Saudi Arabian SARS-CoV-2 genomes implicate a mutant Nucleocapsid protein in modulating host interactions and increased viral load in COVID-19 patients

Tobias Mourier[1,a], Muhammad Shuaib[1,a], Sharif Hala[2,3,a], Sara Mfarrej[1,a], Fadwa Alofi[4,b], Raeece Naeem[1,b], Afrah Alsomali[5,b], David Jorgensen[6,b], Amit Kumar Subudhi[1,b], Fathia Ben Rached[1,b], Qingtian Guan[1,b], Rahul P Salunke[1,b], Amanda Ooi[1,c], Luke Esau[1,c], Olga Douvropoulou[1,c], Raushan Nugmanova[1,c], Sadhasivam Perumal[1,d], Huoming Zhang[1,d], Issaac Rajan[1,d], Awad Al-Omari[7], Samer Salih[7], Abbas Shamsan[7], Abbas Al Mutair[7], Jumana Taha[8], Abdulaziz Alahmadi[9], Nashwa Khotani[10], Abdelrahman Alhamss[11], Ahmed Mahmoud[12], Khaled Alquthami[10], Abdullah Dageeg[13], Asim Khogeer[14], Anwar M. Hashem[15,16], Paula Moraga[17], Eric Volz[6], Naif Almontashiri[12], Arnab Pain[1,18,19,*]

[1]King Abdullah University of Science and Technology (KAUST), Pathogen Genomics Laboratory, Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia

[2]Infectious Disease Research Department, King Abdullah International Medical Research Centre, Ministry of National Guard Health Affairs, Jeddah, Saudi Arabia

[3]King Saud bin Abdulaziz University for Health Sciences, Ministry of National Guard Health Affairs, Jeddah, Saudi Arabia

[4]Infectious Diseases Department, King Fahad Hospital, Madinah, MOH, Saudi Arabia

[5]Infectious Diseases Department, King Abdullah Medical Complex, Jeddah, MOH, Saudi Arabia

[6]School of Public Health, Faculty of Medicine, Imperial College, Norfolk Place, St Mary's Campus, London, United Kingdom

[7]Dr. Suliman Al-Habib Medical Group, Riyadh, Saudi Arabia

[8]Department of Neuroscience, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

[9]Department of Preventive Medicine, Ministry of National Guard - Health Affairs, Riyadh, Saudi Arabia

[10]Infectious Diseases Medical Department, Al Noor Specialist Hospital Makkah, MOH, Saudi Arabia

[11]Gastroenterology Department, King Abdul Aziz Hospital Makkah, MOH, Saudi Arabia

[12]College of Applied Medical Sciences, Taibah University, Madinah, Saudi Arabia

[13]Department of Medicine, King Abdulaziz University Jeddah, Saudi Arabia

[14]Plan and Research Department, General Directorate of Health Affairs Makkah Region, MOH, Saudi Arabia

[15]Vaccines and Immunotherapy Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia

[16]Department of Medical Microbiology and Parasitology, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia

[17]King Abdullah University of Science and Technology (KAUST), Computer, Electrical and Mathematical Science and Engineering Division (CEMSE), Thuwal-Jeddah, 23955-6900, Saudi Arabia

[18]Research Center for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, N20 W10 Kita-ku, Sapporo, 001-0020 Japan

[19]Nuffield Division of Clinical Laboratory Sciences (NDCLS), The John Radcliffe Hospital, University of Oxford, Headington, Oxford, OX3 9DU, United Kingdom

[*]Correspondence: arnab.pain@kaust.edu.sa

Equal contributions:

[a]T.M., M.S., S.H., S.M. co-first authors contributed equally to this work

[b]F.A., R.N., D.J., A.S., A.S., F.B., Q.G., R.S. co-second authors contributed equally to this work

[c]A.O., L.E., O.D., R.N. co-third authors contributed equally to this work

[d]S.P., H.Z., I.R. co-fourth authors contributed equally to this work

Summary

Monitoring SARS-CoV-2 spread and evolution through genome sequencing is essential in handling the COVID-19 pandemic. The availability of patient hospital records is crucial for linking the genomic sequence information to virus function during the course of infections. Here, we sequenced 892 SARS-CoV-2 genomes collected from patients in Saudi Arabia from March to August 2020. From the assembled sequences, we estimate the SARS-CoV-2 effective population size and infection rate and outline the epidemiological dynamics of import and transmission events during this period in Saudi Arabia. We show that two consecutive mutations (R203K/G204R) in the SARS-CoV-2 nucleocapsid (N) protein are associated with higher viral loads in COVID-19 patients. Our comparative biochemical analysis reveals that the mutant N protein displays enhanced viral RNA binding and differential interaction with key host proteins. We found hyper-phosphorylation of the adjacent serine site (S206) in the mutant N protein by mass-spectrometry analysis. Furthermore, analysis of the host cell transcriptome suggests that

the mutant N protein results in dysregulated interferon response genes. We provide crucial information in linking the R203K/G204R mutations in the N protein as a major modulator of host-virus interactions and increased viral load and underline the potential of the nucleocapsid protein as a drug target during infection.

Main

The emergence of novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes the respiratory coronavirus infectious disease 2019 (COVID-19), resulted in a pandemic that has triggered an unparalleled public health emergency [1,2]. The global spread of SARS-CoV-2 depended fundamentally on human mobility patterns. This is highly pertinent to a country like the Kingdom of Saudi Arabia, which as of 22nd February 2021 had a total of 374,691 cases and 6,457 deaths [3]. The kingdom frequently experiences major population movements, particularly religious mass gatherings. For instance, during Umrah and Hajj roughly 9.5 million pilgrims visit two Islamic sites in Makkah and Madinah annually [4,5] and the Ministry of Health takes extraordinary public health measures to keep the pilgrims safe and major outbreaks have been by and large avoided in recent years. Further, an estimated 5 million Shiite Saudi nationals travel to Iran for pilgrimage, which became an early source of COVID-19 infections in the region [5,6]. This movement has been reflected in the early phase of COVID-19 transmission within Saudi, as the first case was officially reported in Qatif (Eastern Region) on March 2nd, 2020 [7].

Genomic epidemiology of emerging viruses has proven to be a useful tool for outbreak investigation and tracking the pathogen's progress [8,9]. Currently, over half a million complete and high coverage genomes are accessible on GISAID[10,11], which aids immensely in tracking the

viral sequences globally [12]. Novel SARS-CoV-2 variants are continuously arising and besides providing signals for epidemiological tracking, a subset of the resulting variants will have a functional impact on transmission and infection [13-15]. It is therefore critical to monitor the genetic viral diversity throughout the pandemic.

In this study, we sequenced 892 SARS-CoV-2 genomes from nasopharyngeal swab samples of patients from the four main cities, Jeddah, Makkah, Madinah, and Riyadh, as well as a small number of patients from the Eastern region of Saudi Arabia (Figure 1, Table S1, Table S2). We analyzed the genomes to investigate the nucleotide changes and multiple mutation events that represent the first 6 months of the locally circulating pandemic lineages of the SARS-CoV-2 in Saudi Arabia and searched for potential association of polymorphic sites in the genome with available hospital records including severe disease and case fatality rates among the COVID-19 patients. We performed phylogenetic analysis to visualize the genetic diversity of SARS-CoV-2 and the nature of transmission lineages during March-August, 2020. We have presented a snapshot of the genomic variation landscapes of the SARS-CoV-2 lineages in our study population and linked specific set of mutation events in the N gene to viral loads in a diverse population of COVID-19 patients in Saudi Arabia (Figure S1). Finally, we experimentally show the functional impact of these mutations in the N protein on the virus' interactions with the host.

