in a high density of the G to U
367 changes in the plus-strand of SARS-CoV-2, while they were nearly absent in rubella. We note
368 that the density of G to U changes in minus-strand (reported as the complementary C to A
369 changes in plus-strand in Fig 2) was similar to other low abundant changes in both viruses. A
370 possible origin of the increased G to U changes in SARS-CoV-2 genomes will be detailed in
371 Discussion.
372
373 **Several types of base substitutions show preference for regions prone to loop formation**
374 **in viral RNA secondary structure**
375 A high abundance of C to U (or G to A) mutations was already noticed in several recent
376 analyses of SARS-CoV-2 mutation data and inferred to either APOBEC mutagenesis or to
377 errors in RdRp copying of the minus-strand [23, 32, 33]. C to U mutations in RNA can be also
378 caused by non-enzymatic deamination of cytidines similar to such deamination described in

379 DNA [15, 34]. Recently it was revealed that APOBEC3A has a preference for deaminating
380 cytosines in regions prone to forming loops in ssDNA secondary structure [35]. Therefore we
381 annotated all positions in the SARS-CoV-2 and rubella genomes for either preference for loop or
382 stem location in potential secondary structure formed by the RNA plus-strand ([29] S1B and
383 S3B Tables and Methods). We then compared mutation counts in loop vs stem for each type of
384 base substitutions (Fig 3 and S5 Table).
385
386 **Fig 3. Comparison of base substitution mutagenesis between locations prone to loop or**
387 **stem formation in viral RNA genomes**
388 Bars represent densities of base substitutions in stem- or in loop-forming sections. Densities are
389 calculated by dividing counts of each base substitution in either loop or in stem by counts of the
390 substituted base in the loop-forming or in stem-forming regions of the reference sequence.
391 Statistical comparison between mutagenesis in stem vs loop for every base substitution was
392 done by two-tailed Fisher's exact test. P-values were considered after correcting by FDR.
393 Brackets indicate pairs passing FDR=0.05.
394 * <0.05
395 ** <0.005
396 *** <0.0005
397 Source data including exact p-values are in S5 Table.
398
399 In both viruses there was a highly significant preference for loop location with C to U changes in
400 plus-strand. The second type of base changes prevalent in both SARS-CoV-2 and in rubella,
401 the U to C changes in plus-strand (corresponding to A to G changes in minus-strand) did not
402 show statistically significant differences between loop and stem. As discussed above, the U to
403 C (A to G) changes were consistent with ADAR adenine deaminase activity in dsRNA minus-
404 strand, where secondary structure effects are not expected. The only other type of changes

showing statistically significant difference between loop and stem locations in all groups of SARS-CoV-2 mutations were G to U changes. Same as for the base substitution spectra, this outstanding feature of the G to U changes showed up in SARS-CoV-2, but not in rubella (see above).

Mutational motif preferences in SARS-CoV-2 and rubella genomes suggest APOBEC cytidine deaminases as a source of C to U base substitutions in the plus RNA strand.

Since base substitution spectra in SARS-CoV-2 and rubella are correlated, it is likely that they also shared common mechanisms which generated these changes. It is well established that several mechanisms of mutagenesis in DNA of tumors and normal cells can have not only distinctive base substitutions spectra but also diagnostic preference for trinucleotide mutation motifs [26, 36, 37]. Currently there is very little information about motif preference in RNA editing or mutagenesis. Therefore, we assessed enrichment using all possible 192 trinucleotide mutation motifs (96 in plus and 96 in minus RNA strand) of each virus. Enrichment values for each motif were calculated based on counts of mutations in a motif normalized for the motif content in the genomic background (see Methods). Statistical evaluation of enrichments showed significance for several motifs even after FDR<0.05 correction to individual P-values was applied (S6A-C Tables). However, base substitutions for the most-enriched motifs were present in low numbers, so these results require validation in independent studies (also see Discussion). We concentrated on the motifs representing the most abundant types of base substitutions present in both viruses, i.e., on the C to U and U to C changes in plus-strand as well as on their respective complementary changes G to A and A to G. For statistically-significant enriched trinucleotide motifs containing one of these four base substitutions, we calculated the minimum estimates of mutation load (MutLoad) that can be assigned to mechanism(s) with preference for a significantly enriched motif (Fig 4, S6A-D Tables and Materials and Methods).

Fig 4. Trinucleotide-centered mutation motifs with statistically significant enrichment over random mutagenesis.

Bars represent minimum estimates of mutation load that can be assigned to motif-specific mutagenic mechanism (MutLoad) as described in Methods. Statistical evaluation of enrichments was done by two-tailed Fisher's exact test and corrected by FDR including p-values for all 192 possible trinucleotide-centered base substitution motifs. MutLoad for FDR>0.05 = 0. Only results for motifs which included the most frequent base substitutions in the plus-strand, C to U, G to A, A to G, U to C, are shown. Reverse complement motifs in the plus-strand corresponding to the statistically significant motifs mutated in the minus-strand are shown in parentheses. If both plus-strand motifs in the reverse complement pair were statistically-significantly enriched in at least one dataset (in rubella or in a filtered SARS-CoV-2 MAF), they are highlighted in red font. Source data including calculations for all 192 motifs are in S6 Table.

The only revealed similarity between statistically-significant enriched motifs in rubella and in SARS-CoV-2 was for the uCn to uUn changes, consistent with the tCn to tTn ssDNA mutagenesis specificity of a subgroup of APOBEC cytidine deaminases. However, even within the APOBEC-like group of motifs there was a difference between strong enrichment with uCa to uUa motif in rubella and the lack of statistically significant preference for this motif in SARS-CoV-2. There were also three groups of motifs significantly enriched in SARS-CoV-2, but not in rubella (see Tab 1 and Discussion for possible mechanistic assignment of these motifs).

We also assessed the potential loop vs stem preference for trinucleotide motifs containing C to U and G to U single base substitutions that showed overall loop vs stem preference. None of trinucleotide motifs containing C to U base substitutions showed loop or stem preference in

457 selection-free SARS-CoV-2 NoDupsNonFunc filtered dataset and in rubella (S2 Fig and S7A,
458 S7D Tables). Several C to U containing trinucleotide motifs in SARS-CoV-2 datasets, where
459 functional selection cannot be excluded, showed statistically significant bias towards mutations
460 in loops (S2 Fig and S7B,S7C Tables), however more data accumulation is required in order to
461 exclude the confounding effects of functional selection in specific sites. No loop vs stem
462 preference was detected in trinucleotide motifs containing G to U substitutions either in SARS-
463 CoV-2 or in rubella (S3 Fig and S8A-C Tables).
464
465 In summary, our agnostic analysis of trinucleotide signature motifs demonstrated that the uCn
466 (tCn) APOBEC-like mutagenesis, which is a major component in rubella hypermutation, also
467 contributes towards the mutations accumulated in the genomes of infectious SARS-CoV-2
468 spreading in the current pandemic.
469
470 **Discussion**
471 Previously, we demonstrated that hypermutation of the inactivated rubella vaccine virus can
472 generate infectious virus particles in immunocompromised children [19]. Based on this work,
473 we hypothesized that similar mutagenic processes may act upon the genomes of other similar
474 plus-strand RNA viruses like SARS-CoV-2. The large-scale sequencing efforts producing
475 genomes of tens of thousands of SARS-CoV-2 isolates allowed us to accurately identify
476 mutations, build a mutation catalog for this virus, highlight similarities with hypermutated rubella
477 and reveal unique features of SARS-CoV-2 mutagenesis (summarized in Table 1).
478

