Generation of a novel SARS-CoV-2 sub-genomic RNA due to the R203K/G204R variant in nucleocapsid: homologous recombination has potential to change SARS-CoV-2 at both protein and RNA level

Shay Leary[1¶], Silvana Gaudieri[1,2,3¶], Matthew D. Parker[4¶], Abha Chopra[1], Ian James[1], Suman Pakala[3], Eric Alves[2], Mina John[1,5], Benjamin B. Lindsey[6,7], Alexander J Keeley[6,7], Sarah L. Rowland-Jones[6,7], Maurice S. Swanson[8], David A. Ostrov[9], Jodi L. Bubenik[8], Suman Das[3], John Sidney[10], Alessandro Sette[10,11], COVID-19 Genomics Consortium UK, Thushan I. de Silva[6,7*], Elizabeth Phillips[1,3*], Simon Mallal[1,3#*]

[1]Institute for Immunology and Infectious Diseases, Murdoch University, Murdoch, Western Australia, Australia.

[2]School of Human Sciences, University of Western Australia, Crawley, Western Australia, Australia.

[3]Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

[4]Sheffield Biomedical Research Centre, Sheffield Bioinformatics Core, The University of Sheffield, Sheffield, UK.

[5]Department of Clinical Immunology, Royal Perth Hospital, Perth, Western Australia, Australia.

[6]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK.

[7]Department of Infection, Immunity and Cardiovascular Disease and The Florey Institute for Host-Pathogen Interactions, Medical School, University of Sheffield, Sheffield, UK.

[8]Department of Molecular Genetics and Microbiology, Center for NeuroGenetics and the Genetics Institute, University of Florida, Gainesville, Florida, USA.

[9]Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, Florida, USA.

[10]Center for Infectious Disease and Vaccine Research, La Jolla Institute for Immunology, La Jolla, California, USA.

[11]Department of Medicine, Division of Infectious Diseases and Global Public Health, University of California, San Diego, La Jolla, California, USA.

[¶]These authors contributed equally to this work.

[*]These authors also contributed equally to this work.

[#]Corresponding author

Prof. Simon Mallal

Email: s.mallal@vumc.org

Keywords: COVID-19; SARS-CoV-2; homologous recombination; sub-genomic RNA transcript; transcription-regulating sequence; viral polymorphism

Abstract

Background: Genetic variations across the SARS-CoV-2 genome may influence transmissibility of the virus and the host's anti-viral immune response, in turn affecting the frequency of variants over-time. In this study, we examined the adjacent amino acid polymorphisms in the nucleocapsid (R203K/G204R) of SARS-CoV-2 that arose on the background of the spike D614G change and describe how strains harboring these changes became dominant circulating strains globally. **Methods:** Deep sequencing data of SARS-CoV-2 from public databases and from clinical samples were analyzed to identify and map genetic variants and sub-genomic RNA transcripts across the genome. **Results:** Sequence analysis suggests that the three adjacent nucleotide changes that result in the K203/R204 variant have arisen by homologous recombination from the core sequence (CS) of the leader transcription-regulating sequence (TRS) rather than by stepwise mutation. The resulting sequence changes generate a novel sub-genomic RNA transcript for the C-terminal dimerization domain of nucleocapsid. Deep sequencing data from 981 clinical samples confirmed the presence of the novel TRS-CS-dimerization domain RNA in individuals with the K203/R204 variant. Quantification of sub-genomic RNA indicates that viruses with the K203/R204 variant may also have increased expression of sub-genomic RNA from other open reading frames. **Conclusions:** The finding that homologous recombination from the TRS may have occurred since the introduction of SARS-CoV-2 in humans resulting in both coding changes and novel sub-genomic RNA transcripts suggests this as a mechanism for diversification and adaptation within its new host.

Introduction

It is believed SARS-CoV-2 originated from a bat coronavirus transmitted to humans, likely via an intermediate host such as a pangolin, acquiring a furin-cleavage site in the process. This new motif allows cleavage at the boundary of the S1 and S2 domains of the spike protein in virus-producing cells (1). A SARS-CoV-2 variant in the spike protein, D614G (B.1 lineage), emerged early in the epidemic and has rapidly became dominant in virtually all areas of the world where it has circulated (2). Several studies have shown this variant to be associated with higher viral RNA levels in the upper respiratory tract, higher titers in pseudoviruses in-vitro (2, 3) and increased infectivity (4, 5). More recently, emerging lineages from this genetic background (B.1.1.7 – 'Alpha or UK variant', B.1.351 – 'Beta or South African variant', or B.1.617.2 - 'Delta variant') have been identified with reported rapid local expansions of these viruses.

The diversification of coronaviruses can occur via point mutations and recombination events (6, 7) that can result in increased prevalence due to selective advantage related to increased infectiousness and transmission of the virus or by chance. Evidence of viral adaptation to selective pressures as a virus spreads among diverse human populations has important implications for the ongoing potential for changes in viral fitness over time, which in turn may impact transmissibility, disease pathogenesis and immunogenicity. Furthermore, the functional impact of new genetic changes need to be considered in the performance of diagnostic tests, ongoing public health measures to contain infection around the world and the development of universal vaccines and antiviral therapies including monoclonal antibodies.

Here we examined a variant of SARS-CoV-2 that emerged within the subset of sequences harboring the D614G variant and contains three adjacent nucleotide changes spanning two residues of the nucleocapsid protein (R203K/G204R; B.1.1 lineage) that has resulted in a novel sub-genomic RNA transcript. Sequence analysis suggests these changes are the result of homologous recombination from the core sequence (CS) of the leader transcription-regulating sequence (TRS). This event introduced a new TRS between the RNA binding and dimerization domains of nucleocapsid providing the template for the generation of a novel sub-genomic RNA transcript. Further novel sub-genomic RNA transcripts arising in association with incorporation of leader sequence and TRS were also observed, suggesting homologous recombination from this region as a potential mechanism for SARS-CoV-2 diversification and adaptation within its new host.

102 **Methods**

103 **Study Design**

104 This study utilized deposited SARS-CoV-2 genomic sequences in public databases, with a
105 further 981 Oxford Nanopore Technology genomes and clinical metadata from Sheffield,
106 UK, as a validation set, to identify and map genetic variants and sub-genomic RNA
107 transcripts across the genome. Accession numbers and links to datasets are in Supplementary
108 Material.

109

110 **SARS-CoV-2 sequence generation from patients with COVID-19**

111 SARS-CoV-2 sequences, with matched clinical metadata, were generated using samples
112 taken for routine clinical diagnostic use from 981 individuals presenting with COVID-19
113 disease to Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. This work
114 was performed under approval by the Public Health England Research Ethics and
115 Governance Group for the COVID-19 Genomics UK consortium (R&D NR0195).
116 Following extraction, samples were processed using the ARTIC Network SARS-CoV-2
117 protocol. After RT-PCR, SARS-CoV-2 specific PCR and library preparation with Oxford
118 Nanopore LSK-109 and barcoding expansion packs NBD-104 and NBD-114 samples were
119 sequenced on an Oxford Nanopore GridION X5 using R9.4.1D flow cells. Bases were called
120 with either fast or high accuracy guppy with demultiplexing enabled and set to --require-
121 both-ends. Samples were then analyzed using ARTIC Network pipeline v1.1.0rc1.

122

123 **SARS-CoV-2 sequence acquisition from public repositories**

124 Complete SARS-CoV-2 genome sequences were downloaded from the GISAID EpiCoV
125 repository on 24th January 2021 (https://www.gisaid.org/). The complete dataset of 455,774
126 sequences with coverage across the genome were aligned in CLCbio Genomics Workbench

127 12 (QIAGEN Bioinformatics) to the GenBank reference sequence NC_045512.2. Aligned
128 sequences were exported in FASTA format and imported into Visual Genomics Analysis
129 Studio (VGAS), an in-house program for visualizing and analyzing sequencing data
130 (http://www.iiid.com.au/software/vgas). The chronological appearance of the sequences was
131 generated using the collection dates for each of the sequences. Of note, our current
132 knowledge of the global circulating variants is dependent on the ability of laboratories in
133 different countries to deposit full genome length SARS-CoV-2 sequences and may be subject
134 to ascertainment bias. As such, the frequencies of specific variants shown may not reflect the
135 size of the outbreak. However, the data does provide the opportunity to predict the presence
136 of specific variants in areas given the known epidemiology within different countries and
137 regions. A subset of subjects also had individual deep sequence reads deposited in the
138 Sequence Read Archive (SRA) at www.ncbi.nlm.nih/sra. These sequence reads were
139 downloaded and aligned as indicated above.
140
141 **Identification of amino acid substitutions**
142 Codon usage output allowed for identification of amino acid substitutions across the SARS-
143 Cov-2 genome. A cut-off of 5% frequency within the consensus SARS-CoV-2 protein
144 sequences was set to obtain the codon usage across all sequences and as shown in S1 Table.
145 The viral polymorphisms detected are present in viral variants sequenced using different
146 NGS platforms (e.g. nanopore, Illumina) and the Sanger-based sequencing method making it
147 unlikely that the new changes are sequence or alignment errors. In addition, different
148 laboratories around the world have deposited sequences with these polymorphisms in the
149 database and examination of individual sequences in the region failed to uncover obvious
150 insertions/deletions likely representing alignment issues or homopolymer slippage.
151

152 **HLA peptide binding prediction**

153 The region containing the adjacent amino acid polymorphisms in the nucleocapsid was
154 divided into sliding windows of 8-14 amino acids. NetMHC 4.0
155 (http://www.cbs.dtu.dk/services/NetMHC/) and NetMHCpan 4.0
156 (http://www.cbs.dtu.dk/services/NetMHCpan/) with default settings were utilized to predict
157 HLA-class I binding scores and binding differences across all HLA class-I alleles for the
158 original 2003 SARS and current SARS-CoV-2 sequences harboring the R203/G204 and
159 K203/R204 polymorphisms in the nucleocapsid (output listed in S2 Table).