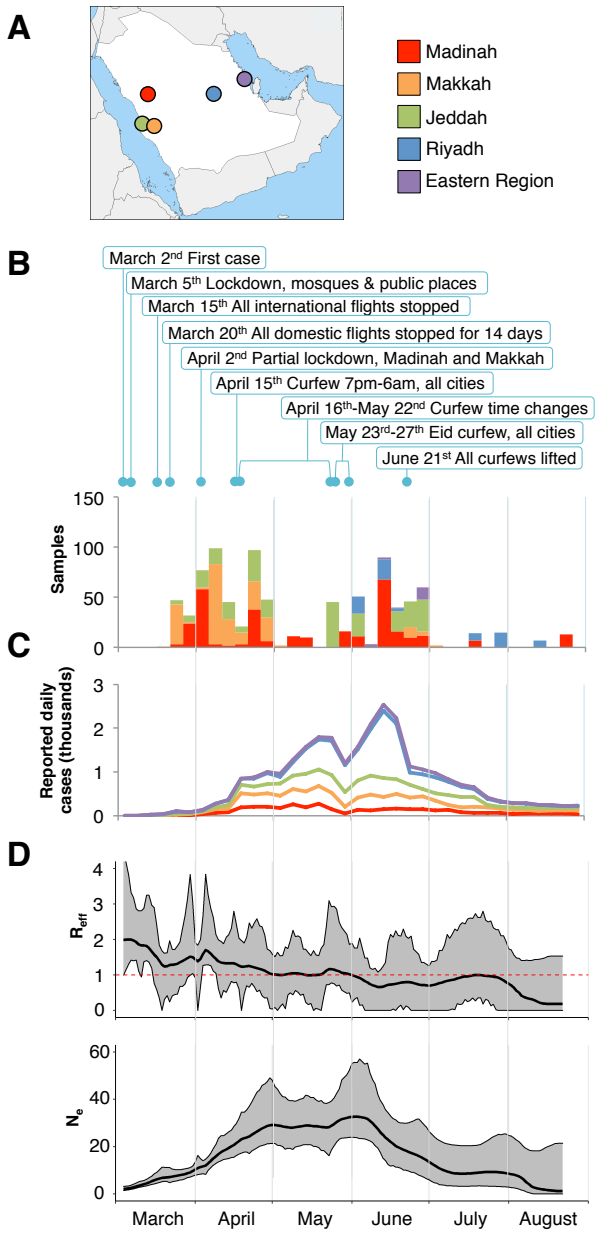

Figure 1. Sample overview and population genetics

A) Locations of the sampling cities within Saudi Arabia. B) Combined numbers of samples retrieved from the 4 cities and the Eastern region during the first six months of the pandemic. Cities are colored as in panel A. Months are shown at the bottom of the figure, and each month is divided into 5-days intervals. New daily cases for the city of Khobar is shown on the Eastern Region plot. Major restrictions imposed by the Ministry of Health and by Royal decrees are indicated above plots. C) Combined average numbers of new daily cases (Supplementary Information). D) Estimate of effective reproduction number [R_{eff}] over time in Saudi Arabia (top) and the estimate of effective population size [Ne], the relative population size required to produce the diversity seen in the sample (bottom). The red horizontal red line represents an R of 1, the level required to sustain epidemic growth. Grey confidence areas denote the 95% credible intervals.

Results

SNP calling and phylodynamics of SARS-CoV-2 samples from Saudi Arabia

We sequenced and assembled SARS-CoV-2 genomes from 892 patient samples. This group includes 144 patients that were placed in quarantine and had either mild symptoms or were asymptomatic. The remaining patients were all hospitalized (Table S2). Data on comorbidities were available for 689 patients, and included diabetes, hypertension, Lung BA, Kidney CDK, Cardiovascular, and Cerebrovascular diseases. Patient outcome data was available for 850 samples, and 199 patients (23%) died during hospitalization.

From the 892 assembled viral genomes collected over a period of 6 months, we found a total of 836 single-nucleotide polymorphisms (SNPs) compared to the Wuhan SARS-CoV-2 reference (GenBank accession: NC_045512) (Figure S2). The observed numbers of SNPs relative to the Wuhan reference follow the numbers observed in global samples (Figure S3). We further detected 41 indels of which 26 reside in coding regions (Table S3). Most indels were specific to a single sample, and no identical indel was found in more than four samples. Compared with global SNP data, seven SNPs were found in higher frequencies (absolute difference > 0.1) in samples from Saudi Arabia (Figure S2). These include the Spike protein D614G (A23403G) and three consecutive SNPs causing the R203K and G204R changes in the nucleoprotein (G28881A, G28882A, and G28883C). Together with all sequences from Saudi Arabia available on GISAID on December 31st 2020, the assembled sequences were used to construct the effective population size and growth rate estimates of SARS-CoV2 over the course of the first wave of the epidemic. The skygrowth model [16] (Figure 1D) shows a downward trend in the effective reproduction number (R) over time with the timely introduction and maintenance of effective non-pharmaceutical interventions by the Saudi Ministry of Health. Following the lifting of restrictions towards the end of June, the model estimates that R remained below or at 1 to the end of the period covered by the genetic data presented in this study. The effective population size (N_e) represents the relative diversity of the sequences collected in Saudi Arabia over the course of the outbreak (Figure 1D). The model predicts a peak in viral diversity at the beginning of June. This is ahead of the peak number of cases reported nationally and is likely influenced by the earlier peak in reported cases in the three cities, which contribute the most viral sequences to this analysis (Madinah, Makkah and Jeddah).

A maximum-likelihood phylogenetic analysis revealed that samples from Saudi Arabia represent 5 major Nextstrain clades[12], 19A-B and 20A-C (Figure 2A). This highlighted the clade 20A that all carried the Nucleocapsid (N) protein R203K/G204R mutations [17] with high incidences of ICU hospitalizations. These samples were predominantly coming from Jeddah. Through time-scaled phylogenies dates of importation events were then estimated for each clade. The majority of importations for all clades were inferred to have occurred early in the outbreak, primarily in March and early April (Figure 2B). Inferring importation events from a phylogenetic tree with estimated dating of nodes we see an early import from Asia followed by multiple imports from different continents (Figure 2C).

Origin of R203K/G204R SNPs and importation into Saudi Arabia

A dated phylogeny of global samples showed that samples with the R203K/G204R SNPs are predominantly found in Nextstrain clades 20A, 20B and 20C, and do not form a monophyletic group (Figure S4). Furthermore, a few samples are further found in the early appearing 19A and 19B clades. However, due to the limited number of mutations separating SARS-CoV-2 genomes constructing a reliable and robust phylogeny is problematic[18], and while different clades may be well supported, the exact relationship between clades is often less easily resolved. Although phylogenetic trees of SARS-CoV-2 genomes may appear to robustly reflect transmission events, collapsing branches with low support will typically result in extensive polytomies[19,20]. Additionally, the placement of individual virus genomes may be hampered by systematic errors, homoplasies, potential recombination, or co-infection of multiple virus strains[18,19,21-23]. It is therefore not clear if the phylogenetic distribution of samples with R203K/G204R SNPs reflects multiple independent origins of the SNPs, although it is evident that the R203K/G204R SNPs

appeared early in the pandemic spread (Figure S4). Consistent with this, we find the earliest estimated importation events of R203K/G204R SNPs in late January 2020, most likely from Italy (Figure S5). This thus suggest a slightly earlier importation date than the estimate of importation events of clade 20B (Figure 2B). Within our sampling window we observe an apparent transient increase in the frequency of R203K/G204R SNPs (Figure 3A) in accordance with earlier observations [17,24]. This peak is similarly observed in global data up until the fall of 2020, where the R203K/G204R SNPs once again increase along with the Spike protein Y501N mutation in the B1.1.17 lineage [25] (Figure 3A).

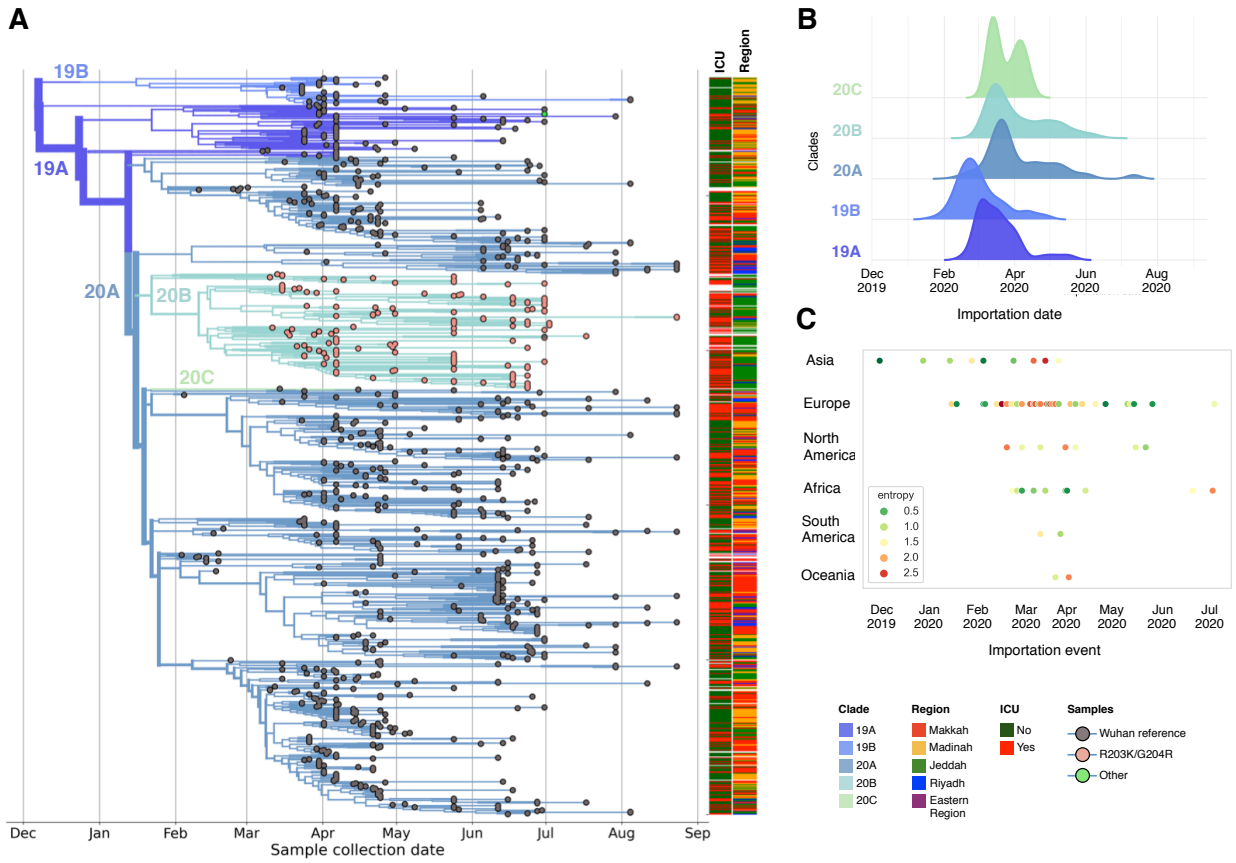

Figure 2. Phylodynamics of SARS-CoV-2 samples in Saudi Arabia

A) Global Time scaled phylogeny of 952 Saudi samples coloured by Nextstrain clades. Samples are shown as circles and coloured according to their genotype at genome positions 28,881-28,883. ICU status and sampling region are indicated on the right of the tree. B) Distributions of

importation dates for the 5 Nextstrain (nextstrain.org) clades found in Saudi Arabia coloured by clade. C) Importation events estimated by traversing a phylogenetic tree to identify branches that resulted in transitions into Saudi Arabia from another country. Events are coloured by their normalised Shannon entropy, which measures the uncertainty inherent in the country of origin for a given importation event.