Table 1. Analyses of base substitutions prevailing in rubella and in SARS-CoV-2 plus-strand genomic RNAs

Feature	Virus	C to U	G to A	A to G	U to C
Prevalence of a base substitution	rubella	High frequency	Frequent; Less frequent than C to U	High frequency; Less frequent than U to C	High frequency; More frequent than A to G
	SARS-CoV-2	High frequency	Frequent; Less frequent than C to U	High frequency	High frequency
Secondary structure element preferred by base substitution	rubella	Prefers loops over stems	ND	ND	ND
	SARS-CoV-2	Prefers loops over stems	ND	ND	ND
Enriched trinucleotide motif(s)	rubella	uCn to uUn	ND	ND	ND
	SARS-CoV-2	uCn to uUn; aCn to aUn	cGn to cAn (reverse complement for nCg to nTg in minus-strand);	nAu to nGu (reverse complement to motif preferred by U to C)	aUn to aCn (reverse complement to motif preferred by A to G)
Suggested mechanism	rubella and/or SARS-CoV-2	(i) Frequent C to U editing by tCn (uCn) - specific APOBEC(s) in plus-strand; (ii) new motif aCn to aUn	(i) Increased C-deamination in nCg (CpG) minus-strand motif (SARS-CoV-2 only); (ii) No statistical support to APOBEC editing in minus-strands of either virus	A to G editing by ADAR(s) in plus-strand	A to G editing by ADAR(s) in minus-strand

ND – not detected

Comparisons to hypermutated rubella strains demonstrated that the base substitution spectra correlate between the two viruses. Two types of base substitutions – C to U (and its complementary G to A) and A to G (and its complementary U to C), expected from endogenous mutagenesis by APOBECs and ADARs, respectively, prevailed in both viruses (Fig 2).

Since RNA can form secondary structures, any mutagenic processes active upon ssRNA would preferentially be formed in the loop regions of the secondary structures. Analysis of A to G and U to C substitutions consistent with the biochemical specificity of ADARs did not reveal any preference for mutagenesis in loops versus stem regions. ADARs are known to act on dsRNA substrates. Thus, if ADARs did in fact contribute to induction of these substitutions, they should be acting on a dsRNA form, wherein we do not expect RNA to fold into secondary structures. We also found in SARS-CoV-2, but not in rubella, statistically significant enrichment of the nAu to nGu mutations along with its reverse complement aUn to aCn. While these could reflect adenine deamination by one of the ADARs in either strand of a dsRNA intermediate, these motifs are different from the motif preference revealed by *in vitro* editing of artificial dsRNA substrate [8]. Also, there was no increase in ADAR scores in the SARS-CoV-2 A to G or U to C mutations (S1 Fig and S1 Data), thus indicating that ADARs may not be the primary source of these changes. Alternatively, it is possible that these enriched mutation motifs are preferred by ADARs only *in vivo.*

Unlike A to G changes, in-depth analysis of C to U substitutions revealed that they were predominantly present in the RNA plus-strands of both viruses and demonstrated a preference for loops versus stems in the RNA secondary structure (Fig 3 and S5 Table). This phenomenon is similar to the preference of ssDNA and mRNA editing in loops by the APOBEC3A cytidine deaminase [35, 38]. Agnostic analysis of enrichments in all 192 possible trinucleotide mutation motifs highlighted statistically significant excess of uCa to uUa motif in rubella, however these

509 changes were not prevalent in SARS-CoV-2. Mutations in plus-strands of both viruses showed
510 statistically significant enrichments with uCg to uUg and uCu to uUu motifs (Fig 4A,B). These
511 motifs belong to a group uCn to uUn (tCn to tTn in DNA) which is characteristic of several
512 APOBEC cytidine deaminases ([14] and references therein). We note that although these
513 signatures were enriched in the non-functional mutations (NoDupsNonFunc), they did not pass
514 the 0.05 FDR threshold in filtered datasets that included mutations with potential functional
515 effects (NoDupsFunc). These differences in the mutation signatures between SARS-Cov-2 and
516 rubella may be due to different APOBEC family members performing editing or due to the
517 confounding presence of other sources of C to U mutagenesis, such as spontaneous cytosine
518 deamination that frequently occurs in ssDNA [39] or oxidative mutagenesis capable of
519 generating C to T mutations in ssDNA *in vivo* [40, 41]. In support of the role of oxidative
520 damage in SARS-CoV-2 genomes, is the increased prevalence of G to U substitutions which is
521 consistent with the oxidation of guanines in the RNA plus-strand (Fig 2). G to U changes could
522 be caused by an increased level of oxidative damage generating 8-oxoG in viral RNA within
523 cells or during sequencing library preparation [42, 43]. Frequent copying of 8-oxoG with A,
524 would show up as G to U changes in the strand, where 8-oxoG was present. However, since
525 we analyze the consensus sequences of the viral genomes and not individual reads, errors
526 during library preparation would most likely be filtered out and would not be represented in the
527 viral genome sequence. On the other hand, G to U changes were present only at low density in
528 hypermutated rubella genomes indicating physiological differences between the two viruses.
529
530 There were two more groups of trinucleotide mutation motifs involving C to U (and
531 complementary G to A) substitutions in plus RNA strand specifically enriched for SARS-CoV-2
532 (Fig 4A,B). The aCn to aUn (reverse complement nGu to nAu) group of motifs may represent a
533 preference previously unknown for APOBECs in RNA or just a mutagenic mechanism yet to be
534 defined. The cGn to cAn group of motifs seen in the plus-strand may be in fact due to mutations

of the reverse complement motif nCg to nUg in the minus-strand. nCg to nTg (CpG to TpG) germline and somatic mutagenesis is universally present in DNA of species with 5-methylcytosine and is generated by systems specialized to mutagenesis in methylated CpG sequences. However various studies have demonstrated that CpG to TpG mutagenesis can occur independent of cytosine methylation [44, 45]. Several studies have shown that CpG dinucleotides are depleted in the genomes of SARS viruses indicating functional selection and/or increased frequency of cytosine deamination in these viral genomes [46-49]. Our study shows with high statistical confidence that nCg to nUg (CpG to UpG) mutagenesis in the minus strand is enriched (Fig 4B) supporting the role of nCg- (CpG)-specific cytosine deamination in minus RNA strand in SARS-CoV-2 genomic mutagenesis.

In summary, comparison of base substitution spectra and signatures between hypermutated rubella isolates and the SARS-CoV-2 multi-genome dataset demonstrates both similarities and differences in the mutational processes active upon the two plus-strand RNA viruses. It is important to understand the mechanisms that contribute to mutagenesis of viral genomes, since hypermutation of even inactivated rubella vaccine virus was shown to generate reactivated viral particles [19]. We demonstrate here that the APOBEC-specific uCa to uUa changes that are highly enriched in hypermutated rubella, are much less prevalent in SARS-CoV-2. We propose that assessment of uCa to uUa signature in viral genomes can provide insights into the potential hypermutation risk of SARS-CoV-2. Moreover, understanding the genomic mutational patterns is important for predicting virus evolution. Our study has highlighted several distinct features of SARS-CoV-2 mutational spectrum that, after validation with independent dataset(s) can be used to build predictive models for this and related SARS viruses.