160

161 **HLA peptide binding assays**

162 MHC was purified from the Steinlin EBV transformed homozygous cell line (IHWG ID:
163 9087; A*01:01, B*08:01 and C*07:01) using the B123.2 (anti-HLA-B, C) and W6/32 (anti-
164 class I) monoclonal antibodies, and classical MHC-peptide inhibition of binding assays
165 performed, as previously described (8). To develop an HLA C*07:01-specific binding assay,
166 the IEDB was utilized to identify candidate peptides reported as HLA-C*07:01 epitopes or
167 eluted ligands. One peptide (3424.0028; sequence IRSSYIRVL, Macaca mulatta and Homo
168 sapiens DNA replication licensing factor MCM5 289-297) was radiolabeled and found in
169 direct binding assays to yield a strong signal with as little as 0.5 nM MHC. Subsequent
170 inhibition of binding assays established that 3424.0028 bound with an affinity of 0.21 nM. To
171 establish that the putative assay was specific for C*07:01, and not co-purified B*08:01, two
172 additional peptides previously reported as HLA-C*07:01 ligands were also tested, with one
173 found to bind with high affinity (IC50 67 nM) and the other with intermediate (IC50 1600
174 nM). At the same time, a panel of known B*08:01 ligands were not found to have the
175 capacity to inhibit binding of radiolabeled 3424.0028 (S3 Table). By contrast, when the same
176 panel of peptides was tested in the previously validated B*08:01 assay (9), 3424.0028 was

177 found to bind with about 1500-fold lower affinity, all of the known B*08:01 ligands bound
178 with IC50s <10 nM, and the C*07:01 ligands with affinities >1000 nM.
179
180 **Sub-genomic RNA classification & quantification in the Validation Dataset**
181 We developed a tool, "periscope" (v0.0.0), to classify and quantify sub-genomic RNA in the
182 Sheffield ARTIC network Nanopore dataset (10). The tool can be downloaded from git-hub
183 at https://github.com/sheffield-bioinformatics-core/periscope. Briefly, this tool uses local
184 alignment to identify putative sub-genomic RNA supporting reads and uses genomic reads
185 from the same amplicon to normalize.
186
187 **RNA structure modeling**
188 The RNAfold program from the ViennaRNA Web Server (http://rna.tbi.univie.ac.at/) was
189 used for structural predictions using the default settings and the minimum free energy
190 structures were acquired using the base-pairing probability color scheme. The Dot-bracket
191 folding notations were obtained for each of the R203K/G204R sequences and used for
192 Junction Explorer (nature.njit.edu/biosoft/Junction-Explorer/) and CHS-align
193 (nature.njit.edu/biosoft/CHSalign/).
194

195 **Statistical Analysis**

196 Fisher exact test was used to compare the proportion of subjects with specific sub-genomic
197 RNA transcripts. P values less than 0.05 was used as the statistical threshold. Comparisons
198 between sub-genomic and genomic RNA expression in R203/G204 compared to K203/R204
199 containing sequences was made using the Mann-Whitney U test, corrected for multiple
200 comparisons using the Holm method. Logistic and linear regression modeling used to explore
201 the impact of K203/R204 and other co-variates on hospitalization, CT values and sub-
202 genomic RNA expression.

203

204

205 **Results and Discussion**

206 **Adjacent nucleocapsid polymorphisms emerged from the existing spike protein D614G**
207 **variant**

208 We utilized publicly available SARS-CoV-2 sequences from the GISAID database (available
209 on the 24th of January 2021; www.gisaid.org) to identify amino acid polymorphisms arising
210 in global circulating forms of the virus in relation to region and time of collection. Of the
211 455,774 circulating variants there were 29 amino acid polymorphisms present in >5% of the
212 deposited sequences (of a total of 9413 sites; S1 Table) including the spike D614G variant
213 (B.1 lineage) that emerged early in the pandemic and the adjacent R203K/G204R variants
214 (B.1.1 lineage) in the nucleocapsid protein (11) that formed one of the main variants
215 emerging from Europe in early 2020. As of the end of January 2021, the K203/R204 variant
216 comprises 37.4% of globally reported SARS-CoV-2 sequences (Fig 1) and almost
217 exclusively occurs on the D614G genetic background (S4 Table).

218

219 Although the D614G change rapidly increased in prevalence in almost all regions, the
220 prevalence rates of the K203/R204 subset of the D614G variant are variable in different
221 geographic areas and over-time (Fig 2). For example, an almost complete replacement of
222 D614 by G614 was noted in South America between March and April 2020 and a similar
223 replacement pattern was seen with the K203/R204 variant most marked in Chile, Argentina
224 and Brazil (12). A closer examination of the deposited sequences in the UK shows the
225 K203/R204 variant increasing in prevalence early in 2020 but the second wave later in the
226 year shows a shift in the proportion of deposited sequences with the R203/G204 subset of the
227 D614G variant (B.1.177 lineage) until the recent appearance of the B.1.1.7 'Alpha or UK
228 variant' that harbors the K203/R204 polymorphisms (S1 Fig and S4 Table); supporting a
229 likely increased infectivity of this variant.

230

231 **Amino acid polymorphisms due to three adjacent nucleotide changes in the**
232 **nucleocapsid likely due to homologous recombination**

233 Of the publicly available sequences examined with the two amino acid polymorphisms
234 K203/R204, all showed the three adjacent nucleotide changes from AGG GGA to AAA
235 CGA. There was no differential codon usage for the K203/R204 variant in the database.
236 However, there was evidence of low frequency alternative codon usage for arginine at 203
237 (AGA) for the R203/G204 variant and for lysine (AAG) at 203 for the K203/G204 variant
238 (S5 Table). Overall, circulating variants that contain the intermediate codon as the consensus
239 that could facilitate a single step from the AGG arginine codon to the AAA lysine codon at
240 position 203 appear rare among captured variants to date (S5 Table). Furthermore, a K203
241 polymorphism alone was seen in 0.3% and an R204 polymorphism alone seen in only 0.02%
242 of sequences (S5 Table). The low frequency K203/L204 and K203/P204 variants are both
243 one nucleotide step from the K203/R204 variant, have been deposited into the public
244 databases (November 2020) well after the emergence of the K203/R204 variant (February
245 2020) and accordingly likely arose from this genetic background.

246

247 The rapid emergence of closely linked polymorphisms in viruses can also reflect strong
248 selection pressure on this region of the genome in which the original mutation incurred a
249 replicative capacity, or other fitness cost, which could be restored by a linked compensatory
250 mutation. Evidence for such adaptations with closely linked compensatory mutations are
251 known to occur under host immune pressure as is well established for other RNA viruses
252 such as HIV (13-15) and Hepatitis C virus (16). In the absence of anti-viral treatment, these
253 viruses have such a high rate of viral replication, error-prone polymerases and lack associated
254 proofreading, mismatch repair, and other nucleic acid repair pathways generating a swarm of

255 viral variants with ongoing recombination between variants (in the case of HIV) being
256 generated continuously. As a result, selection pressure exerted by immune responses or other
257 selective pressures effectively operate on each separate residue independently (15). In
258 contrast, coronaviruses encode proofreading machinery and have a propensity to adapt by
259 homologous recombination between viruses (6) rather than necessarily by classic stepwise
260 individual mutations driven by selective pressures effectively operating on individual viral
261 residues. Furthermore, a simulation based on the nucleocapsid genomic region and allowing
262 up to 10 random mutations indicates the likelihood of observing three consecutive nucleotide
263 changes is less than 0.0005. These findings argue against stepwise change of the nucleotides
264 for the R203K/G204R variant.
265
266 The introduction of the AAACGA motif by homologous or heterologous recombination is a
267 more parsimonious mechanistic explanation and would have immediately resulted in both an
268 R to K change and adjacent G to R change at the positions 203 and 204, respectively. It is
269 critical to determine if the introduction of the AAACGA motif has induced any replicative or
270 other fitness change for the virus as a result of either structural or functional changes in the
271 RNA or the concomitant change of amino acids from R203/G204 to K203/R204 and any
272 related structural or functional impact on the nucleocapsid protein.
273
274 **SARS-CoV-2 itself as likely source for homologous recombination**
275 To identify possible viral sources for homologous recombination with SARS-CoV-2,
276 we initially performed a search of the motif in the nucleocapsid in related beta coronaviruses
277 from human and other species in the public databases and only found the presence of the
278 R203/G204 combination. We performed a similar search in our metatranscriptome data
279 generated from a cohort study consisting of 65 subjects of whom 43 had acute respiratory

280 infections and 22 were asymptomatic. From the data we assembled near complete and coding
281 complete viral genomes of the Coronavirus (NL63 - alpha, OC43 - beta, 229E - alpha), RSV
282 (A, B), Rhinovirus (A, B, C), Influenza (A - H3N2), and Bocavirus family. None of the alpha
283 coronaviruses had the R203/G204 or K203/R204 combination or indeed any variation at
284 these sites (n=14; sequence depth >3000). We then performed a search for stretches of
285 similarity using varying window sizes (>14 base-pair (bp) including the motif) in all
286 sequences. A 14bp window was selected as 14bp has been shown to be the minimum amount
287 of homology required for homologous recombination in mammalian cells (17). No significant
288 hits were identified. However, the AAACGA sequence encoding the K203/R204 amino acids
289 overlaps with the CTAAACGAAC motif of the leader transcription-regulating sequences
290 (TRS; core underlined) (18) of SARS-CoV-2 itself and this core sequence motif is also found
291 near the start codon of the protein for surface glycoprotein (S), ORF3a, E, M, ORF6, ORF7a,
292 ORF8, ORF10 and nucleocapsid, in keeping with its known roles in mediating template
293 switching and discontinuous transcription (18).
294

295 **Deep sequencing confirms quasi-species with the leader sequence linked to known or**
296 **introduced TRS region**