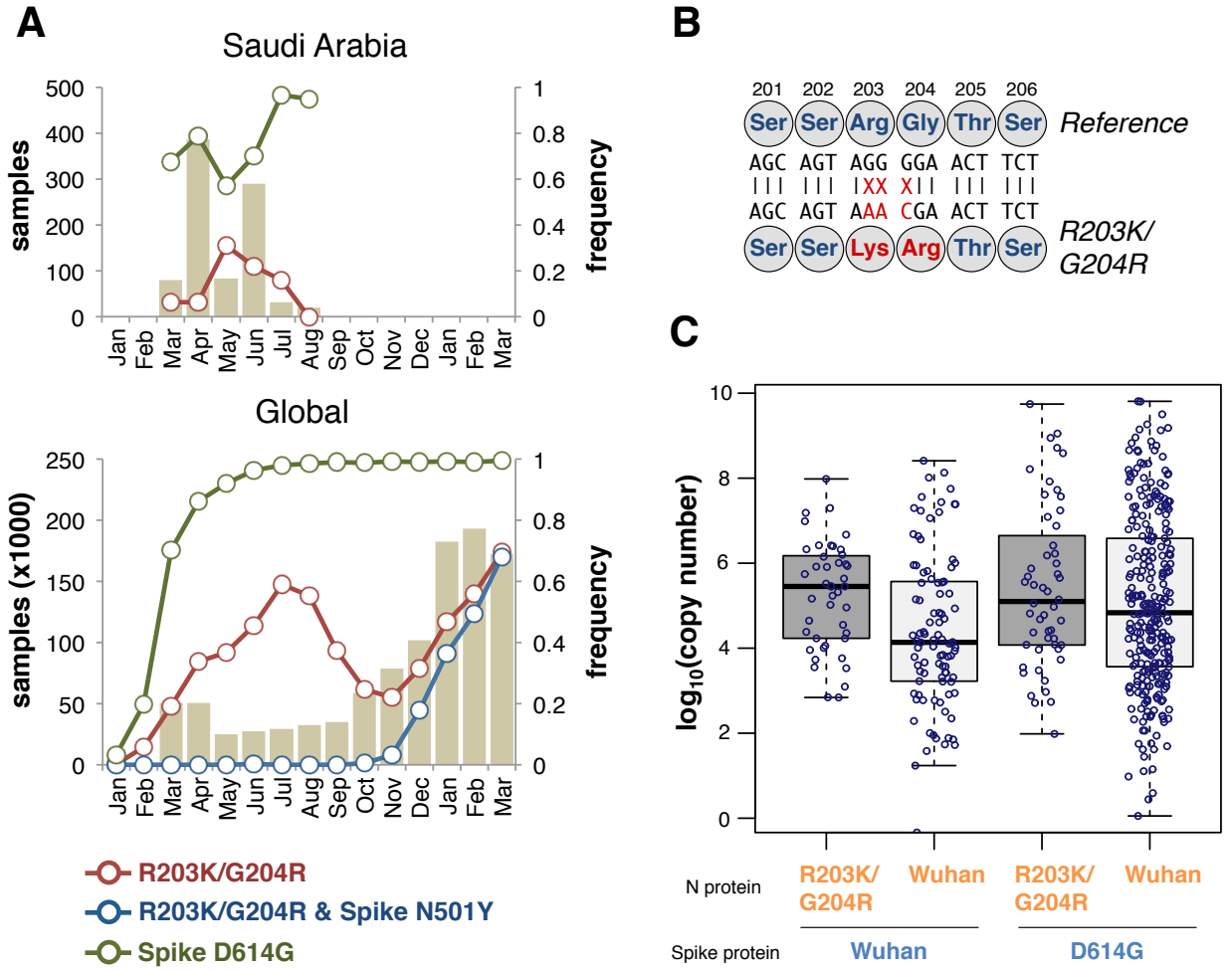

Figure 3. Higher viral loads in samples with R203K/G204R SNPs

A) Top: The numbers of samples from Saudi Arabia presented in this study are shown as bars by their sampling date (January 2020-March 2021). Bottom: Samples deposited in GISAID. On both plots, lines show the fraction of samples having the R203K/G204R SNPs (red line), having both the R203K/G204R SNPs and the Spike protein N501Y SNP (blue line), and having the Spike protein D614G SNP (green line). B) Overview of the three SNPs underlying the N protein R203K/G204R changes. Amino acid numbers in the N protein is shown above. C) Boxplot

showing the distribution of virus copy number derived from Ct measurements. Ct values from the N1 primer pairs were normalized by RNase P primer pair values and converted to copy numbers from a standard curve. Only samples processed using the TaqPath™ kit (Thermofisher) were included (see Supplementary Information). Copy numbers are shown for four different haplotypes (as indicated below the plot) corresponding to virus genome positions 28,881-28,883 (orange text) and 23,403 (blue text). 'Wuhan' denotes the genotypes in the reference genome (NC_045512).

A mutant form of the Nucleocapsid (N) protein associated with higher viral loads in COVID-19 patients in Saudi Arabia

A genome-wide association study between SARS-CoV-2 SNPs and patient mortality identified the three consecutive SNPs (G28881A, G28882A, G28883C) underlying the R203K/G204R mutations (Figure 3B, Figure S6). Of the 892 assembled genomes, 882 (98.9%) genomes either have the three reference alleles, GGG, or the three mutant alleles, AAC, at positions 28,881-28,883. This is similarly found in global samples deposited in GIASID in 2020, where 99.7% of samples with SNPs at positions 28,881-28,883 contain all three SNPs (Figure S7). In our samples, no other SNPs co-occur with the R203K/G204R SNPs (Figure S8). The frequency of the R203K/G204R SNPs is markedly higher in samples from Jeddah, where the observed frequency of 0.38 is more than 10-fold higher than the average of the other cities (Table S2). Within-host polymorphism has been observed for the R203K/G204R SNPs either resulting from co-infection of multiple strains or cross-sample contamination[21]. Co-infection of SARS-CoV-2 is demonstrated through observations of recombination between genetically distinct lineages[26]. To rule out cross-sample contamination, we investigated the levels of within-host polymorphisms in a range of SNP positions and found this more consistent with cases of co-infection among patients rather than contamination issues (Supplementary Information, Figure S9).

Using multivariable regression, we next evaluated the effect of the R203K/G204R SNPs on mortality, severity, and viral load in our COVID-19 patients samples for which limited amount

of clinical meta-datasets were available. Disease severity was defined as deceased patients and patients admitted to ICU. For mortality and severity, we first fitted a linear model using R203K/G204R SNPs as a covariate. Then we fitted adjusted models by including gender, age, comorbidities, hospital and time. Additionally, the Spike protein D614G SNP that is associated with higher viral load [13] was included. Age and time were included using smoothing splines to allow for potential non-linear relationships [27]. Using an unadjusted logistic regression, we observed a positive and statistically significant association between R203K/G204R SNPs and severity. Specifically, we found that the log-odds of severity increased by 1.16, 95% CI 0.70-1.64. In the adjusted model, the log-odds decreased to 0.66, 95% CI 0.01-1.32. That is, we found a borderline significant association between R203K/G204R SNPs and severity. The relationship between mortality and R203K/G204R SNPs was positive and statistically significant in the unadjusted model with log-odds equal to 1.39, 95% 0.99-1.79. We also found a positive and statistically significant relationship adjusting for D614G SNP, gender, age, comorbidities, and hospital (log-odds = 0.62, 95% 0.13-1.10). However, after adjusting for time as a variable, there was no longer any association between R203K/G204R SNPs and mortality (log-odds: 0.25, 95% CI -0.30-0.80). The models thus suggest a temporal component in our observations, and it is important to note that the recorded mortalities from Jeddah are concentrated on just a few dates (Figure S10). Unfortunately, our data set does not allow us to assess if the observed mortality rates are the result of shifts in treatment regimes or admission procedures during the sampling window.