559 **Acknowledgments**

560 We thank Dr. Natalya Degtyareva and Mr. Adam Burkholder for critical reading of the
561 manuscript and Mr. Michael Doubintchik for help in the initial evaluation of the on-line databases
562 of SARS-CoV-2 genomes. We gratefully acknowledge the authors, as well as the originating
563 laboratories for submitting the sequences to GISAID's EpiCoV™ Database on which this
564 research is based; complete list of contributors is provided in S9 Table.

565

References

1. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. Journal of virology. 2010;84(19):9733-48. Epub 2010/07/28. doi: 10.1128/JVI.00694-10. PubMed PMID: 20660197; PubMed Central PMCID: PMCPMC2937809.

2. Domingo E, Perales C. Viral quasispecies. PLoS Genet. 2019;15(10):e1008271. Epub 2019/10/18. doi: 10.1371/journal.pgen.1008271. PubMed PMID: 31622336; PubMed Central PMCID: PMCPMC6797082.

3. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA biology. 2011;8(2):270-9. Epub 2011/05/20. doi: 10.4161/rna.8.2.15013. PubMed PMID: 21593585; PubMed Central PMCID: PMCPMC3127101.

4. Posthuma CC, Te Velthuis AJW, Snijder EJ. Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes. Virus Res. 2017;234:58-73. Epub 2017/02/09. doi: 10.1016/j.virusres.2017.01.023. PubMed PMID: 28174054; PubMed Central PMCID: PMCPMC7114556.

5. Domingo E. Molecular Basis of Genetic Variation of Viruses. In: Domingo E, editor. Virus as Populations. Boston: Academic Press; 2016. p. 35-71.

6. Samuel CE. Adenosine deaminase acting on RNA (ADAR1), a suppressor of double-stranded RNA-triggered innate immune responses. J Biol Chem. 2019;294(5):1710-20. Epub 2019/02/03. doi: 10.1074/jbc.TM118.004166. PubMed PMID: 30710018; PubMed Central PMCID: PMCPMC6364763.

7. Conticello SG. Creative deaminases, self-inflicted damage, and genome evolution. Annals of the New York Academy of Sciences. 2012;1267:79-85. Epub 2012/09/08. doi: 10.1111/j.1749-6632.2012.06614.x. PubMed PMID: 22954220.

592 8. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded
593 RNA. Nat Commun. 2011;2(1):319. Epub 2011/05/19. doi: 10.1038/ncomms1324. PubMed
594 PMID: 21587236; PubMed Central PMCID: PMCPMC3113232.
595 9. Blango MG, Bass BL. Identification of the long, edited dsRNAome of LPS-stimulated
596 immune cells. Genome research. 2016;26(6):852-62. Epub 2016/05/20. doi:
597 10.1101/gr.203992.116. PubMed PMID: 27197207; PubMed Central PMCID:
598 PMCPMC4889969.
599 10. Niavarani A, Currie E, Reyal Y, Anjos-Afonso F, Horswell S, Griessinger E, et al.
600 APOBEC3A Is Implicated in a Novel Class of G-to-A mRNA Editing in WT1 Transcripts. PloS
601 one. 2015;10(3):e0120089. doi: 10.1371/journal.pone.0120089. PubMed PMID: 25807502;
602 PubMed Central PMCID: PMC4373805.
603 11. Sharma S, Patnaik SK, Taggart RT, Kannisto ED, Enriquez SM, Gollnick P, et al.
604 APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nat
605 Commun. 2015;6:6881. Epub 2015/04/22. doi: 10.1038/ncomms7881. PubMed PMID:
606 25898173; PubMed Central PMCID: PMCPMC4411297.
607 12. Smith HC, Bennett RP, Kizilyer A, McDougall WM, Prohaska KM. Functions and
608 regulation of the APOBEC family of proteins. Seminars in cell & developmental biology.
609 2012;23(3):258-68. Epub 2011/10/18. doi: 10.1016/j.semcdb.2011.10.004. PubMed PMID:
610 22001110; PubMed Central PMCID: PMC4017262.
611 13. Chan K, Gordenin DA. Clusters of Multiple Mutations: Incidence and Molecular
612 Mechanisms. Annual review of genetics. 2015;49:243-67. Epub 2015/12/04. doi:
613 10.1146/annurev-genet-112414-054714. PubMed PMID: 26631512; PubMed Central PMCID:
614 PMCPMC4710516.
615 14. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A
616 hypermutation signature is distinguishable from the signature of background mutagenesis by

617 APOBEC3B in human cancers. Nature genetics. 2015;47(9):1067-72. Epub 2015/08/11. doi:
618 10.1038/ng.3378. PubMed PMID: 26258849; PubMed Central PMCID: PMCPMC4594173.
619 15. Chan K, Sterling JF, Roberts SA, Bhagwat AS, Resnick MA, Gordenin DA. Base
620 damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous
621 environmental agent. PLoS Genet. 2012;8(12):e1003149. Epub 2012/12/29. doi:
622 10.1371/journal.pgen.1003149. PubMed PMID: 23271983; PubMed Central PMCID:
623 PMCPMC3521656.
624 16. Harris RS, Dudley JP. APOBECs and virus restriction. Virology. 2015;479-480(0):131-
625 45. Epub 2015/03/31. doi: 10.1016/j.virol.2015.03.012. PubMed PMID: 25818029; PubMed
626 Central PMCID: PMCPMC4424171.
627 17. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and
628 mechanisms. Nature reviews Cancer. 2014;14(12):786-800. Epub 2014/11/25. doi:
629 10.1038/nrc3816. PubMed PMID: 25568919; PubMed Central PMCID: PMCPMC4280484.
630 18. Saraconi G, Severi F, Sala C, Mattiuz G, Conticello SG. The RNA editing enzyme
631 APOBEC1 induces somatic mutations and a compatible mutational signature is present in
632 esophageal adenocarcinomas. Genome Biol. 2014;15(7):417. Epub 2014/08/03. doi:
633 10.1186/s13059-014-0417-z. PubMed PMID: 25085003; PubMed Central PMCID:
634 PMCPMC4144122.
635 19. Perelygina L, Chen MH, Suppiah S, Adebayo A, Abernathy E, Dorsey M, et al. Infectious
636 vaccine-derived rubella viruses emerge, persist, and evolve in cutaneous granulomas of
637 children with primary immunodeficiencies. PLoS pathogens. 2019;15(10):e1008080. Epub
638 2019/10/29. doi: 10.1371/journal.ppat.1008080. PubMed PMID: 31658304; PubMed Central
639 PMCID: PMCPMC6837625.
640 20. Lee JY, Bowden DS. Rubella virus replication and links to teratogenicity. Clin Microbiol
641 Rev. 2000;13(4):571-87. Epub 2000/10/12. doi: 10.1128/cmr.13.4.571-587.2000. PubMed
642 PMID: 11023958; PubMed Central PMCID: PMCPMC88950.