297 Discontinuous transcription of SARS-CoV-2 results in sub-genomic RNA (sgRNA)
298 transcripts containing 5'-leader sequence-TRS-start codon-ORF-3'. These RNA transcripts
299 should also be captured from reads generated from NGS platforms. We therefore reasoned we
300 should be able to find such sequences within deep sequencing reads at the sites of known
301 sub-genomic regions (corresponding to the ORFs) and adjacent to position 203/204 of the
302 nucleocapsid in subjects infected with the K203/R204 variant but not in those with the
303 R203/G204 variant (Fig 3).
304

305 We searched for sgRNAs in sequence data generated from n=981 patients with COVID-19
306 based on the ARTIC network protocol (www.artic.network/ncov-2019; Fig 3) and subsequent
307 Nanopore sequencing in Sheffield, UK. As expected, the most frequent sgRNA transcripts in
308 each subject, irrespective of variant, corresponded to the known regions containing the start
309 codon of the SARS-CoV-2 proteins (Fig 4A). However, out of a total of 550 K203/R204
310 sequences, 231 had evidence (>=1 read containing leader sequence at the novel TRS site) of
311 the non-canonical nucleocapsid sgRNA (42%) but only 1 out of a total of 431 R203/G204
312 subjects had evidence of the novel sgRNA (likely a false positive as described in S2 Fig).
313
314 We confirmed the presence of the novel non-canonical nucleocapsid sgRNA in 27/45
315 individuals with the K203/R204 variant but in none of 45 individuals with the R203/G204
316 variant (Fisher test, p=5.0e-11; S6 Table) from the sequence read archive (SRA) database
317 (www.ncbi.nlm.nih/sra). Interestingly, we also found the presence of 23 other non-canonical
318 sgRNA transcripts with the 5'-leader-TRS-start codon-3' at low frequency in the 90 subjects
319 (irrespective of variant) due to multiple adjacent changes to the consensus sequence across
320 the genome generating new core TRS motifs (including with minor mismatches) (S6 Table).
321 It should be noted that none of these changes are present in the consensus sequence of the
322 SARS-CoV-2 genomes downloaded and represent low frequency quasispecies within
323 individuals. It does, however, suggest other instances of the introduction of the core
324 sequences from the leader TRS elsewhere in the SARS-CoV-2 genome.
325
326 **SARS-CoV-2 viruses with K203/R204 are not associated with greater hospitalization**
327 **with COVID-19 or higher virus levels in the upper respiratory tract**
328 The same dataset from COVID-19 patients in Sheffield, UK, was used to explore whether the
329 K203/R204 variant had any association with clinical outcome. The median age of this cohort

330 was 54 years (IQR 38 to 74) and 59.8% were female. Of these, 440 (44.9%) were
331 hospitalized COVID-19 patients and 42 (4.3%) subsequently required critical care support. A
332 multivariable logistic regression model including 203/204 status, age and sex showed no
333 association of K203/R204 with hospitalization (OR 0.82, 95% confidence intervals (CI 0.58 –
334 1.16), p=0.259). As expected, higher age and male sex were significantly association with
335 hospitalization with COVID-19 (OR 1.09, 95% CI 1.08 – 1.11, p <2e-16 for age and OR
336 4.47, 95% CI 3.13 – 6.43, p=2.91e-16 for male sex). Male sex, but not age or 203/204 status,
337 was associated with risk of critical care admission (S7 Table).
338
339 We explored whether K203/R204 was associated with greater virus levels in the upper
340 respiratory tract as estimated by cycle threshold (CT) values from the diagnostic RT-PCR. As
341 day of illness will impact CT value, we focused on a subset of the cohort (n=478) where this
342 information was available (all non-hospitalized patients, median symptom day 3, range 1 – 13
343 days). Data were analyzed with sequences stratified by spike 614 and nucleocapsid 203/204
344 status (D614/R203/G204, G614/R203/G204 and G614/K203/R204). Multivariable linear
345 regression models showed no impact of G614/K203/R204 compared to G614/R203/G204
346 status on CT values (p= 0.83, S6B Table), but as expected, later day of symptom onset was
347 significantly associated with higher CT values, therefore lower viral load (S8 Table,
348 p=2.05E-05). Consistent with recent findings (2), presence of a spike D614G variant was
349 significantly associated with lower CT values (higher viral loads) in the same subset of
350 individuals, even when day of illness at sampling is included in the model (S8A Table,
351 D614/R203/G204 vs G614/R203/G204, p=0.00011, Fig 5A & B).
352
353 **SARS-CoV-2 viruses with K203/R204 have evidence of higher sub-genomic RNA**
354 **expression**

355 We hypothesized that the amount of sgRNA at each of the ORF TRS positions in the SARS-
356 CoV-2 genome in ARTIC nanopore sequencing data could serve as a proxy for expression
357 levels of each of the ORFs due to their positions in the amplicons (Fig 3). To test this
358 hypothesis we developed a tool, periscope (19), which quantifies the number of sgRNA and
359 genomic RNA reads at each ORF TRS position in ARTIC network nanopore sequencing
360 data. We applied periscope to the 981 sequences in the Sheffield validation dataset. To
361 control for the sequencing depth differences evident between amplicons, we determined the
362 amplicon that shares the 3' primer with the sgRNA reads and used the total count of genomic
363 RNA at this amplicon to calculate the proportion of sgRNA for each ORF. The N ORF
364 sgRNA is expressed at high levels in all samples. ORF10 sgRNA was absent as others have
365 shown (20). A significant increase in sgRNA levels for several ORFs in samples with
366 K203/R204 compared to R203/G204 samples is apparent (Fig 4B). N is the most striking
367 example (Fig 4C, Mann-Whitney U test p value, adjusted for multiple testing p = 2.06e-37),
368 but sgRNA from ORFs E, M and ORF6 are also significantly increased. There is no
369 significant difference in genomic RNA levels (Fig 4D, normalized to total mapped reads)
370 between these two groups.
371
372 As discussed above, the K203/R204 variants appear to have emerged within the subset of
373 SARS-CoV-2 sequences with a D614G variant in the spike protein, which has been
374 associated with infections with a higher viral load in the upper respiratory tract. To explore
375 whether the differences between K203/R204 and R203/G204 sequences in sgRNA quantities
376 were due to D614 compared to G614 variant differences, we repeated the comparisons
377 following further stratification of sequences. Interestingly, G614/R203/G204 variants showed
378 *lower* total sgRNA expression than D614/R203/G204 samples (S3 Fig). Of note, sgRNA for
379 spike (S), membrane (M) and envelope (E) ORFs were significantly higher in samples with

380 D614/R203/G204 compared to those with G614/R203/G204 (adjusted p values 1.02e-4 for S,
381 0.0495 for M and 0.00696 for E). Total sgRNA in G614/K203/R204-containing samples was
382 still significantly higher than in G614/R203/G204 samples (S3A Fig, Mann-Whitney U test p
383 value, adjusted for multiple testing p = 3.5e-6). Similar increases in some individual ORF
384 sgRNA quantities in G614/K203/R204 compared to G614/R203/G204 sequences were also
385 seen, most notably for nucleocapsid (S3B Fig, adjusted p value 1.34e-12).
386
387 To ensure that the increase in sgRNA in K203/R204-containing sequences was not due to
388 confounding by differences in sampling date compared to date of symptom onset, we
389 evaluated the impact of K203/R204 and day of illness on sgRNA expression in a
390 multivariable linear regression model using the subset of 478 sequences described above
391 (stratified by D614/R203/G204, G614/R203/G204 and G614/K203/R204 status). Higher
392 sgRNA levels were significantly associated with later day from symptom onset (S9 Table,
393 p=9.9E-08). G614/R203/G204 compared to D614/R203/G204 was again associated with a
394 reduction in sgRNA levels (p=0.011, S9A Table), whereas a K203/R204 change on the
395 background of spike G614-containing sequences was associated with a significant increase in
396 sub-genomic RNA (p=4.51E-05, S9B Table). Spike canonical sub-genomic RNA was higher
397 in D614/R203/G204 samples, whereas nucleocapsid canonical sub-genomic RNA was higher
398 in G614/K203/R204 samples (Fig 5C and D, S3 Fig).
399
400 RT-PCR assays have been developed to directly assess sub-genomic mRNA (sgRNA) as a
401 measure of replicative intermediates of SARS-CoV-2 representing putative replication in
402 cells rather than RNA packaged in virions or residual viral RNA (21, 22). A decline in
403 sgRNA in sputum typically occurs from day 10 to 11 after onset of symptoms (22). Our
404 finding that a variant can emerge that is associated with a novel sub-genomic RNA or may

405 differentially impact the level of different sgRNAs suggest that the viral sequences should be
406 analyzed to ensure the primers or probes used are appropriate and analysis of short read deep
407 sequences with the periscope tool considered to help interpret results obtained from different
408 variants.
409
410 **Potential impact of introduced TRS sequences on RNA structure**
411 Modeling of the region around the mRNA encoding position 203 and 204 of the nucleocapsid
412 using RNAfold (23) predicts the presence of a three-way junction in the RNA (S4 Fig),
413 which was also predicted using Junction-Explorer (24). Three-way junction motifs are
414 common throughout biology and are found both in pure RNAs, such as riboswitches or
415 ribozymes, and in RNA-protein complexes, including the ribosome (25). RNA three-way
416 junctions are often stabilized via terminal loop interactions with distant tertiary contacts
417 while the junctions act like flexible hinges. These attributes allow these structures to sample
418 unusual conformational spaces and they often form platforms for interactions with other
419 molecules such as proteins, RNAs or small molecule ligands (25), and these folds often have
420 an essential role in either the function or assembly of the molecules in which they are
421 contained.
422
423 RNAfold predicts the mutation from AGGGGA to AAACGA strongly disrupts this structure
424 as the lengths of the predicted helices and each of the junctions are altered and the stability of
425 Helix 2 is undermined (S4 Fig). A comparison of the two-modeled sequences using
426 CHSalign (26) also indicates that none of the junctions are maintained. Given these
427 widespread alterations, this modeling predicts that the AGGGGA to AAACGA mutation
428 would have a strong impact on the local RNA structure of this region, and likely impacts the
429 normal function of this three-way junction motif. Interestingly, the RNA modeling shown in