We then tested if R203K/G204R SNPs were associated with higher viral copy numbers as indicated by the cycle threshold (Ct) values obtained through quantitative PCRs (see Methods). From the unadjusted regression we found a positive and statistically significant relationship

between R203K/G204R SNPs and $\log_{10}$(viral copy number), with the mean of $\log_{10}$(viral copy number) values increasing by 1.03 units (95% CI 0.67-1.46). The significance was still observed in the adjusted model, although the relationship decreased to 0.57 units (95% CI 0.13-1.01) (Figure 3C). Similarly, the adjusted model showed a significant relationship between D614G SNPs and $\log_{10}$(viral copy number), with the mean values increasing by 0.32 units (95% CI 0.13-1.01), consistent with earlier reports [13,28]. The positive and statistically significant association of R203K/G204R SNPs with higher viral load in critical COVID-19 patients suggests its functional implications during viral infection.

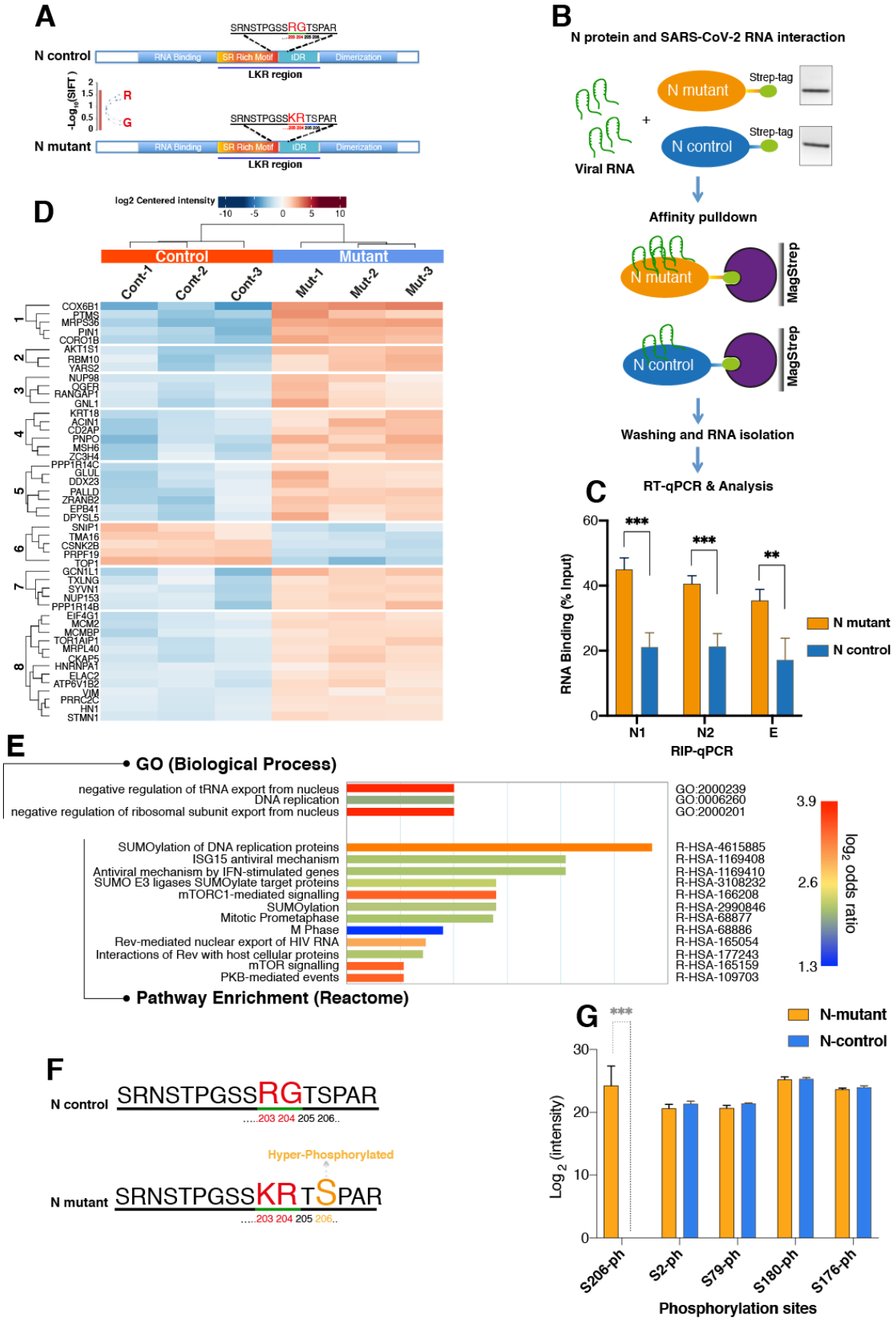
A
N control
SRNSTPGSSRGTSPAR
RNA Binding
SR Rich Motif
IDR
Dimerization
LKR region
-Log10(SIFT)
R
G
N mutant
SRNSTPGSSKRTSPAR
B
N protein and SARS-CoV-2 RNA interaction
Viral RNA
N mutant
N control
Strep-tag
Affinity pulldown
MagStrep
Washing and RNA isolation
RT-qPCR & Analysis
C
RNA Binding (% Input)
N1
N2
E
RIP-qPCR
N mutant
N control
D
log2 Centered intensity
Control
Mutant
Cont-1
Cont-2
Cont-3
Mut-1
Mut-2
Mut-3
E
GO (Biological Process)
negative regulation of tRNA export from nucleus
DNA replication
negative regulation of ribosomal subunit export from nucleus
SUMOylation of DNA replication proteins
ISG15 antiviral mechanism
Antiviral mechanism by IFN-stimulated genes
SUMO E3 ligases SUMOylate target proteins
mTORC1-mediated signalling
SUMOylation
Mitotic Prometaphase
M Phase
Rev-mediated nuclear export of HIV RNA
Interactions of Rev with host cellular proteins
mTOR signalling
PKB-mediated events
Pathway Enrichment (Reactome)
log2 odds ratio
F
N control
SRNSTPGSSRGTSPAR
Hyper-Phosphorylated
N mutant
SRNSTPGSSKRTSPAR
G
N-mutant
N-control
Log2 (intensity)
S206-ph
S2-ph
S79-ph
S180-ph
S176-ph
Phosphorylation sites

Figure 4. RNA binding and Affinity Purification Mass-Spectrometry (AP-MS) analysis of mutant and control SARS-CoV-2 N protein

A) A schematic diagram showing the SARS-CoV-2 N protein different domains (Upper: control, Lower: mutant) and highlighting the mutation site (R203K and G204R) and the linker region (LKR) containing a serine-arginine rich motif (SR-motif). The bar-plot (lower panel) indicates the SIFT [29] predicted deleteriousness score of substitution at position 204 from G to R. B) Sketch of In-vitro RNA immunoprecipitation (RIP) procedure used for analysis of viral RNA interaction with mutant and control N protein (See methods for details). Isolated RNAs were analyzed by RT-qPCR using specific primers for viral N gene (N1 and N2) and E gene. C) Bar chart shows level of viral RNA retrieval (% input) with mutant and control N protein (± SD from 4 experiments, [t-test, p value 0.00016 (***), 0.00019 (***), and 0.003 (**)]). D) Identification of host-interacting partners of mutant and control SARS-CoV-2 N protein by Affinity Mass-Spectrometry. Heatmap showing significantly differentially changed human proteins (3 replicates) interactome in mutant versus control N protein AP-MS analysis. E) Gene Ontology (GO)-enrichment analysis of significantly changed terms between mutant and control proteins in terms of biological process and pathway enrichment. The scale shows p-value adjusted Log2 of odds ration mutant-versus-control. F) Profiling of phosphorylation status of mutant and control N protein by Mass-Spectrometry. Sketch showing part of SR-rich motif of SARS-CoV-2 N protein containing the KR mutation site (R203K and G204R) (Lower). The hyper-phosphorylated serine 206 (as shown in G) in the mutant N protein near the KR mutation site is indicated in orange color. G) Phosphorylation status of mutant and control N protein was analyzed by mass spectrometry (3 biological replicates per affinity condition). Bar-plot shows the Log2 intensities of selected phosphorylated peptides in control and mutant condition. Serine 206 is hyper-phosphorylated in mutant N protein (± SD from 3 experiments, p-value 0.00017 (***) t-test).

<u>*N mutant protein has high oligomerization potential and RNA binding affinity*</u>

The SARS-CoV-2 N protein binds the viral RNA genome and is central to viral replication [30]. Protein structure predictions have shown that the R203K/G204R mutations result in significant changes in protein structure [24], theoretically destabilizing the N structure [31], and potentially enhancing the protein's ability to bind RNA and alter its response to serine phosphorylation events [32]. The R203K/G204R mutations in the SARS-CoV-2 N protein are within the linkage region (LKR) containing the serine/arginine-rich motif (SR-rich motif) (Figure 4A), known to be involved in the oligomerization of N proteins [33,34]. Protein cross-linking shows that N mutant protein (with the R203K/G204R mutations) has higher oligomerization potential compared to the

control N protein (without the changed amino acids) at low protein concentration (Figure S11A-B).