643 21. de Wilde AH, Snijder EJ, Kikkert M, van Hemert MJ. Host Factors in Coronavirus
644 Replication. In: Tripp RA, Tompkins SM, editors. Roles of Host Gene and Non-coding RNA
645 Expression in Virus Infection. Cham: Springer International Publishing; 2018. p. 1-42.
646 22. Paul D, Bartenschlager R. Architecture and biogenesis of plus-strand RNA virus
647 replication factories. World journal of virology. 2013;2(2):32-48. Epub 2013/11/01. doi:
648 10.5501/wjv.v2.i2.32. PubMed PMID: 24175228; PubMed Central PMCID: PMCPMC3785047.
649 23. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-
650 dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 2020;6(25):eabb5813.
651 Epub 2020/07/01. doi: 10.1126/sciadv.abb5813. PubMed PMID: 32596474; PubMed Central
652 PMCID: PMCPMC7299625.
653 24. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile
654 and open software for comparing large genomes. Genome Biol. 2004;5(2):R12. Epub
655 2004/02/05. doi: 10.1186/gb-2004-5-2-r12. PubMed PMID: 14759262; PubMed Central PMCID:
656 PMCPMC395750.
657 25. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from
658 high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. Epub 2010/07/06. doi:
659 10.1093/nar/gkq603. PubMed PMID: 20601685; PubMed Central PMCID: PMCPMC2938201.
660 26. Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, et al. The Impact of
661 Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts.
662 PLoS Genet. 2016;12(10):e1006385. Epub 2016/10/28. doi: 10.1371/journal.pgen.1006385.
663 PubMed PMID: 27788131.
664 27. Saini N, Sterling JF, Sakofsky CJ, Giacobone CK, Klimczak LJ, Burkholder AB, et al.
665 Mutation signatures specific to DNA alkylating agents in yeast and cancers. Nucleic Acids Res.
666 2020;48(7):3692-707. Epub 2020/03/07. doi: 10.1093/nar/gkaa150. PubMed PMID: 32133535.
667 28. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
668 occurrences of k-mers. Bioinformatics. 2011;27(6):764-70. Epub 2011/01/11. doi:

669 10.1093/bioinformatics/btr011. PubMed PMID: 21217122; PubMed Central PMCID:
670 PMCPMC3051319.
671 29. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al.
672 ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):26. Epub 2011/11/26. doi:
673 10.1186/1748-7188-6-26. PubMed PMID: 22115189; PubMed Central PMCID:
674 PMCPMC3319429.
675 30. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision
676 to reality. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European
677 communicable disease bulletin. 2017;22(13). Epub 2017/04/07. doi: 10.2807/1560-
678 7917.ES.2017.22.13.30494. PubMed PMID: 28382917; PubMed Central PMCID:
679 PMCPMC5388101.
680 31. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-
681 time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3. Epub 2018/05/24. doi:
682 10.1093/bioinformatics/bty407. PubMed PMID: 29790939; PubMed Central PMCID:
683 PMCPMC6247931.
684 32. Chiara M, Horner DS, Gissi C, Pesole G. Comparative genomics suggests limited
685 variability and similar evolutionary patterns between major clades of SARS-CoV-2.
686 2020:2020.03.30.016790. doi: 10.1101/2020.03.30.016790 %J bioRxiv.
687 33. Matyasek R, Kovarik A. Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13
688 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid
689 Evolution in Their Hosts. Genes. 2020;11(7). Epub 2020/07/11. doi: 10.3390/genes11070761.
690 PubMed PMID: 32646049.
691 34. Lindahl T. Instability and decay of the primary structure of DNA. Nature.
692 1993;362(6422):709-15. Epub 1993/04/22. doi: 10.1038/362709a0. PubMed PMID: 8469282.
693 35. Jalili P, Bowen D, Langenbucher A, Park S, Aguirre K, Corcoran RB, et al. Quantification
694 of ongoing APOBEC3A activity in tumor cells by monitoring RNA editing at hotspots. Nat

695 Commun. 2020;11(1):2971. Epub 2020/06/14. doi: 10.1038/s41467-020-16802-8. PubMed
696 PMID: 32532990.
697 36. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The
698 repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94-101. Epub
699 2020/02/07. doi: 10.1038/s41586-020-1943-3. PubMed PMID: 32025018; PubMed Central
700 PMCID: PMCPMC7054213.
701 37. Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, et al. The
702 genome as a record of environmental exposure. Mutagenesis. 2015;30(6):763-70. doi:
703 10.1093/mutage/gev073. PubMed PMID: 26443852; PubMed Central PMCID:
704 PMCPMC4637815.
705 38. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al. Passenger
706 hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science.
707 2019;364(6447):eaaw2872. Epub 2019/06/30. doi: 10.1126/science.aaw2872. PubMed PMID:
708 31249028; PubMed Central PMCID: PMCPMC6731024.
709 39. Lindahl T, Nyberg B. Heat-induced deamination of cytosine residues in deoxyribonucleic
710 acid. Biochemistry. 1974;13(16):3405-10. Epub 1974/07/30. doi: 10.1021/bi00713a035.
711 PubMed PMID: 4601435.
712 40. Degtyareva NP, Heyburn L, Sterling J, Resnick MA, Gordenin DA, Doetsch PW.
713 Oxidative stress-induced mutagenesis in single-strand DNA occurs primarily at cytosines and is
714 DNA polymerase zeta-dependent only for adenines and guanines. Nucleic Acids Res.
715 2013;41(19):8995-9005. Epub 2013/08/09. doi: 10.1093/nar/gkt671. PubMed PMID: 23925127;
716 PubMed Central PMCID: PMCPMC3799438.
717 41. Degtyareva NP, Saini N, Sterling JF, Placentra VC, Klimczak LJ, Gordenin DA, et al.
718 Mutational signatures of redox stress in yeast single-strand DNA and of aging in human
719 mitochondrial DNA share a common feature. PLoS biology. 2019;17(5):e3000263. Epub

720 2019/05/09. doi: 10.1371/journal.pbio.3000263. PubMed PMID: 31067233; PubMed Central
721 PMCID: PMCPMC6527239.
722 42. Boo SH, Kim YK. The emerging role of RNA modifications in the regulation of mRNA
723 stability. Exp Mol Med. 2020;52(3):400-8. Epub 2020/03/27. doi: 10.1038/s12276-020-0407-z.
724 PubMed PMID: 32210357; PubMed Central PMCID: PMCPMC7156397.
725 43. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery
726 and characterization of artifactual mutations in deep coverage targeted capture sequencing data
727 due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41(6):e67.
728 Epub 2013/01/11. doi: 10.1093/nar/gks1443. PubMed PMID: 23303777; PubMed Central
729 PMCID: PMCPMC3616734.
730 44. Alsoe L, Sarno A, Carracedo S, Domanska D, Dingler F, Lirussi L, et al. Uracil
731 Accumulation and Mutagenesis Dominated by Cytosine Deamination in CpG Dinucleotides in
732 Mice Lacking UNG and SMUG1. Sci Rep. 2017;7(1):7199. Epub 2017/08/05. doi:
733 10.1038/s41598-017-07314-5. PubMed PMID: 28775312; PubMed Central PMCID:
734 PMCPMC5543110.
735 45. Behringer MG, Hall DW. Genome-Wide Estimates of Mutation Rates and Spectrum in
736 Schizosaccharomyces pombe Indicate CpG Sites are Highly Mutagenic Despite the Absence of
737 DNA Methylation. G3 (Bethesda, Md). 2015;6(1):149-60. Epub 2015/11/14. doi:
738 10.1534/g3.115.022129. PubMed PMID: 26564949; PubMed Central PMCID:
739 PMCPMC4704713.
740 46. Xia X. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral
741 defense. Molecular biology and evolution. 2020. Epub 2020/04/15. doi:
742 10.1093/molbev/msaa094. PubMed PMID: 32289821; PubMed Central PMCID:
743 PMCPMC7184484.
744 47. Woo PC, Wong BH, Huang Y, Lau SK, Yuen KY. Cytosine deamination and selection of
745 CpG suppressed clones are the two major independent biological forces that shape codon