430 S4 Fig also suggests that pairing of specific nucleotides to maintain these RNA structures
431 may require the preferential codon usage by RG (AGGGGA) and KR (AAACGA) and be a
432 contributory factor to preferential codon usage in RNA viruses more generally even in
433 protein coding regions.
434
435 While it is not possible to determine the impact of this proposed structural alteration on
436 SARS-CoV-2 without a defined function for this structure, there are precedents where minor
437 changes in a three-way junction have large functional consequences for their host viruses. For
438 example, Flaviviruses such as Dengue and West Nile virus utilize the host cell machinery to
439 degrade viral genomes until they encounter structures near the 3' end that are resistant to
440 XRN1 5'-3' exonuclease (27). The resulting small flaviviral RNAs (sfRNAs) are non-coding
441 RNAs that induce cytopathicity and pathogenicity. The resistance of sfRNA to XRN1 is
442 dependent on the structure of a three-way junction and a single nucleotide change at the
443 junction alters the fold sufficiently to prevent the accumulation of disease-related sfRNAs.
444 Thus, small changes at the nucleotide level can have profound functional consequences for
445 viral RNA three-way junctions.
446
447 **Lack of evidence that the RG to KR change at positions 203 and 204 of nucleocapsid**
448 **was driven by HLA-restricted immune selective pressure**
449 Selection of viral adaptations to polymorphic host responses mediated by T cells, NK-cells
450 and antibodies are well described for other RNA viruses such as HIV and HCV (15, 28).
451 HIV-1 adaptations to human leucocyte antigen (HLA)-restricted T-cell responses have also
452 been shown to be transmitted and accumulate over time (29, 30). As previously shown for
453 SARS-CoV, T-cell responses against SARS-CoV-2 are likely to target the nucleocapsid (31).
454 Notably, SARS-CoV-2 R203K/G204R polymorphisms modify the predicted binding of

455 putative HLA-restricted T-cell epitopes containing these residues (S2 Table). One of the
456 predicted T-cell epitopes is restricted by the HLA-C*07 allele; and we therefore considered
457 whether escape from HLA-C-restricted T-cell responses may conceivably confer a fitness
458 advantage for SARS-CoV-2, particularly in European populations where HLA-C*07 is
459 prevalent and carried by >40% of the population (www.allelefrequencies.net). However,
460 using HLA-C*07:01 purified from the Steinlin cell line (IHWG ID: 9087; A*01:01, B*08:01
461 and C*07:01) and the anti-HLA Class I B123.2 mAb in inhibition assays we were not able to
462 detect binding of either of the SARS-CoV-2 peptides SRGTSPARM or SKRTSPARM (S3
463 Table). We therefore have, as yet no evidence of any impact or selective advantage to the
464 virus at the protein level of a change at position 203/204 from the RG to KR residues.
465

466 **SARS-CoV-2 and Host Adaptation: Implications for global viral dynamics,**
467 **pathogenesis and immunogenicity**

468 Currently the possible functional effect(s) of the introduction of the AAACGA motif from the
469 leader TRS into the RNA encoding position 203 and 204 of the nucleocapsid at the RNA and
470 protein level are not known. TRS sites are usually intergenic and it has been assumed that
471 recombination events at such sites are more likely to be viable. It has also been shown
472 recently that recombination breakpoint hotspots in coronaviruses are more frequently co-
473 located with TRS-B sites than expected (32). Our findings suggest that a novel TRS-B site
474 can be introduced in a recombination breakpoint from the leader TRS, and that this can occur
475 within an ORF and remain viable. The exact mechanism by which the AAA CGA codons
476 could have been incorporated from the TRS-L into the nucleocapsid is not known but may
477 have first required the AAACGA to be captured from the TRS-L and then for replication to
478 be reinitiated at the nucleocapsid to generate a full-length genomic RNA.
479

480 The nucleocapsid protein is a key structural protein critical to viral transcription and
481 assembly (33), suggesting that changes in this protein could either increase or decrease
482 replicative fitness. The K203/R204 polymorphism is located between the RNA
483 binding/serine-rich domains and the dimerization structural domain (S5 Fig) in a part of the
484 protein that has not been characterized in terms of 3-dimensional structure. The sequence of
485 this region is not similar enough to solved structures to allow prediction of the influence of
486 the K203/R204 polymorphisms on the structure or function of the protein. However, it is
487 known that SARS-CoV-2 is exquisitely sensitive to interferons and that it depends on the
488 nucleocapsid and M proteins to maintain interferon antagonism (34, 35). Specifically the C
489 terminus (aa 362 to 422) of the nucleocapsid, which is predicted to be expressed at higher
490 levels in those with the KR variant and novel sgRNA, has been shown to interact with the
491 SPRY domain of TRIM25 disturbing its interaction with CARDs of RIG-I inhibiting RIG-I
492 ubiquitination and Type 1 interferon signaling (36). Importantly the cells expressing the C-
493 terminal nucleocapsid protein in that study produced lower viral titer, suggesting the
494 incorporation of this protein into the nucleocapsid may reduce the formation of functional
495 virus. This raises the possibility that any enhancement of inhibition of interferon signaling
496 associated with the novel K203/R204 sgRNA may be offset by less efficient replication,
497 potentially accounting for the lack of association with higher viral load in the upper
498 respiratory tract and absence of epidemiologic evidence of increased transmission. It is also
499 possible that the increase in sgRNA directly inhibits RIG-I signaling and downstream Type I
500 interferon responses as has been described for Dengue serotype 2 (37). Finally, the central
501 region of coronavirus nucleocapsid (aa 117 to 268) has been shown to have RNA chaperone
502 activity that enhances template switching and efficient transcription possibly accounting for
503 the increase in sgRNA for the E and M proteins and ORF6 in KR-sequences compared to

504 RG-sequences (38). Note we cannot exclude that the novel sgRNA may also use the
505 downstream ATG in the ORF9c reading frame.
506
507 The adaptive potential of differential expression of sgRNAs is supported by a recent study by
508 Thorne and colleagues that demonstrates that the B.1.1.7 ('Alpha' or UK variant) isolate
509 containing the R203K/G204R substitutions is associated with enhanced antagonism of the
510 innate immune response (39). Specifically, this study showed that in-vitro infection of human
511 lung epithelial (Calu-3) cells by B.1.1.7 isolates showed diminished RNA and protein
512 expression of IFNβ and reduced induction of interferon sensitive genes relative to other
513 isolates without these defining mutations in the nucleocapsid (normalized for intracellular
514 viral RNA). This effect was independent of the reduced sensitivity to type I and III IFNs
515 described for isolates carrying the D614G spike mutation (40). Further evaluation of this
516 system showed that infection with the B.1.1.7 isolate resulted in significant changes in
517 protein expression of known innate immune regulators such as ORF9b (41), ORF6 (42) and
518 nucleocapsid (36, 43), as well as increased levels of the N* sgRNA described in this study
519 and was again confirmed to be unique to those isolates with the R203K/G204R mutations.
520 These increased levels of sgRNAs and protein support the findings in this study showing
521 increased sgRNA levels for N, ORF6 and N* in clinical samples from B.1.1.7-infected
522 subjects relative to subjects infected with other SARS-CoV-2 isolates. Interestingly, the
523 increased levels of ORF9b may be due to the D3L mutation in the nucleocapsid that we have
524 proposed to have arisen similarly to the R203G/G204R mutations and is associated with
525 increased levels of B.1.1.7 sgRNA encoding ORF9b in clinical samples (44).
526

527 The B.1.617.2 ('Delta') variant appears to be more transmissible even in the context of
528 previous vaccination and is now replacing other variants. This variant has acquired an
529 R203M substitution as a result of a single nucleotide change while retaining an arginine (G)
530 at position 204. This raises the possibility that the residue 203 is critical to the interaction of
531 nucleocapsid with TRIM25 decreasing the Type 1 interferon response or increases
532 transmissibility in some other way (36).

533

534 Other contemporary concerns include the fall in antibody levels following infection or
535 vaccination, the potential limited durability of protection afforded by currently available
536 vaccines and the risk of reinfection by variants after vaccination (45, 46). At low levels of
537 antibodies, the lungs appear to remain relatively protected against severe disease presumably
538 by some combination of antibodies and amnestic responses restimulated by the time the lung
539 is involved. In contrast, the early establishment of infection in the upper respiratory tract
540 appears possible if antibody levels are low (47). We therefore postulate that variants that are
541 more effective in interfering with Type 1 interferon responses would be more transmissible,
542 but not necessarily cause severe disease in the context of waning immunity at an individual or
543 population level.

544

545 **Conclusion**

546 Marked viral diversity and adaptation of other RNA viruses such as HIV, HCV and influenza
547 to host selective pressures have been a barrier to successful treatment and vaccination to date.
548 Although SARS-CoV-2 is less diverse and adaptable, the D614G variant and the K203/R204
549 and Delta variants have emerged by either nucleotide mutation or homologous recombination
550 during its rapid, widespread global spread and do appear to have functional impact. It will

551 therefore be critical to continue molecular surveillance of the virus and elucidate the
552 functional consequences of any newly emerging viral genetic changes to guide development
553 of diagnostics, antivirals and universal vaccines and to target conserved and potentially less
554 mutable SARS-CoV-2 elements. The ability of SARS-CoV-2 to introduce new TRS motifs
555 throughout its genome with the potential to introduce both novel sub-genomic RNA
556 transcripts and coding changes in its proteins may add to these challenges.
557
558
559
560

561 **Acknowledgments:** We thank colleagues at the Institute for Immunology and Infectious
562 Diseases, Murdoch University, Australia and the Department of Medicine, Division of
563 Infectious Diseases, Vanderbilt University Medical Center, USA. We would like to
564 acknowledge additional members of the Sheffield COVID-19 Genomic Group who
565 contributed to the generation of the sequence data: Adrienne Angyal, Rebecca L. Brown,
566 Laura Carrilero, Cariad M Evans, Luke R. Green, Danielle C. Groves, Katie J Johnson, Paul J
567 Parsons, David Partridge, Mohammad Raza, Rachel M. Tucker, Dennis Wang, Matthew D.
568 Wyles. Ethics approval and consent to participate (COG-UK CONSORTIUM; R&D
569 NR0195).
570 We thank the reviewers and editor for their helpful comments.
571
572 **Funding Sources:** SG, SL and EA were supported by a grant awarded by the National
573 Health and Medical Research Council (NHMRC; APP1148284). SM was supported by a
574 National Institutes of Health (NIH)-funded Tennessee Center for AIDS Research (P30
575 AI110527). MDP was funded by the NIHR Sheffield Biomedical Research Centre (BRC - IS-
576 BRC-1215-20017). Sequencing of SARS-CoV-2 samples was undertaken by the Sheffield
577 COVID-19 Genomics Group as part of the COG-UK CONSORTIUM. COG-UK and
578 supported by funding from the Medical Research Council (MRC) part of UK Research &
579 Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research
580 Limited, operating as the Wellcome Sanger Institute. TIdS is supported by a Wellcome Trust
581 Intermediate Clinical Fellowship (110058/Z/15/Z).
582
583 **Conflicts of interests:** The authors declare that they have no conflicts of interests.