Given that the oligomerization of N protein acts as a platform for viral RNA interactions [35], we sought to examine the binding affinity of mutant and control N protein with viral RNA isolated from COVID-19 patient swabs. The RNA-binding activity of mutant and control N proteins was examined by pulled-down viral RNA through an *in vitro* RIP assay (Figure 4B), and our data revealed that the mutant N protein enriched significantly higher level of viral RNA compared to control protein (Figure 4C). This indicates a strong binding capability of mutant N proteins with viral RNA, which could potentially impact the essential roles of N protein at various stages of viral life cycle and its interaction with the host.

The R203K/G204R mutations in the N protein affect its interaction with host proteins

According to the SIFT tool [29], a substitution at position 204 from G to R in the N protein is predicted to affect functional properties (Figure 4A). Therefore, we decided to investigate how the two amino acids substitution (R203K and G204R) in the N protein impact its functional interaction with the host that could modulate viral pathogenesis and rewiring of host cell pathways and processes. HEK-293T cells (3 biological replicates, Supplementary Information) were used for affinity-purification followed by mass spectrometry analysis (AP-MS) to identify host proteins associated with the control and mutant N protein (Figure S12). The majority (87%) of non-differentially interacting proteins overlapped with the previously reported[36] N protein interacting partners (Figure S12D and Table S4). We identified 48 human proteins that displayed significant (adjusted p-value ≤ 0.05, and Log_2 fold change ≥ 1) differential interactions with the mutant and control N protein (Figure 4D, Figure S12E, Table S5). Among these, 43 proteins

showed increased interaction and 5 proteins showed decreased interaction with the N mutant (Figure 4D, Figure S12E). Among the group with increased interaction, we identified many proteins associated with TOR and other signaling pathways (such as AKT1S1 and PIN1), proteins associated with the viral process, viral transcription, and negative regulation of RNA nuclear export (NUP153 and NUP98), and proteins involved in apoptotic and cell death processes (PAWR, ACIN1, and PDCD5) (Figure 4D, Figure S12E). We also identified proteins in the mutant condition that are linked with the immune system processes (PTMS), kinase activity (GCN1), and translation (e.g. MRPS36) (Figure 4D, Figure S12E). In the group with decreased interaction, we identified SNIP1 (NF-kappaB signaling), TMA16 (translation), and CSNK2B (casein kinase II) (Figure 4D, Figure S12E). Gene ontology analysis showed that the most enriched biological processes are associated with negative regulation of tRNA and ribosomal subunit export from the nucleus (Figure 4E). This finding suggests that the mutant virus may more efficiently inhibit and hijack the host translation to facilitate viral replication and pathogenesis. Further, many viruses can manipulate the host sumoylation process to enhance viral survival and pathogenesis[37]. By pathway enrichment analysis of differentially interacting proteins, we identified pathways associated with the sumoylation of host proteins and antiviral mechanisms (Figure 4E).

Serine 206 (S206) displays hyper-phosphorylation in the mutant N protein

In SARS-CoV, it has been shown that phosphorylation of the N protein is more prevalent during viral transcription and replication [38] and inhibition of phosphorylation diminishes viral titer and cytopathogenic effects [39]. Recent elegant studies elaborated the role of N protein phosphorylation in modulating RNA binding and phase separation in SARS-CoV-2 [35,40-42]. Thus, phosphorylation

of N protein in the LKR region is critical for regulating both viral genome processing (transcription and replication) and nucleocapsid assembly [35,40]. To further understand the functional relevance of KR mutation in the N protein, we performed phosphoproteomic analysis in control and mutant conditions. We consistently found that the serine 206 (S206) site, which is next to the KR mutation site (Figure 4F), is highly phosphorylated, specifically in the mutant N protein (Figure 4G, Table S6). Notably, the phosphorylation level at serine 2 (S2) and other serine sites (S79, S176, and S180) within the LKR region did not change between mutant and control conditions (Figure 4F).

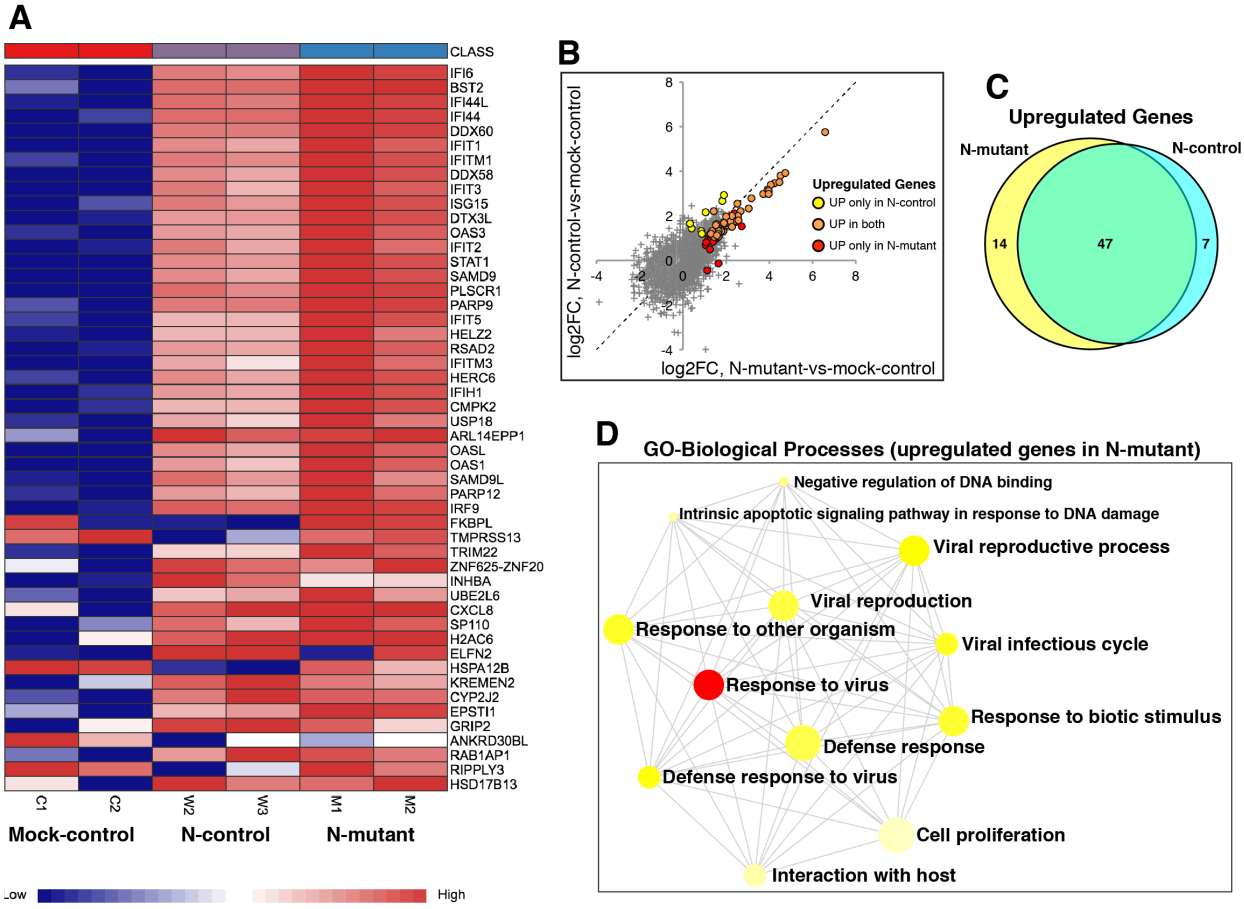

Figure 5. Transcriptional profiling of mutant and control N transfected cells.
HEK293T cells were transfected with plasmids expressing the full-length N-control and N-mutant protein along with mock control. 48-hour post-transfection total RNA was isolated and

subjected to RNA-sequencing using illumina NovaSeq 6000 platform. A) Heatmap shows normalized expression of top significantly differentially expressed genes in N-mutant and N-control conditions (adj p-value <0.05 and log2 fold-change cutoff ≥1). Genes enriched in interferon and immune related processes are overexpressed in the N-mutant transfected cells. The heatmap was generated by the visualization module in the NetworkAnalyst. B) Plot showing comparison of fold-changes for up-regulated genes in N-mutant and N-control conditions. Differentially expressed genes display higher up-regulation in the N-mutant condition (as orange dots that represent common up-regulated genes are skewed towards the lower half of the diagonal). C) Venn diagram shows the common and unique up-regulated genes in both conditions. D) GO-enrichment analysis of uniquely up-regulated genes in the N-mutant condition. The enriched GO BP (Biological Processes) term is related to interferon response. The enriched terms display an interconnected network with overlapping gene sets (from the list). Each node represents an enriched term and colored by its p-value. The size of each node corresponds to number of linked genes from the list.