usage bias in coronaviruses. Virology. 2007;369(2):431-42. Epub 2007/09/21. doi: 10.1016/j.virol.2007.08.010. PubMed PMID: 17881030; PubMed Central PMCID: PMCPMC7103290.

48. Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. eLife. 2014;3:e04531. Epub 2014/12/10. doi: 10.7554/eLife.04531. PubMed PMID: 25490153; PubMed Central PMCID: PMCPMC4383024.

49. Grigoriev A. Mutational patterns correlate with genome organization in SARS and other coronaviruses. Trends in genetics : TIG. 2004;20(3):131-5. Epub 2004/03/31. doi: 10.1016/j.tig.2004.01.009. PubMed PMID: 15049309; PubMed Central PMCID: PMCPMC7127256.

Supporting Information

S1 Fig. Mean values of ADAR scores calculated as described in Methods.

Source data are in S1 Dataset.

S2 Fig. Comparison of trinucleotide-centered motif C to U mutation densities between locations prone to loop or to stem formation in viral RNA genomes

Bars represent densities of base substitutions in stem- or in loop-forming regions. Densities are calculated by dividing counts of each motif mutations in either loop or in stem by counts of this motif in the loop-forming or in stem-forming regions of the reference sequence. Statistical comparison between mutagenesis in stem vs loop for every base substitution was done by two-tailed Fisher's exact test. P-values were corrected by FDR including 16 motifs containing C to U base substitution. Brackets indicate pairs passing FDR=0.05.

* <0.05

** <0.005

Source data are in S7 Table.

S3 Fig. Comparison of trinucleotide-centered motif G to U mutation densities between locations prone to loop or to stem formation in viral RNA genomes

Bars represent densities of base substitutions in stem- or in loop-forming regions. Densities are calculated by dividing counts of each motif mutations in either loop or in stem by counts of this motif in the loop-forming or in stem-forming regions of the reference sequence. Statistical comparison between mutagenesis in stem vs loop for every base substitution was done by two-tailed Fisher's exact test. P-values were corrected by FDR including 16 motifs containing G to U base substitution. Brackets indicate pairs passing FDR=0.05.

* <0.05

Source data are in S8 Table.

S1 Data. ADAR scores and complete ADAR analysis for SARS-CoV-2 filtered MAFs (contains source data for S1 Fig).

S1 Table. Mutation calls and RNA fold prediction in SARS-CoV-2 genomes

S1A Table. Complete list of mutation calls in all SARS-CoV-2 genomes in TCGA compatible Mutation Annotation Format (MAF); nucleotides named as in DNA.

S1B Table. Annotation of predicted RNA-fold in SARS-CoV-2 reference positions.

S2 Table. SARS-CoV-2 filtered Mutation Annotation Files (MAFs)

S2A_Table. NoDupsNonFunc - de-duplicated set of mutations from all samples of the dataset; non-functional

S2B_Table. NoDupsFunc - de-duplicated set of mutations from all samples of the dataset; aminoacid changes or protein-truncating

S2C_Table. NoDups - de-duplicated set of mutations from all samples of the dataset

S3 Table. Mutation calls and RNA fold prediction in rubella genomes

S3A Table. The list of 993 mutations in six rubella isolates (from [19])

Sequences are shown in DNA format (T instead of U) to maintain compatibility with other outputs of the mutation signature R-script

S1B Table. Annotation of predicted RNA fold in rubella reference

S4 Table. Counts and densities of single base substitutions in SARS-CoV-2 and in rubella

S4A Table. Counts of base substitutions

S4B Table. Densities of base substitutions. (Source data for Fig 2)

S5 Table. Comparison of base substitution densities between locations prone to loop or stem formation in viral RNA genomes (Source data for Fig. 3)

S5A Table. SARS-CoV-2, NoDupsNonFunc - de-duplicated set of mutations from all samples of the dataset; non-functional

S5B Table. SARS-CoV-2, NoDupsFunc - de-duplicated set of mutations from all samples of the dataset; aminoacid changes or protein-truncating

S5C Table. SARS-CoV-2, NoDups - de-duplicated set of mutations from all samples of the dataset

S5D Table. Rubella, all mutations

S6 Table. Statistical evaluation of mutagenesis in 192 trinucleotide-centered mutation motifs

Sequences are shown in DNA format (T instead of U) to maintain compatibility with other outputs of the mutation signature R-script

S6A Table. SARS-CoV-2, NoDupsNonFunc - de-duplicated set of mutations from all samples of the dataset; non-functional

S6B Table. SARS-CoV-2, NoDupsFunc - de-duplicated set of mutations from all samples of the dataset; aminoacid changes or protein-truncating

S6C Table. SARS-CoV-2, NoDups - de-duplicated set of mutations from all samples of the dataset

S6D Table. Rubella, all mutations

836 **S7 Table. Comparison of C to U trinucleotide motif substitution densities between**
837 **locations prone to loop or stem formation in viral RNA genomes**
838
839 S7A Table. SARS-CoV-2, NoDupsNonFunc - de-duplicated set of mutations from all samples of
840 the dataset; non-functional
841 S7B Table. SARS-CoV-2, NoDupsFunc - de-duplicated set of mutations from all samples of the
842 dataset; aminoacid changes or protein-truncating
843 S7C Table. SARS-CoV-2, NoDups - de-duplicated set of mutations from all samples of the
844 dataset
845 S7D Table. Rubella, all mutations
846
847 **S8 Table. Comparison of G to U trinucleotide motif substitution densities between**
848 **locations prone to loop or stem formation in viral RNA genomes**
849
850 S8A Table. SARS-CoV-2, NoDupsNonFunc - de-duplicated set of mutations from all samples of
851 the dataset; non-functional
852 S8B Table. SARS-CoV-2, NoDupsFunc - de-duplicated set of mutations from all samples of the
853 dataset; aminoacid changes or protein-truncating
854 S8C Table. SARS-CoV-2, NoDups - de-duplicated set of mutations from all samples of the
855 dataset
856 S8D Table. Rubella, all mutations
857
858 **S9 Table. Acknowledgments to research groups and individuals provided SARS-CoV-2**
859 **genome sequences to GISAID's EpiCoV™ Database**.