584 References:

585

586 1. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with
587 the COVID-19 Outbreak. Curr Biol. 2020;30(7):1346-51 e2. Epub 2020/03/21. doi:
588 10.1016/j.cub.2020.03.022. PubMed PMID: 32197085; PMCID: PMC7156161.

589 2. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner
590 N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM,
591 Freeman TM, de Silva TI, Sheffield C-GG, McDanal C, Perez LG, Tang H, Moon-Walker A,
592 Whelan SP, LaBranche CC, Saphire EO, Montefiori DC. Tracking Changes in SARS-CoV-2
593 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020. Epub
594 2020/07/23. doi: 10.1016/j.cell.2020.06.043. PubMed PMID: 32697968; PMCID:
595 PMC7332439.

596 3. Grubaugh ND, Hanage WP, Rasmussen AL. Making Sense of Mutation: What
597 D614G Means for the COVID-19 Pandemic Remains Unclear. Cell. 2020. Epub 2020/07/23.
598 doi: 10.1016/j.cell.2020.06.040. PubMed PMID: 32697970; PMCID: PMC7332445.

599 4. Yurkovetskiy L, Pascal KE, Tompkins-Tinch C, Nyalile T, Wang Y, Baum A, Diehl
600 WE, Dauphin A, Carbone C, Veinotte K, Egri SB, Schaffner SF, Lemieux JE, Munro J,
601 Sabeti PC, Kyratsous C, Shen K, Luban J. SARS-CoV-2 Spike protein variant D614G
602 increases infectivity and retains sensitivity to antibodies that target the receptor binding
603 domain. bioRxiv. 2020. Epub 2020/07/09. doi: 10.1101/2020.07.04.187757. PubMed PMID:
604 32637944; PMCID: PMC7337374.

605 5. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, Farzan M, Choe H.
606 The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases
607 infectivity. bioRxiv. 2020. Epub 2020/06/27. doi: 10.1101/2020.06.12.148726. PubMed
608 PMID: 32587973; PMCID: PMC7310631.

609 6. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike:
610 mechanisms of coronavirus cross-species transmission. Journal of virology. 2010;84(7):3134-
611 46. Epub 2009/11/13. doi: 10.1128/JVI.01394-09. PubMed PMID: 19906932; PMCID:
612 2838128.
613 7. Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly
614 identified coronavirus 2019-nCoV. Journal of medical virology. 2020;92(4):433-40. Epub
615 2020/01/23. doi: 10.1002/jmv.25682. PubMed PMID: 31967321.
616 8. Sidney J, Southwood S, Moore C, Oseroff C, Pinilla C, Grey HM, Sette A.
617 Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture.
618 Curr Protoc Immunol. 2013;Chapter 18:Unit 18 3. Epub 2013/02/09. doi:
619 10.1002/0471142735.im1803s100. PubMed PMID: 23392640; PMCID: PMC3626435.
620 9. Sidney J, del Guercio MF, Southwood S, Engelhard VH, Appella E, Rammensee HG,
621 Falk K, Rotzschke O, Takiguchi M, Kubo RT, et al. Several HLA alleles share overlapping
622 peptide specificities. Journal of immunology. 1995;154(1):247-59. Epub 1995/01/01.
623 PubMed PMID: 7527812.
624 10. Parker MD, Lindsey BB, Leary S, Gaudieri S, Chopra A, Wyles M, Angyal A, Green
625 LR, Parsons P, Tucker RM, Brown R, Groves D, Johnson K, Carrilero L, Heffer J, Partridge
626 DG, Evans C, Raza M, Keeley AJ, Smith N, Filipe ADS, Shepherd JG, Davis C, Bennett S,
627 Sreenu VB, Kohl A, Aranday-Cortes E, Tong L, Nichols J, Thomson EC, Wang D, Mallal S,
628 De Silva TI. Subgenomic RNA identification in SARS-CoV-2 genomic sequencing data.
629 Genome Research. 2021;31(4):645-58. doi: 10.1101/gr.268110.120.
630 11. Jenjaroenpun P, Wanchai V, Ono-Moore KD, Laudadio J, James LP, Adams SH,
631 Prior F, Nookaew I, Ussery DW, Wongsurawat T. Two SARS-CoV-2 Genome Sequences of
632 Isolates from Rural U.S. Patients Harboring the D614G Mutation, Obtained Using Nanopore

633 Sequencing. Microbiol Resour Announc. 2020;10(1). Epub 2020/12/19. doi:
634 10.1128/MRA.01109-20. PubMed PMID: 33334896.
635 12. Franco-Muñoz C, Álvarez-Díaz DA, Laiton-Donato K, Wiesner M, Escandón P,
636 Usme-Ciro JA, Franco-Sierra ND, Flórez-Sánchez AC, Gómez-Rangel S, Rodríguez-
637 Calderon LD, Barbosa-Ramirez J, Ospitia-Baez E, Walteros DM, Ospina-Martinez ML,
638 Mercado-Reyes M. Substitutions in Spike and Nucleocapsid proteins of SARS-CoV-2
639 circulating in South America. Infect Genet Evol. 2020;85:104557-. Epub 2020/09/17. doi:
640 10.1016/j.meegid.2020.104557. PubMed PMID: 32950697.
641 13. Leslie A, Kavanagh D, Honeyborne I, Pfafferott K, Edwards C, Pillay T, Hilton L,
642 Thobakgale C, Ramduth D, Draenert R, Le Gall S, Luzzi G, Edwards A, Brander C, Sewell
643 AK, Moore S, Mullins J, Moore C, Mallal S, Bhardwaj N, Yusim K, Phillips R, Klenerman
644 P, Korber B, Kiepiela P, Walker B, Goulder P. Transmission and accumulation of CTL
645 escape variants drive negative associations between HIV polymorphisms and HLA. J Exp
646 Med. 2005;201(6):891-902. Epub 2005/03/23. doi: 10.1084/jem.20041455. PubMed PMID:
647 15781581; PMCID: 2213090.
648 14. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, Tang Y,
649 Holmes EC, Allen T, Prado JG, Altfeld M, Brander C, Dixon C, Ramduth D, Jeena P,
650 Thomas SA, St John A, Roach TA, Kupfer B, Luzzi G, Edwards A, Taylor G, Lyall H,
651 Tudor-Williams G, Novelli V, Martinez-Picado J, Kiepiela P, Walker BD, Goulder PJ. HIV
652 evolution: CTL escape mutation and reversion after transmission. Nature medicine.
653 2004;10(3):282-9. Epub 2004/02/11. doi: 10.1038/nm992. PubMed PMID: 14770175.
654 15. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of
655 HIV-1 adaptation to HLA-restricted immune responses at a population level. Science.
656 2002;296(5572):1439-43. Epub 2002/05/25. doi: 10.1126/science.1069660. PubMed PMID:
657 12029127.

658 16. Fitzmaurice K, Petrovic D, Ramamurthy N, Simmons R, Merani S, Gaudieri S, Sims
659 S, Dempsey E, Freitas E, Lea S, McKiernan S, Norris S, Long A, Kelleher D, Klenerman P.
660 Molecular footprints reveal the impact of the protective HLA-A*03 allele in hepatitis C virus
661 infection. Gut. 2011;60(11):1563-71. Epub 2011/05/10. doi: 10.1136/gut.2010.228403.
662 PubMed PMID: 21551190; PMCID: 3184218.
663 17. Rubnitz J, Subramani S. The minimum amount of homology required for homologous
664 recombination in mammalian cells. Molecular and cellular biology. 1984;4(11):2253-8. Epub
665 1984/11/01. doi: 10.1128/mcb.4.11.2253. PubMed PMID: 6096689; PMCID: 369052.
666 18. Sola I, Moreno JL, Zuniga S, Alonso S, Enjuanes L. Role of nucleotides immediately
667 flanking the transcription-regulating sequence core in coronavirus subgenomic mRNA
668 synthesis. Journal of virology. 2005;79(4):2506-16. Epub 2005/02/01. doi:
669 10.1128/JVI.79.4.2506-2516.2005. PubMed PMID: 15681451; PMCID: 546574.
670 19. Parker MD, Lindsey BB, Leary S, Gaudieri S, Chopra A, Wyles M, Angyal A, Green
671 LR, Parsons P, Tucker RM, Brown R, Groves D, Johnson K, Carrilero L, Heffer J, Partridge
672 DG, Evans C, Raza M, Keeley AJ, Smith N, Filipe ADS, Shepherd JG, Davis C, Bennett S,
673 Sreenu VB, Kohl A, Aranday-Cortes E, Tong L, Nichols J, Thomson EC, Consortium C-GU,
674 Wang D, Mallal S, de Silva TI. Subgenomic RNA identification in SARS-CoV-2 genomic
675 sequencing data. Genome Res. 2021;31(4):645-58. Epub 2021/03/17. doi:
676 10.1101/gr.268110.120. PubMed PMID: 33722935; PMCID: PMC8015849.
677 20. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-
678 CoV-2 Transcriptome. Cell. 2020;181(4):914-21 e10. Epub 2020/04/25. doi:
679 10.1016/j.cell.2020.04.011. PubMed PMID: 32330414; PMCID: PMC7179501.
680 21. Chandrashekar A, Liu J, Martinot AJ, McMahan K, Mercado NB, Peter L, Tostanoski
681 LH, Yu J, Maliga Z, Nekorchuk M, Busman-Sahay K, Terry M, Wrijil LM, Ducat S,
682 Martinez DR, Atyeo C, Fischinger S, Burke JS, Slein MD, Pessaint L, Van Ry A,