The N mutant (R203K/G204R) induces overexpression of interferon related genes in transfected host cells

To understand whether the R203K/G204R mutations in the N gene affect host cell transcriptome, we transfected HEK293T cells with plasmids expressing the full-length N-control and N-mutant protein along with mock-transfection control. The transcriptome profile of N-mutant and N-control transfected cells displays a distinct pattern from the mock-control (Figure S13A). We identified 83 and 67 differentially expressed genes (DEGs) in the N-mutant and N-control transfected cells, respectively, with adjusted p-value < 0.05 and $\log_2$ fold-change ≥ 1 (Figure S13B-C and Table S7). Among the DEGs, numerous interferon, cytokine, and immune-related genes are up-regulated, some of which are shown in Figure 5 (for complete list see Table S8). We found a robust overexpression of interferon-related genes in the N-mutant compared to N-control transfected cells (Figure 5A-B) after adjusting for fold change (Figure S13D). Indeed, strong overexpression of interferon and chemokine related genes (Table S7) were reported in critical COVID-19 patients [43,44]. Recent reports further indicate a link between increased expression of interferon-related genes and higher viral load in severe COVID-19 patients [45-47].

Also, we found overexpression of other genes such as ACE2, STAT1 [44], and TMPRSS13 [48] (Figure 5A and Table S7) that are elevated in critical COVID-19 disease.

Pathway enrichment analysis of the uniquely up-regulated genes in the N-mutant condition (Figure 5C) shows an overrepresentation of biological process pathways associated with response to the virus (Figure 5D). Similarly, all up-regulated genes were related to substantially enriched pathways, such as interferon-related response, cytokine production, and viral reproductive processes (Figure S13E). The enriched GO terms display an interconnected network highlighting the relationships between up-regulated overlapping genes sets in these pathways (Figure 5D and Figure S13E). Taken together, these results suggest that the R203K/G204R mutations in the N protein may enhance its function in provoking a hyper-expression of interferon-related genes that contribute to the cytokine storm in exacerbating COVID-19 pathogenesis.

Discussion

From 892 samples collected across the country over the course of approximately 6 months we have analyzed the dynamics of transmission and diversity of SARS-CoV-2 in Saudi Arabia. The lineage analysis of assembled genomes highlights the repeated influx of SARS-CoV-2 lineages into the Kingdom. The earliest estimated importation dates point to an entry during the early stages of the pandemic (Figure 2B, Figure 2C), with the first importation likely to have an Asian origin (Figure 2C). From estimates of viral genetic diversity and reproduction rate, we find that decreased diversity and reproduction rate coincides with imposed national curfews and is followed by an observed drop in reported COVID-19 cases (Figure 1D).

Our COVID-19 patient data allowed to us detect three SNPs – underlying the N protein R203K and G204R mutations – significantly associated with higher viral load. It is worth noting that two studies have found higher viral load has in infected patients to be associated with severity and mortality [49,50]. Among our samples we initially observed an apparent association between the R203K/G204R mutations and mortality, however, the association was no longer statistically significant when correcting for sampling time.

A dated phylogenetic approach suggests that the R203K/G204R mutations arose early in the SARS-CoV-2 pandemic – perhaps even as multiple independent events – and that the mutations entered Saudi Arabia during late January 2020, most likely through Italy.

The N protein of SARS-CoV-2, an abundant viral protein within infected cells, serves multiple functions during viral infection, which besides RNA binding, oligomerization, and genome packaging, playing essential roles in viral transcription, replication, and translation [30,51]. Also, the N protein can evade immune response and perturbs other host cellular processes such as translation, cell cycle, TGFβ signaling, and induction of apoptosis [52] to enhance virus survival. The critical functional regulatory hub within the N protein is a conserved serine-arginine (SR) rich-linker region (LKR), which is involved in RNA and protein binding [53], oligomerization [33,34], and phospho-regulation [35,40].

We show that the mutant N protein containing R203K and G204R changes has higher oligomerization and stronger viral RNA binding ability, suggesting a potential link of these mutations with efficient viral genome packaging. The R203K and G204R mutations are in close proximity to the recently reported RNA-mediated phase separation domain (aa 210–246) [42] that is involved in viral RNA packaging through phase separation. This domain was thought to enhance phase-separation also through protein-protein interactions [42]. Further studies are needed

to examine any definite link between KR mutation and phase-separation; however, the differential interaction of host proteins, as shown in our study could affect this process.

Moreover, the functional activities of the N protein at different stages of the viral life cycle are regulated by phosphorylation-dependent physiochemical changes in the LKR region [40]. Although all individual phosphorylation sites may not be functionally important [32,54], the specific enhancement of phosphorylation at serine 206 in the mutant N protein shown in this study hints at its functional significance. The serine 206 can form a phosphorylation-dependent binding site for protein 14-3-3, involved in cell cycle regulatory pathways regulating human and virus protein expression [55]. Multiple lines of evidence show that N protein phosphorylation is critical for its dynamic localization and function at replication-transcription complexes (RTC), where it promotes viral RNA transcription and translation by recruiting cellular factors [38-40,56-59]. The enrichment of glycogen synthase kinase 3 A (GSK3A) with the mutant N protein, could specifically phosphorylate serine 206 in the R203K/G204R mutation background. GSK3 was shown to be a key regulator of SARS-CoV replication due to its ability to phosphorylate N protein [39]. Phosphorylation of serine 206 acts as priming site for initiating a cascade of GSK-3 phosphorylation events [39,40]. Also, GSK3 inhibition dramatically reduces the production of viral particles and the cytopathic effect in SARS-CoV-infected cells [39]. Finally, our analysis of the transcriptome in transfected cells suggests that the mutant N protein induce overexpression of interferon-related genes that can aggravate the viral infection by inducing cytokine storm.

As the COVID-19 pandemic is still ongoing, there is a need for novel therapeutic strategies to treat severe infections in patients. Our identified interaction pathways and inhibition of serine 206 phosphorylation could be used as potential targets for therapies.

In conclusion, our results highlight the major influence of the R203K/G204R mutations on the essential properties and phosphorylation status of SARS-CoV-2 N protein that lead to increased host response and efficacy of viral infection.

Methods

Sample Collection

As part of the study, nasopharyngeal swab samples were collected in 1ml of TRIzol (Ambion, USA) from 892 COVID-19 patients with various grades of clinical disease manifestations – consisting of severe, mild and asymptomatic symptoms. The anonymized samples were amassed from 8 hospitals and one quarantine hotel located in Madinah, Makkah, Jeddah and Riyadh. Ethical approvals were obtained from the Institutional review board of the Ministry of Health in Makkah region with the numbers H-02-K-076-0420-285 and H-02-K-076-0320-279, as well as the Institutional review board of Dr. Sulaiman Al Habib Hospital number RC20.06.88 for samples from Riyadh and the Eastern regions respectively.

RNA Isolation

RNA was extracted using the Direct-Zol RNA Miniprep kit (Zymo Research, USA) following the manufacturer's instructions, along with several optimization steps to improve quality and quantity of RNA from clinical samples. The optimization included extending the TRIzol incubation period, and the addition of chloroform during initial lysis step to obtain the aqueous RNA layer. The quality control of purified RNA was performed using Broad Range Qubit kit (Thermo Fisher, USA) and RNA 6000 Nano LabChip kit (Agilent, USA) respectively. RT-PCR was conducted using the one-step Super Script III with Platinum Taq DNA Polymerase (Thermo

Fisher, USA) and TaqPath COVID-19 kit (Applied Biosystems, USA) on the QuantStudio 3 Real-Time PCR instrument (Applied Biosystems, USA) and 7900 HT ABI machine. The primers and probes used were targeting two regions in the nucleocapsid gene (N1 and N2) in the viral genome following the Centre for Disease Control and prevention diagnostic panel, along with primers and probe for human RNase P gene (CDC; fda.gov/media/134922/download). Samples were considered COVID positive once the cycle threshold (Ct) values for both N1 and N2 regions were less than 40. For amplicon seq purposes, the samples chosen were of Ct less than 35 to ensure successful genome assembly in order to upload on GISAID.

Sequencing and Data analysis

cDNA and amplicon libraries were prepared using the COVID-19 ARTIC-V3 protocol, producing ~ 400bp amplicons tiling the viral genome using V3 nCoV-2019 primers (Wellcome Sanger Institute, UK; dx.doi.org/10.17504/protocols.io.beuzjex6). Amplicons were then processed for deep, paired-end sequencing with the Novaseq 6000 platform on the SP 2 x 250 bp flow cell type (Illumina, USA).

Genome assembly, SNP and indel calling

Illumina adapters and low quality sequences were trimmed using Trimmomatic v0.38 [60]. Reads were mapped to SARS-CoV-2 Wuhan-Hu-1 NCBI reference sequence NC_045512.2 using BWA [61]. Mapped reads were processed using GATK v 4.1.7 pipeline commands MarkDuplicatesSpark, HaplotypeCaller, VariantFiltration, SelectVariants, BaseRecalibrator, ApplyBQSR, and HaplotypeCaller to identify variants [62].

High quality SNPs were filtered using the filter expression:

"QD<2.0 || FS > 60.0 || SOR > 3.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"

High quality Indels were filtered using the filter expression:

"QD<2.0 || FS > 200.0 || SOR > 10.0 || ReadPosRankSum < -20.0"

Consensus sequences were generated by applying the good quality variants from GATK on the reference sequence using bcftools consensus command [63]. Regions which are covered by less than 30 reads are masked in the final assembly with 'N's.