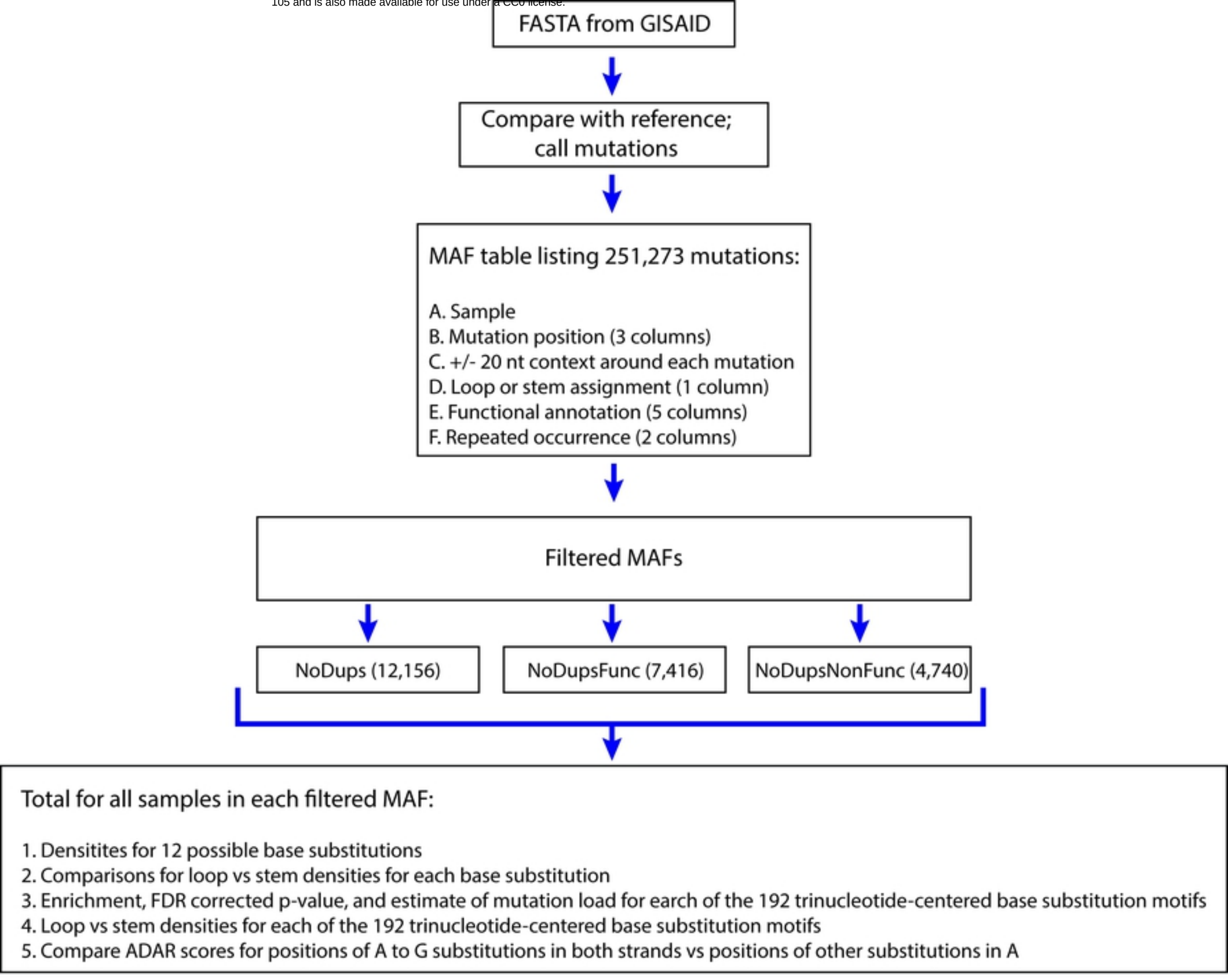
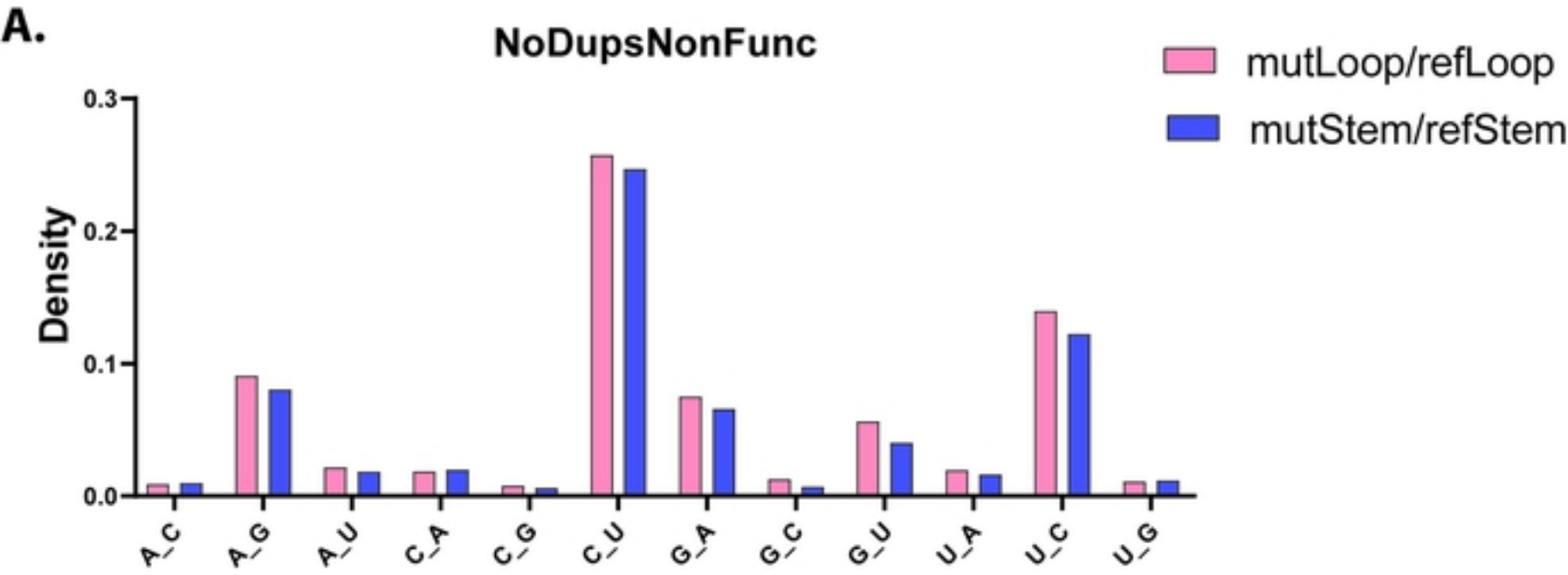