683 Greenhouse J, Taylor T, Blade K, Cook A, Finneyfrock B, Brown R, Teow E, Velasco J,
684 Zahn R, Wegmann F, Abbink P, Bondzie EA, Dagotto G, Gebre MS, He X, Jacob-Dolan C,
685 Kordana N, Li Z, Lifton MA, Mahrokhian SH, Maxfield LF, Nityanandam R, Nkolola JP,
686 Schmidt AG, Miller AD, Baric RS, Alter G, Sorger PK, Estes JD, Andersen H, Lewis MG,
687 Barouch DH. SARS-CoV-2 infection protects against rechallenge in rhesus macaques.
688 Science. 2020;369(6505):812-7. doi: 10.1126/science.abc4776.
689 22. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, Niemeyer
690 D, Jones TC, Vollmar P, Rothe C, Hoelscher M, Bleicker T, Brünink S, Schneider J, Ehmann
691 R, Zwirglmaier K, Drosten C, Wendtner C. Virological assessment of hospitalized patients
692 with COVID-2019. Nature. 2020;581(7809):465-9. doi: 10.1038/s41586-020-2196-x.
693 23. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF,
694 Hofacker IL. ViennaRNA Package 2.0. Algorithms for molecular biology : AMB. 2011;6:26.
695 Epub 2011/11/26. doi: 10.1186/1748-7188-6-26. PubMed PMID: 22115189; PMCID:
696 3319429.
697 24. Laing C, Wen D, Wang JT, Schlick T. Predicting coaxial helical stacking in RNA
698 junctions. Nucleic acids research. 2012;40(2):487-98. Epub 2011/09/16. doi:
699 10.1093/nar/gkr629. PubMed PMID: 21917853; PMCID: 3258123.
700 25. de la Pena M, Dufour D, Gallego J. Three-way RNA junctions with remote tertiary
701 contacts: a recurrent and highly versatile fold. Rna. 2009;15(11):1949-64. Epub 2009/09/11.
702 doi: 10.1261/rna.1889509. PubMed PMID: 19741022; PMCID: 2764472.
703 26. Hua L, Song Y, Kim N, Laing C, Wang JT, Schlick T. CHSalign: A Web Server That
704 Builds upon Junction-Explorer and RNAJAG for Pairwise Alignment of RNA Secondary
705 Structures with Coaxial Helical Stacking. PloS one. 2016;11(1):e0147097. Epub 2016/01/21.
706 doi: 10.1371/journal.pone.0147097. PubMed PMID: 26789998; PMCID: 4720362.

707 27. Chapman EG, Moon SL, Wilusz J, Kieft JS. RNA structures that resist degradation by
708 Xrn1 produce a pathogenic Dengue virus RNA. eLife. 2014;3:e01892. Epub 2014/04/03. doi:
709 10.7554/eLife.01892. PubMed PMID: 24692447; PMCID: 3968743.
710 28. Gaudieri S, Rauch A, Park LP, Freitas E, Herrmann S, Jeffrey G, Cheng W, Pfafferott
711 K, Naidoo K, Chapman R, Battegay M, Weber R, Telenti A, Furrer H, James I, Lucas M,
712 Mallal SA. Evidence of viral adaptation to HLA class I-restricted immune pressure in chronic
713 hepatitis C virus infection. J Virol. 2006;80(22):11094-104. Epub 2006/10/31. doi:
714 10.1128/JVI.00912-06. PubMed PMID: 17071929; PMCID: 1642167.
715 29. Brumme ZL, Kinloch NN, Sanche S, Wong A, Martin E, Cobarrubias KD, Sandstrom
716 P, Levett PN, Harrigan PR, Joy JB. Extensive host immune adaptation in a concentrated
717 North American HIV epidemic. Aids. 2018;32(14):1927-38. Epub 2018/07/27. doi:
718 10.1097/QAD.0000000000001912. PubMed PMID: 30048246; PMCID: 6125742.
719 30. Katoh J, Kawana-Tachikawa A, Shimizu A, Zhu D, Han C, Nakamura H, Koga M,
720 Kikuchi T, Adachi E, Koibuchi T, Gao GF, Brumme ZL, Iwamoto A. Rapid HIV-1 Disease
721 Progression in Individuals Infected with a Virus Adapted to Its Host Population. PloS one.
722 2016;11(3):e0150397. Epub 2016/03/10. doi: 10.1371/journal.pone.0150397. PubMed
723 PMID: 26953793; PMCID: 4783116.
724 31. Peng H, Yang LT, Wang LY, Li J, Huang J, Lu ZQ, Koup RA, Bailer RT, Wu CY.
725 Long-lived memory T lymphocyte responses against SARS coronavirus nucleocapsid protein
726 in SARS-recovered patients. Virology. 2006;351(2):466-75. Epub 2006/05/13. doi:
727 10.1016/j.virol.2006.03.036. PubMed PMID: 16690096.
728 32. Yang Y, Yan W, Hall B, Jiang X. Characterizing transcriptional regulatory sequences
729 in coronaviruses and their role in recombination. bioRxiv. 2020. Epub 2020/06/27. doi:
730 10.1101/2020.06.21.163410. PubMed PMID: 32587968; PMCID: PMC7310624.

731 33. Sola I, Almazan F, Zuniga S, Enjuanes L. Continuous and Discontinuous RNA
732 Synthesis in Coronaviruses. Annual review of virology. 2015;2(1):265-88. Epub 2016/03/10.
733 doi: 10.1146/annurev-virology-100114-055218. PubMed PMID: 26958916; PMCID:
734 6025776.
735 34. Kopecky-Bromberg SA, Martinez-Sobrido L, Frieman M, Baric RA, Palese P. Severe
736 acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and
737 nucleocapsid proteins function as interferon antagonists. Journal of virology. 2007;81(2):548-
738 57. Epub 2006/11/17. doi: 10.1128/JVI.01782-06. PubMed PMID: 17108024; PMCID:
739 PMC1797484.
740 35. Lokugamage KG, Hage A, Schindewolf C, Rajsbaum R, Menachery VD. SARS-
741 CoV-2 is sensitive to type I interferon pretreatment. bioRxiv. 2020. Epub 2020/06/09. doi:
742 10.1101/2020.03.07.982264. PubMed PMID: 32511335; PMCID: PMC7239075.
743 36. Hu Y, Li W, Gao T, Cui Y, Jin Y, Li P, Ma Q, Liu X, Cao C. The Severe Acute
744 Respiratory Syndrome Coronavirus Nucleocapsid Inhibits Type I Interferon Production by
745 Interfering with TRIM25-Mediated RIG-I Ubiquitination. Journal of virology. 2017;91(8).
746 Epub 2017/02/06. doi: 10.1128/JVI.02143-16. PubMed PMID: 28148787; PMCID:
747 PMC5375661.
748 37. Manokaran G, Finol E, Wang C, Gunaratne J, Bahl J, Ong EZ, Tan HC, Sessions OM,
749 Ward AM, Gubler DJ, Harris E, Garcia-Blanco MA, Ooi EE. Dengue subgenomic RNA
750 binds TRIM25 to inhibit interferon expression for epidemiological fitness. Science.
751 2015;350(6257):217-21. Epub 2015/07/04. doi: 10.1126/science.aab3369. PubMed PMID:
752 26138103; PMCID: PMC4824004.
753 38. Zuniga S, Cruz JL, Sola I, Mateos-Gomez PA, Palacio L, Enjuanes L. Coronavirus
754 nucleocapsid protein facilitates template switching and is required for efficient transcription.

755 Journal of virology. 2010;84(4):2169-75. Epub 2009/12/04. doi: 10.1128/JVI.02011-09.
756 PubMed PMID: 19955314; PMCID: PMC2812394.
757 39. Thorne LG, Bouhaddou M, Reuschl A-K, Zuliani-Alvarez L, Polacco B, Pelin A,
758 Batra J, Whelan MVX, Ummadi M, Rojc A, Turner J, Obernier K, Braberg H, Soucheray M,
759 Richards A, Chen K-H, Harjai B, Memon D, Hosmillo M, Hiatt J, Jahun A, Goodfellow IG,
760 Fabius JM, Shokat K, Jura N, Verba K, Noursadeghi M, Beltrao P, Swaney DL, Garcia-
761 Sastre A, Jolly C, Towers GJ, Krogan NJ. Evolution of enhanced innate immune evasion by
762 the SARS-CoV-2 B.1.1.7 UK variant. bioRxiv. 2021.
763 40. Guo K, Barrett BS, Mickens KL, Hasenkrug KJ, Santiago ML. Interferon Resistance
764 of Emerging SARS-CoV-2 Variants. bioRxiv. 2021.
765 41. Jiang HW, Zhang HN, Meng QF, Xie J, Li Y, Chen H, Zheng YX, Wang XN, Qi H,
766 Zhang J, Wang PH, Han ZG, Tao SC. SARS-CoV-2 Orf9b suppresses type I interferon
767 responses by targeting TOM70. Cell Mol Immunol. 2020;17(9):998-1000. Epub 2020/07/31.
768 doi: 10.1038/s41423-020-0514-8. PubMed PMID: 32728199; PMCID: PMC7387808.
769 42. Miorin L, Kehrer T, Sanchez-Aparicio MT, Zhang K, Cohen P, Patel RS, Cupic A,
770 Makio T, Mei M, Moreno E, Danziger O, White KM, Rathnasinghe R, Uccellini M, Gao S,
771 Aydillo T, Mena I, Yin X, Martin-Sancho L, Krogan NJ, Chanda SK, Schotsaert M, Wozniak
772 RW, Ren Y, Rosenberg BR, Fontoura BMA, Garcia-Sastre A. SARS-CoV-2 Orf6 hijacks
773 Nup98 to block STAT nuclear import and antagonize interferon signaling. Proc Natl Acad
774 Sci U S A. 2020;117(45):28344-54. Epub 2020/10/25. doi: 10.1073/pnas.2016650117.
775 PubMed PMID: 33097660; PMCID: PMC7668094.
776 43. Oh SJ, Shin OS. SARS-CoV-2 Nucleocapsid Protein Targets RIG-I-Like Receptor
777 Pathways to Inhibit the Induction of Interferon Response. Cells. 2021;10(3). Epub
778 2021/04/04. doi: 10.3390/cells10030530. PubMed PMID: 33801464; PMCID: PMC7999926.