Consensus assembly sequences were deposited to GISAID (Table S1) [11]. To retrieve high-confidence SNPs assembled sequences were re-aligned against the Wuhan-Hu-1 reference sequence (NC_045512.2), and only positions in the sample sequences with unambiguous bases in a 7-nucleotide window centred around the SNP position were kept for further analysis.

Phylogenetic analysis

To generate the phylogeny of Saudi samples with a global context, a total of 308,012 global sequences were downloaded from GISAID on 31 December 2020, filtered and processed using Nextstrain pipeline[12]. Global sequences were grouped by country and sample collection month and 20 sequences per group were randomly sampled which resulted in 10,873 global representative sequences and 952 Saudi sequences. The phylogeny was constructed using IQ-TREE[64], clades were assigned using Nextclade and internal node dates were inferred and sequences pruned using TreeTime[65]. Nextstrain protocol was followed for the above-mentioned steps. The resulting global phylogenetic tree was reduced to retain the branches that lead to Saudi leaf nodes and visualised using baltic library (https://github.com/evogytis/baltic).

Phylodynamic analysis

Phylodynamic analyses use the same sequence subset used in the full phylogenetic analysis, extracted from the GISAID SARSCoV-2 database [11]. Wrapper functions for the importation date estimates and skygrowth model are provided in the sarscov2 R package as 'compute_timports' and 'skygrowth1' respectively (github.com/emvolz-phylodynamics/sarscov2Rutils).

Importation date estimates for Nextstrain clades

Sequences corresponding to each Nextstrain [12] clade were extracted using the Nextstrain_clade parameter in the GISAID metadata table. A subset of 500 international sequences were select for each clade based on Tamura Nei 93 distance with tn93 (github.com/veg/tn93) and stratified over time[66]. A maximum likelihood phylogeny with an HKY substitution model for each clade was estimated with IQtree[64,67]. Time-scaled phylogenies were estimated from this using treedater with a strict molecular clock constrained between 0.0009 and 0.0015 substitutions per site per year[68]. 15 Variations of each dated phylogeny were produced by collapsing small branches and resolving polytomies. The state of each internal node was reconstructed by maximum parsimony with the phangorn R package[69]. Importation events are estimated at the midpoint of a branch along which a location change is inferred to occur by this method.

Estimation of donor countries behind importation events

To identify import events that resulted in new introductions into Saudi Arabia, 25,198 sequences were subsampled from 590K global sequences available on GISAID on February 24th 2021. Samples with closer genetic distance to Saudi Samples were preferred. The phylogeny was constructed using IQ-TREE[64], internal nodes dates and possible country for internal nodes were inferred using TreeTime[65]. Nextstrain protocol was followed for the above-mentioned steps[12]. In

house scripts were used to traverse the global phylogenetic tree to identify branches that resulted in transitions into Saudi Arabia from another country.

Skygrowth model

Sequences from Saudi Arabia available on GISAID on December 31st 2020 were used to construct effective population size and growth rate of SARS-CoV2 in Saudi Arabia over the course of the first wave of the epidemic (March to September 2020). As with the importation date estimates, a maximum likelihood phylogeny was produced, time-scaled and variation introduced by resolving polytomies to give a sample of 15 phylogenies.

We modelled growth rate and effective population size over time on these phylogenies using the R package skygrowth[16]. Skygrowth is a non-parametric Bayesian approach which applies a stochastic process on estimates of growth rate and effective population size. The model included mean-centred, unit variance estimates of travel rates from google mobility data (google.com/covid19/mobility/) as a covariate (transit stations percent change from baseline), 60 timesteps and a tau (precision) value corresponding to a 1% change in growth per week. The growth rate output was converted to an estimate of R over time using an infectious period of 9.5 days[70].

Origin of R203K/G204R SNPs

A total of 590K samples submitted to GISAID until February 24 were downloaded and SNPs indentifed by mapping against the Wuhan reference using minimap2[71]. The variants were queried to count the distribution of triplets among various Nextstrain clades (Figure S7). To identify if there are lineages of triplet SNPs in clades other than 20B, a phylogenetic tree was

constructed by including all R203K/G204R samples found in other clades outside 20B and its subclades (Figure S4). As it was already evident that 20B and its subclades contains lineages of R203K/G204R samples, subsamples from 20B and its subclades were sufficient to obtain a total of 16,386 samples.

Statistical analysis

Statistical analyses were performed with the statistical software R version 4.0.3 [72] and the R package mgcv version 1.8.33. Model with response ($\log_{10}$ copynumber) used 473 observations (samples processed with the TaqPath kit). Models for mortality and severity used 892 observations (all data).

Plasmid and cloning

The pLVX-EF1alpha-SARS-CoV-2-N-2xStrep-IRES-Puro was a gift from Nevan Krogan (Addgene plasmid # 141391; http://n2t.net/addgene:141391; RRID:Addgene_141391)[36]. The three consecutive SNPs (G28881A, G28882A, G28883C), corresponding to N protein mutation sites R203K and G204R, were introduced by megaprime PCR mutagenesis using the primers listed in Table S8.

Cell culture and transfection

HEK293T (ATCC; CRL-3216) cells were grown in Dulbecco's modified Eagle's medium (DMEM) (4.5 g/l d-glucose and Glutamax, 1 mM sodium pyruvate) (GIBCO) and 10% fetal bovine serum (FBS; GIBCO) with penicillin–streptomycin supplement, according to standard protocols (culture condition 37 °C and 5% CO2). Transfection of ten million cells per 15-cm dish

with 2XStrep-tagged N plasmid (20ug/transfection) was performed using lipofectamine-2000 according standard protocol.

Affinity purification and on-bead digestion

Cell lysis and affinity purification with MagStrep beads (IBA Lifesciences) was manually performed according to the published protocol[36] with minor modifications. Briefly, after transfection (48 hours) cells were collected with 10mM EDTA in 1xPBS and washed twice with cold PBS (1x). The cell pellets were stored at −80 °C. Cells were lysed in lysis buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.5% NP40, supplemented with protease and phosphatase inhibitor cocktails) for 30 minutes while rotating at 4 °C and then centrifuge at high speed to collect the supernatant. The cell lysate was incubated with prewashed MagStrep beads (30 µl per reaction) for 3 hours at 4 °C. The beads were then washed four times with wash buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.05% NP40, supplemented with protease and phosphatase inhibitor cocktails) and then proceed with on-bead digestion. The on-bead digestion was carried out as described before[36]. For affinity confirmation, bound proteins were eluted using buffer BXT (IBA Lifesciences) and after running on SDS-PAGE were subjected to silver staining and western-blot using anti-strep-II antibody (ab76949). To purify clean 2xStrep-tagged N protein (mutant and control), we applied stringent washing and double elution strategy.

MS analysis using Orbitrap Fusion Lumos

The MS analysis was performed as described previously [73,74] with slight modifications. For mass spectrometry analysis an Orbitrap Fusion mass spectrometer (MS) (Lumos, Thermo Fisher

Scientific) was used in data-dependent acquisition (DDA) mode. For injection, 0.5 µg peptide mixture was used and desalting was performed for 5 minutes in 0.1% FA in water. The gradient and all other steps were essentially the same as described[73].

Protein identification analysis from the raw mass spectrometry data was performed using the Maxquant software (version 1.5.3.30)[75] as described[73].

For phosphorylated peptides, we used Maxquant label-free quantification (LFQ)[75]. The analysis and quantification of phosphorylated peptides was performed according to published protocol[76].

Analysis of differential interaction

The normalized LFQ data were processed for statistical analysis on the LFQ-Analyst a web-based tool[77] to performed pair-wise comparison between mutant and control N protein AP-MS data. The significant differentially changed proteins between mutant and control conditions were identified. The threshold cut-off of adjusted p-value <= 0.05, and Log fold change >= 1 were used. Among the replicates, outliers were removed based on correlation and PCA analysis. The GO enrichment analysis was performed on the LFQ-Analyst[77].

BS3 cross-linking

Bis(sulfosuccinimidyl) suberate (BS3, Thermo Scientific Pierce) was used for cross-linking of control and mutant N protein to analyse the oligomerization properties. The experiment was performed as reported previously[78].

RNA-sequencing and differential gene expression analysis

HEK293T cells were transfected with plasmids expressing the full-length N-control and N-mutant protein along with mock control. After 48-hour cells were harvested in Trizol and total

RNA was isolated using Zymo-RNA Direct-Zol kit (Zymo, USA) according to the manufacture's instruction. The concentration of RNA was measured by Qubit (Invitrogen), and RNA integrity was determined by Bioanalyzer 2100 system (Agilent Technologies, CA, USA). The RNA was then subjected to library preparation using Ribozero-plus kit (Illumina). The libraries were sequenced on NovaSeq 6000 platform (Illumina, USA) with 150 bp paired-end reads.

The raw reads from HEK293T RNA-sequencing were processed and trimmed using trimmomatic [60] and mapped to annotated ENSEMBL transcripts from the human genome (hg19) [79,80] using kallisto[81]. Differential expression analysis was performed after normalization using EdgeR integrated in the NetworkAnalyst [82]. GO biological process and pathway enrichment analyses on up-regulated genes were performed using NetworkAnalyst [82].