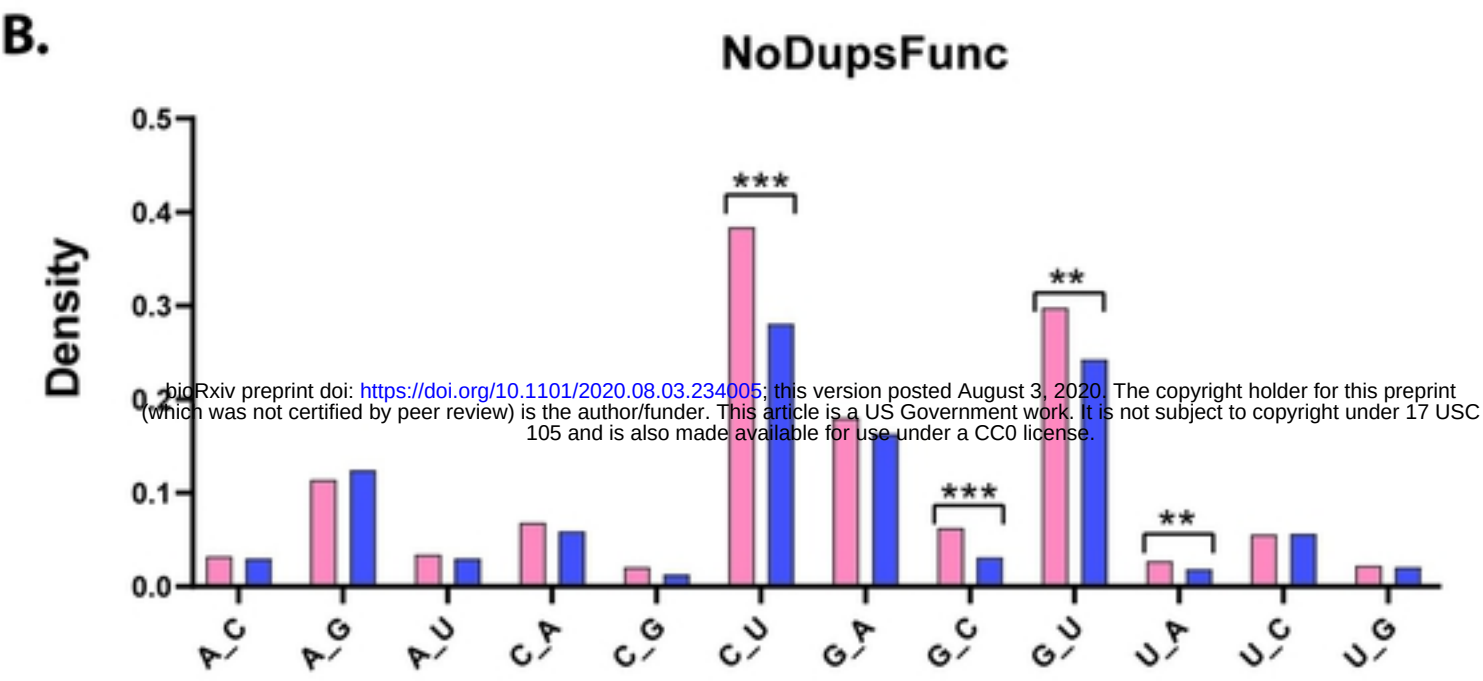
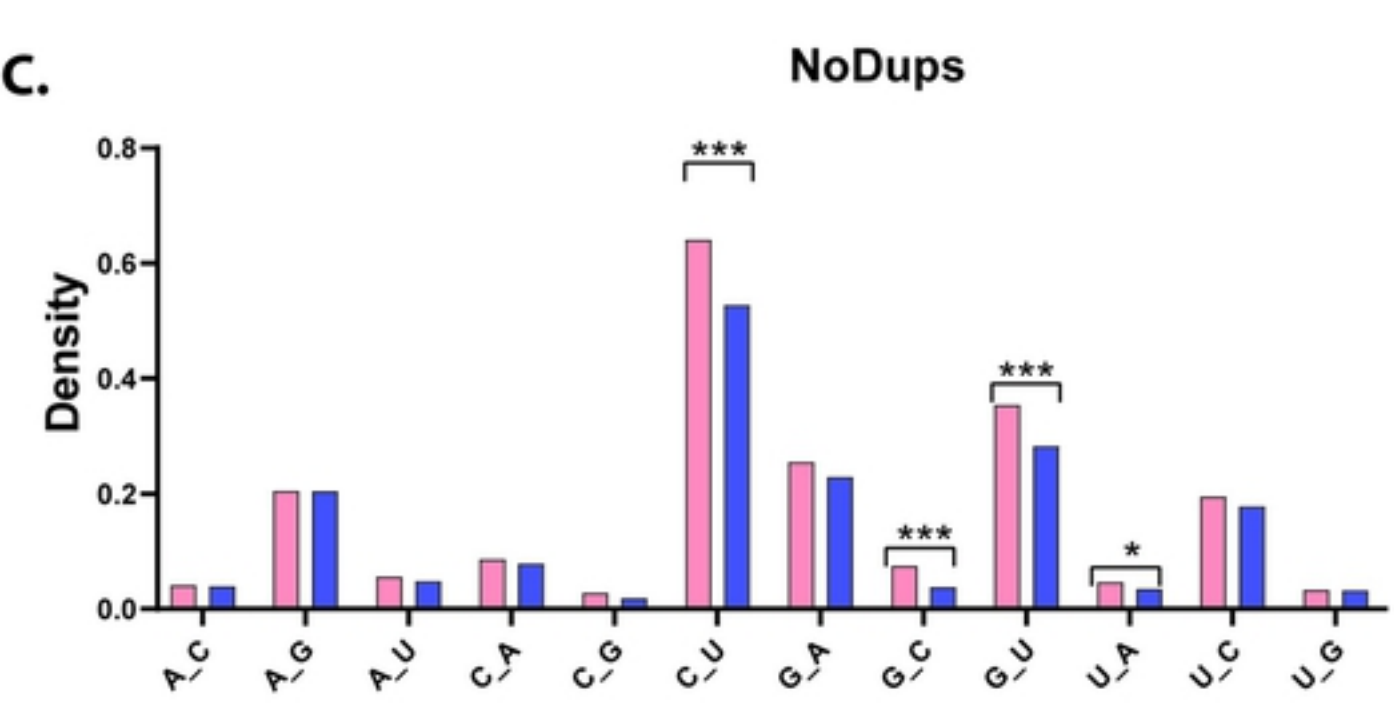