779 44. Parker MD, Lindsey BB, Shah DR, Hsu S, Keeley AJ, Partridge DG, Leary S, Cope
780 A, State A, Johnson K, Ali N, Raghei R, Heffer J, Smith N, Zhang P, Gallis M, Louka SF,
781 Whiteley M, Foulkes BH, Christou S, Wolverson P, Pohare M, Hansford SE, Green LR,
782 Evans C, Raza M, Wang D, Gaudieri S, Mallal S, , de Silva TI. Altered Subgenomic RNA
783 Expression in SARS-CoV-2 B.1.1.7 Infections. bioRxiv. 2021.
784 45. Earle KA, Ambrosino DM, Fiore-Gartland A, Goldblatt D, Gilbert PB, Siber GR,
785 Dull P, Plotkin SA. Evidence for antibody as a protective correlate for COVID-19 vaccines.
786 Vaccine. 2021;39(32):4423-8. doi: 10.1016/j.vaccine.2021.05.063.
787 46. Khoury DS, Cromer D, Reynaldi A, Schlub TE, Wheatley AK, Juno JA, Subbarao K,
788 Kent SJ, Triccas JA, Davenport MP. Neutralizing antibody levels are highly predictive of
789 immune protection from symptomatic SARS-CoV-2 infection. Nature medicine.
790 2021;27(7):1205-11. doi: 10.1038/s41591-021-01377-8.
791 47. Bergwerk M, Gonen T, Lustig Y, Amit S, Lipsitch M, Cohen C, Mandelboim M, Gal
792 Levin E, Rubin C, Indenbaum V, Tal I, Zavitan M, Zuckerman N, Bar-Chaim A, Kreiss Y,
793 Regev-Yochay G. Covid-19 Breakthrough Infections in Vaccinated Health Care Workers.
794 New England Journal of Medicine. 2021. doi: 10.1056/nejmoa2109072.

795

796

FIGURES

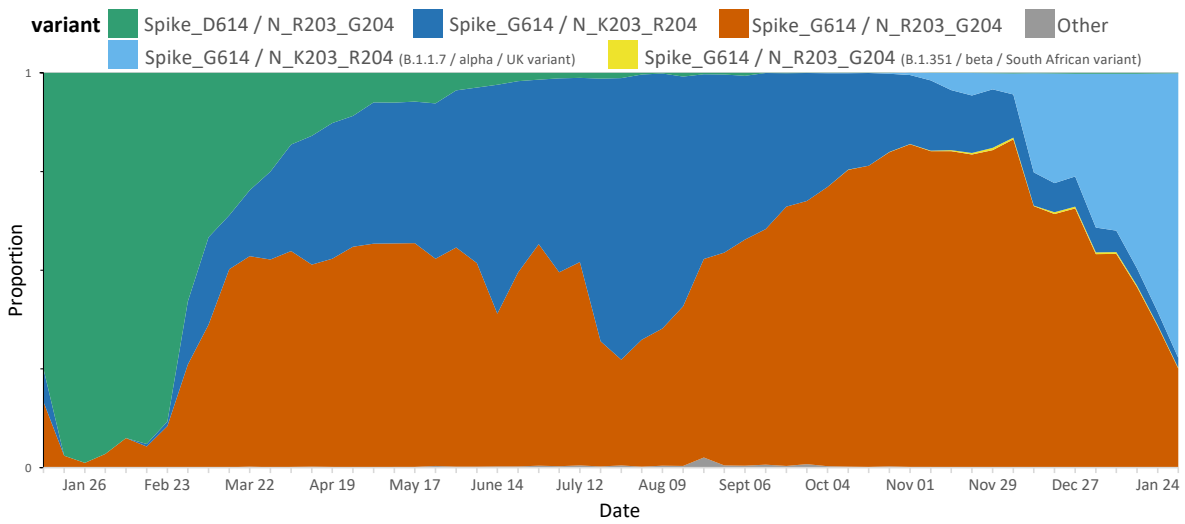

Fig 1. Proportion of weekly deposited SARS-CoV-2 sequences globally (n=455774). The D614G (B.1) variant has become one of the dominant forms globally. Note a small proportion of deposited sequences did not include information regarding specific collection date and as such were excluded.

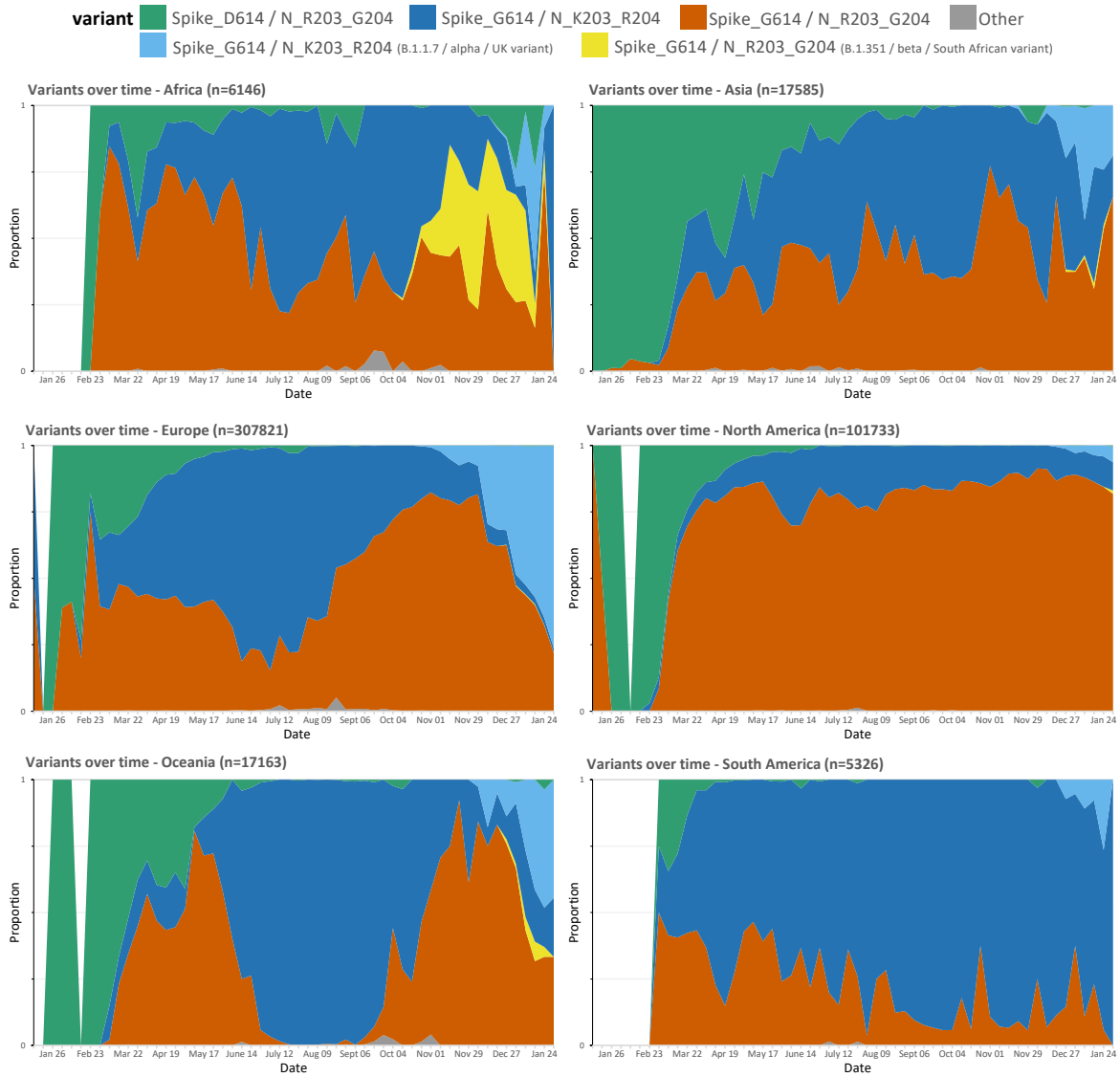

Fig 2. Proportion of weekly deposited SARS-CoV-2 sequences by region. The proportion of R203/G204 to K203/R204 sub-variants of the D614G variant differs in different regions with recent increases in the frequency of new variants.