Fig 1

Fig 2







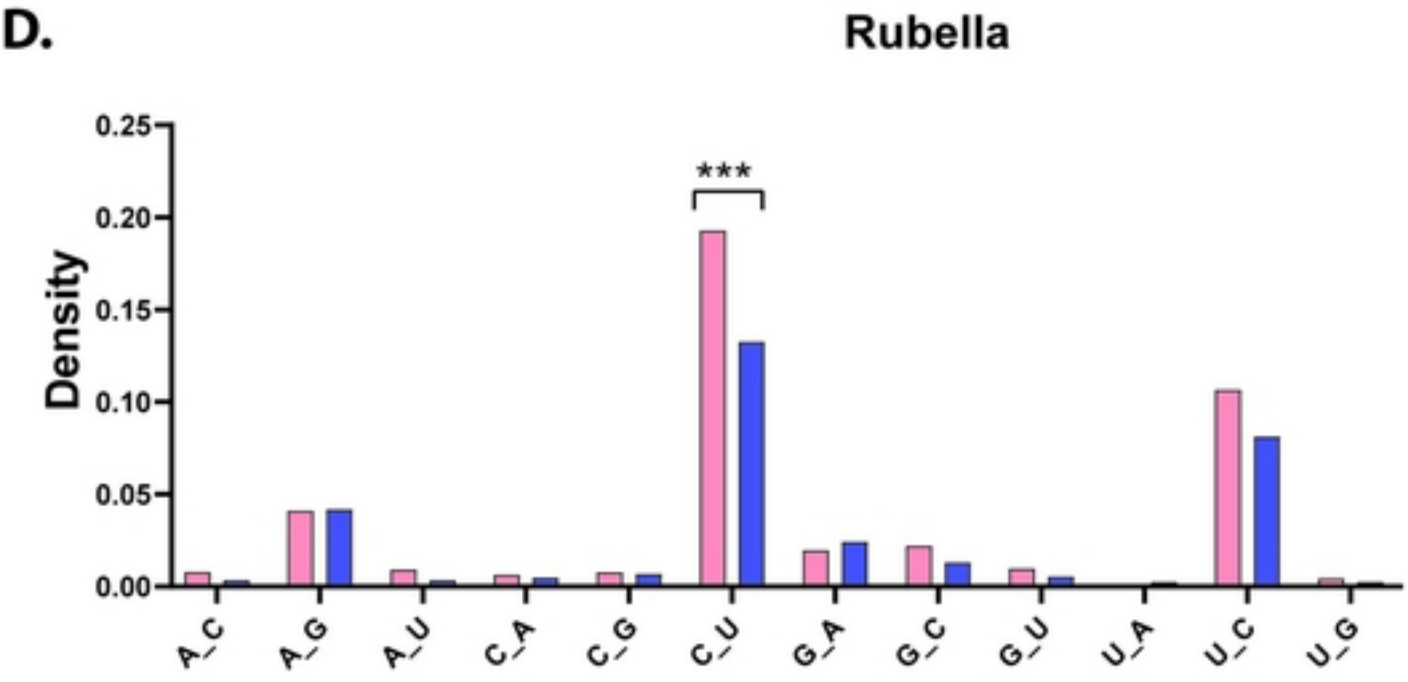


Fig 3

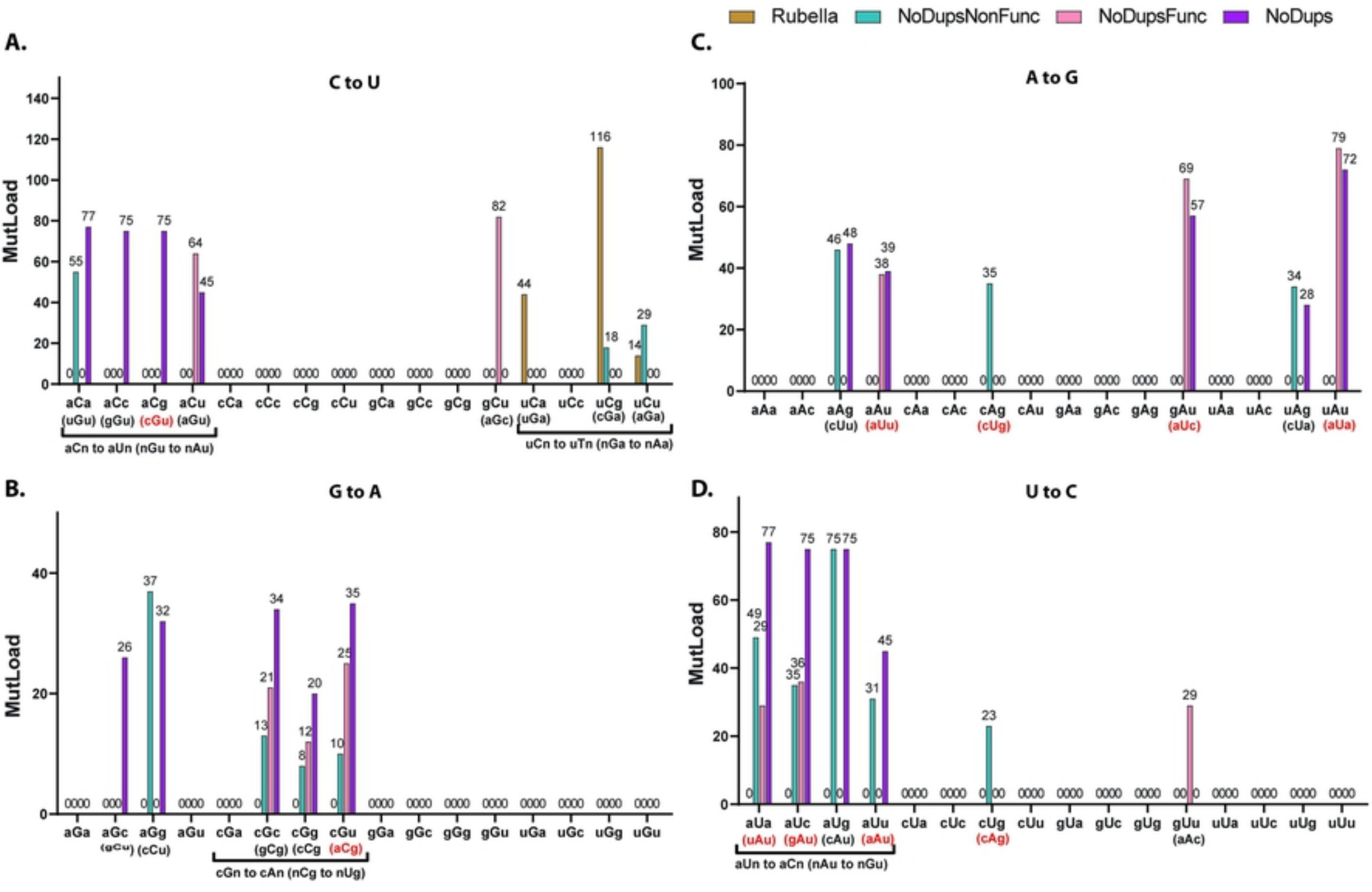

Rubella
NoDupsNonFunc
NoDupsFunc
NoDups
A.
C to U
MutLoad
aCa (uGu)
aCc (gGu)
aCg (cGu)
aCu (aGu)
cCa
cCc
cCg
cCu
gCa
gCc
gCg
gCu (aGc)
uCa (uGa)
uCc
uCg (cGa)
uCu (aGa)
aCn to aUn (nGu to nAu)
uCn to uTn (nGa to nAa)
B.
G to A
MutLoad
aGa
aGc (gCu)
aGg (cCu)
aGu
cGa
cGc (gCg)
cGg (cCg)
cGu (aCg)
gGa
gGc
gGg
gGu
uGa
uGc
uGg
uGu
cGn to cAn (nCg to nUg)
C.
A to G
MutLoad
aAa
aAc
aAg (cUu)
aAu (aUu)
cAa
cAc
cAg (cUg)
cAu
gAa
gAc
gAg
gAu (aUc)
uAa
uAc
uAg (cUa)
uAu (aUa)
D.
U to C
MutLoad
aUa (uAu)
aUc (gAu)
aUg (cAu)
aUu (aAu)
cUa
cUc
cUg (cAg)
cUu
gUa
gUc
gUg
gUu (aAc)
uUa
uUc
uUg
uUu
aUn to aCn (nAu to nGu)

Fig 4