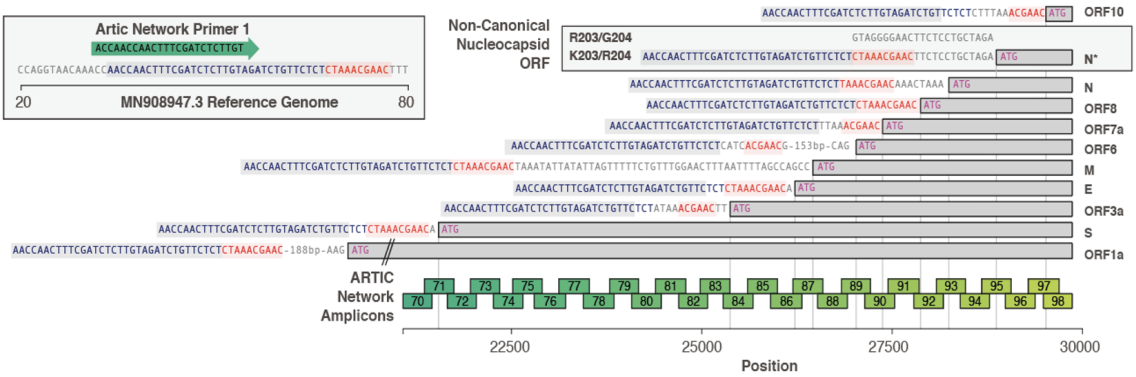

Fig 3. The configuration of canonical sgRNAs and the novel non-canonical nucleocapsid sgRNA (N*) in SARS-CoV-2. The bottom bar illustrates the presence of the leader sequence (blue text) followed by the transcription-regulating sequence (TRS; red text) within the genomic sequence that continues into the first ORF 1a. The presence of other canonical sgRNA transcripts in which the leader sequence and TRS precede the start codon (methionine; pink) of the other proteins are shown. The presence of the novel non-canonical sgRNA transcript containing the K203/R204 polymorphisms (N*) is shown. The ARTIC primer locations and resultant amplicons are shown.

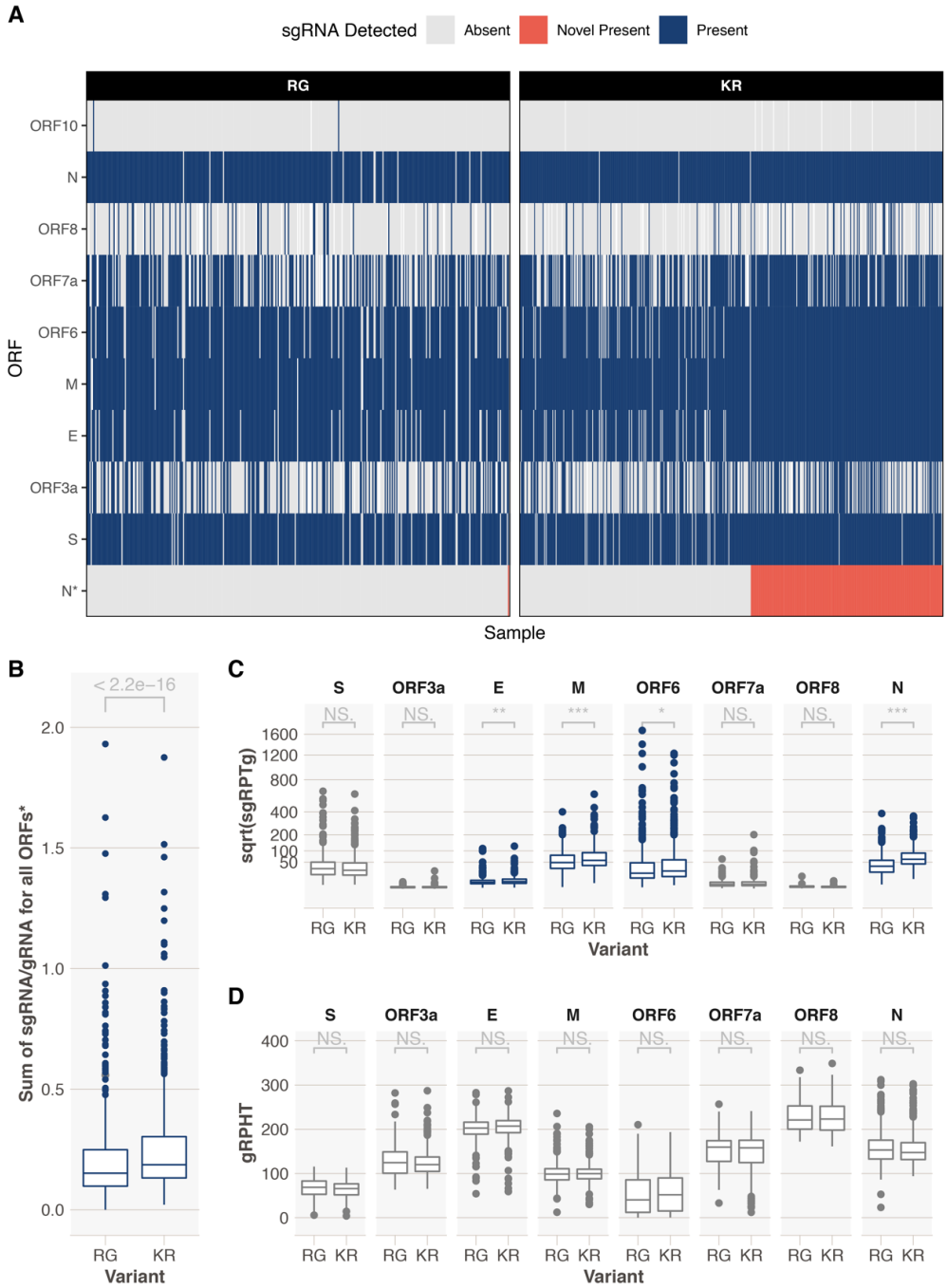

Fig 4. Exploration of sgRNAs in 981 samples from Sheffield, UK. A. A heatmap showing presence or absence of sgRNAs from different ORFs. K203/R204 (KR)-containing sequences have evidence of the novel truncated N ORF sgRNA (N*, red, 233/553, 42%). An ORF sgRNA was deemed present if we could find >=1 read in support. Heatmap is ordered by the presence or absence of the novel sgRNA. There were a total of 448 R203/G204 (RG)-containing sequences and 1 had evidence of a novel sgRNA (likely false positive, Fig S2). **B**. Significantly higher (Mann-Whitney U p < 2.2e-16) total sgRNA in KR-containing compared to RG-containing sequences. **C**. Sub-genomic RNA is significant increased in KR-containing

compared to RG-containing sequences for a number of ORFs, most notably nucleocapsid (N; Mann-Whitney U p = 2.06e-37 corrected for multiple testing using the Holm method). Y-axis denotes square root transformed sub-genomic reads normalized to 100,000 genomic reads from the same ARTIC amplicon. **D.** There is no difference in genomic RNA levels (normalized to total mapped reads) between KR- and RG-containing sequences. *novel sgRNA, ORF10 and ORF1a are excluded from this analysis due to ORF10 not being expressed, difficulty in discriminating ORF1a sgRNA from genomic RNA and the novel truncated N sgRNA is only being present in KR-containing sequences. ***** < 0.001, ** < 0.01, * < 0.05.** All p values shown are following correction for multiple testing with the Holm method.

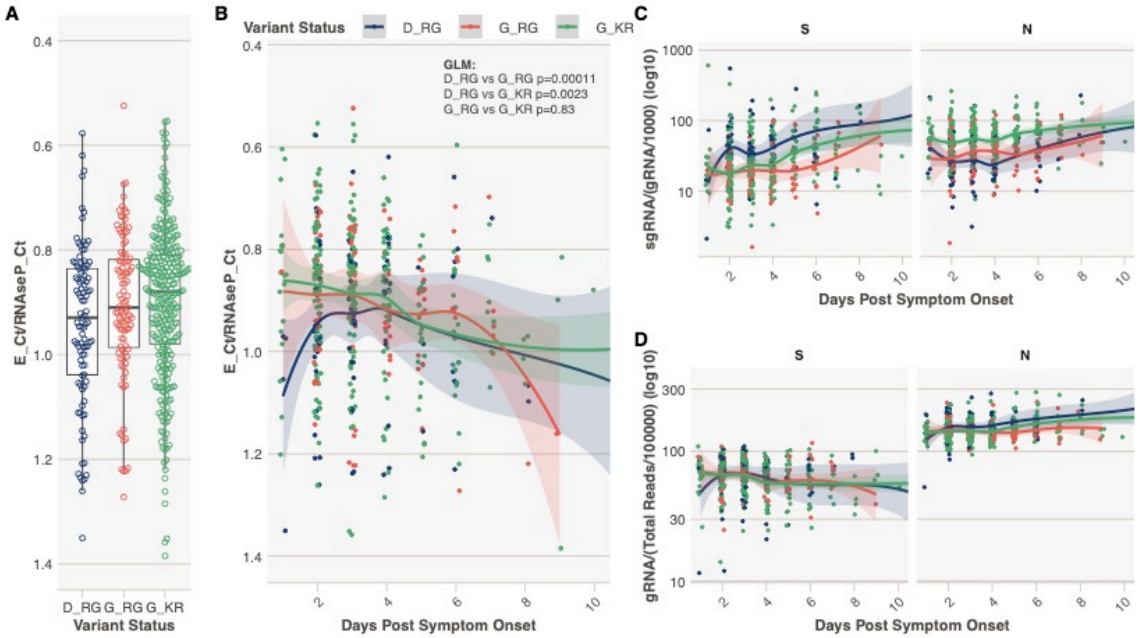

Fig 5. Spike 614 and Nucleocapsid 203/204 Status, Diagnostic Metrics and level of sub-genomic and genomic RNA. A. E gene cycle threshold (CT) normalized to RNAseP CT stratified by variant status in N = 478 individuals from Sheffield dataset with day of symptom onset data available. This normalization was done to combine and display E gene CT data from two different extraction protocols. Y-axis reversed to aid interpretation, as lower normalized CT values equal higher virus levels. **B.** Normalized E gene CT vs the day of sampling from day of symptom onset. P values provided are from a generalized multivariable linear regression model (GLM) for the difference in normalized E gene CT value between samples containing each variant, with extraction method and day of illness included in the model (Table S6) **C.** Normalized (per 1000 genomic reads) sgRNA levels for ORFs S and N. **D.** Normalized (per 100,000 mapped reads) genomic RNA levels for ORFs S and N.