Acknowledgments

We are sincerely grateful to all hospital members for providing samples and collating metadata in such an unprecedented pandemic, along with the MOH and ethical committee, which rendered it permissible. We thank the KAUST Rapid Research Response Team (R3T) under the Vice President – Research (VPR) office in KAUST for generous financial support. We also thank Erik Talley from KAUST Health Safety and Environment (HSE) and Hani Bukhari from KAUST Security for providing timely logistical support for samples transport during COVID-19 Curfew restrictions in the Kingdom.

We extend our thanks and appreciation to GDRS director, PH. Athari Alotaibi (General Director for Research and Studies, MOH) for her vigorous facilitation of the research project, and Mohammad Fawzi (General Directorate of Health Affairs) for his help with the metadata collection. We also deeply thank Dr. Wael Hamzah Motair, Dr. Nader Hamzah Motair, Dr.

Hatim Khogeer and the General Directorate of Health Affairs of Makkah Region (GDHAMR), MOH for all their help and assistance to our study.

We gratefully acknowledge all of the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and publicly shared via the GISAID Initiative, on which was partially used for additional support for some of the conclusions drawn in this study.

Grants:

KAUST Rapid Research Response Team (R3T) by Vice President – Research (VPR) office in KAUST.

KAUST faculty baseline fund (BAS/1/1020-01-01) to AP.

KACST Grants, Proposal number: 5-20-01-002-0008

MOH COVID-19 project grants number 341

MOH COVID-19 project grants number 754

The deputyship for Research and Innovation, Ministry of Education in Saudi Arabia, project number (436) to AMH.

IRBs:

This project was conducted under the IRB approvals of the MOH (H-02-K-076-0420-285), KAUST (20IBEC14) and at the Dr. Suliman Al-Habib Medical group (HAP-01-R-082) in KSA.

Author Contributions

A.P. conceived the study and directed the work and acquired funding from KAUST and supplemental funding from King Abdulaziz City for Science and Technology (KACST). A.P., T.M., M.S., S.H., and S.M. designed the research. IRB and ethical approvals from MOH were acquired by A.K., A.H., N.A., A.M., and S.H. to cover the collection from several cities in the Kingdom. S.H. and A.K. acquired funding from the Saudi Ministry of Health (MOH) numbers 754 and 341, utilized in the study. S.H. organized and directed sample collection and metadata collections with aid from F.A., A.S., A.O., S.S., J.T., A.A., N.K., K.K., K.A., and A.D.; S.M. directed the wet lab work involving sample reception, metadata record-keeping, RNA extraction, quality control, and library preparation, with aid from A.S., F.B., R.S., M.S., A.O., L.E., O.D., S.H. and R.N. Illumina sequencing runs, Mass spectrometry and raw data processing were performed by S.M., L.E., S.P., and I.R. respectively. Genome assemblies and submission to GISAID was done by R.N. Phylogenetic and lineage analysis was done by R.N., Q.G., D.J., and E.V. In-depth SNP data analysis was performed by T.M. Statistical analysis done by P.E.M. and E.V. Functional validation of this link was established by M.S.; T.M. wrote the initial draft of the manuscript with input from M.S., S.M., S.H., R.N., and Q.G., followed by edits from A.P. The final version was produced by T.M., M.S. and A.P. after input from all co-authors.

Data availability

Supplementary Information is available for this paper. Assembled virus genomes are available at GISAID (Table S1). Reads from RNA-Seq analysis of transfected HEK293T cells have been uploaded to European Nucleotide Archive (https://www.ebi.ac.uk/ena/) under the Study accession number PRJEB44716.

Correspondence and requests for materials should be addressed to Arnab Pain (arnab.pain@kaust.edu.sa).

References

1 Organization, W. H. *Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update*, <www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> (2020).

2 Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* **20**, 533-534, doi:10.1016/S1473-3099(20)30120-1 (2020).

3 Center, J. H. U. M. C. R. *COVID-19 Dashboard*, <https://coronavirus.jhu.edu/map.html> (2020).

4 Ebrahim, S. H. & Memish, Z. A. COVID-19: preparing for superspreader potential among Umrah pilgrims to Saudi Arabia. *Lancet* **395**, e48, doi:10.1016/S0140-6736(20)30466-9 (2020).

5 Memish, Z. A., Aljerian, N. & Ebrahim, S. H. Tale of three seeding patterns of SARS-CoV-2 in Saudi Arabia. *Lancet Infect Dis*, doi:10.1016/S1473-3099(20)30425-4 (2020).

6 Tuite, A. R. *et al.* Estimation of Coronavirus Disease 2019 (COVID-19) Burden and Potential for International Dissemination of Infection From Iran. *Ann Intern Med* **172**, 699-701, doi:10.7326/M20-0696 (2020).

7 News, A. *Saudi Arabia announces first case of coronavirus*, <https://www.arabnews.com/node/1635781/saudi-arabia> (2020).

8 Gussow, A. B. *et al.* Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 15193-15199, doi:10.1073/pnas.2008176117 (2020).

9 Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003 e1009, doi:10.1016/j.cell.2020.04.023 (2020).

10 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33-46, doi:https://doi.org/10.1002/gch2.1018 (2017).

11 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, doi:10.2807/1560-7917.ES.2017.22.13.30494 (2017).

12 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).

13 Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75 e11, doi:10.1016/j.cell.2020.11.020 (2021).

14 Davies, N. G. *et al.* Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*, doi:10.1038/s41586-021-03426-1 (2021).

15 Lin, J. W. *et al.* Genomic monitoring of SARS-CoV-2 uncovers an Nsp1 deletion variant that modulates type I interferon response. *Cell Host Microbe* **29**, 489-502 e488, doi:10.1016/j.chom.2021.01.015 (2021).

16 Volz, E. M. & Didelot, X. Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. *Syst Biol* **67**, 719-728, doi:10.1093/sysbio/syy007 (2018).
17 Leary, S. *et al.* Three adjacent nucleotide changes spanning two residues in SARS-CoV-2 nucleoprotein: possible homologous recombination from the transcription-regulating sequence. *bioRxiv*, 2020.2004.2010.029454, doi:10.1101/2020.04.10.029454 (2020).
18 Morel, B. *et al.* Phylogenetic analysis of SARS-CoV-2 data is difficult. *Molecular biology and evolution*, doi:10.1093/molbev/msaa314 (2020).
19 Turakhia, Y. *et al.* Stability of SARS-CoV-2 phylogenies. *PLoS genetics* **16**, e1009175, doi:10.1371/journal.pgen.1009175 (2020).
20 consortium, T. C.-G. U. C.-U. An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe* **1**, e99-e100, doi:10.1016/S2666-5247(20)30054-9 (2020).
21 De Maio, N. *et al.* (virological.org, 2020).
22 Yi, H. 2019 Novel Coronavirus Is Undergoing Active Recombination. *Clinical Infectious Diseases* **71**, 884-887, doi:10.1093/cid/ciaa219 (2020).
23 Richard, D., Owen, C. J., van Dorp, L. & Balloux, F. No detectable signal for ongoing genetic recombination in SARS-CoV-2. *bioRxiv*, 2020.2012.2015.422866, doi:10.1101/2020.12.15.422866 (2020).
24 Wu, S. *et al.* Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions. *J Med Virol*, doi:10.1002/jmv.26597 (2020).
25 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407, doi:10.1038/s41564-020-0770-5 (2020).
26 Jackson, B. *et al. Recombinant SARS-CoV-2 genomes involving lineage B.1.1.7 in the UK*, <https://virological.org/t/recombinant-sars-cov-2-genomes-involving-lineage-b-1-1-7-in-the-uk/658> (2021).
27 Wood, S. *Generalized Additive Models: An Introduction with R, 2 edition*. (Chapman and Hall/CRC, 2017).
28 Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827 e819, doi:10.1016/j.cell.2020.06.043 (2020).
29 Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814, doi:10.1093/nar/gkg509 (2003).
30 McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991-3018, doi:10.3390/v6082991 (2014).
31 Rahman, M. S. *et al.* Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *J Med Virol*, doi:10.1002/jmv.26626 (2020).
32 Guan, Q. *et al.* A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. *Int J Infect Dis* **100**, 216-223, doi:10.1016/j.ijid.2020.08.052 (2020).
33 He, R. T. *et al.* Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem Bioph Res Co* **316**, 476-483, doi:10.1016/j.bbrc.2004.02.074 (2004).
34 Chang, C. K., Chen, C. M. M., Chiang, M. H., Hsu, Y. L. & Huang, T. H. Transient Oligomerization of the SARS-CoV N Protein - Implication for Virus Ribonucleoprotein Packaging. *Plos One* **8**, doi:ARTN e65045

10.1371/journal.pone.0065045 (2013).
35 Chao Wu, A. J. Q., Asmaa Hachim, Niloufar Kavian, Aidan R. Cole, Austin B. Moyle, Nicole D. Wagner, Joyce Sweeney-Gibbons, Henry W. Rohrs, Michael L. Gross, J. S. Malik Peiris, Christopher F. Basler, Christopher W. Farnsworth, Sophie A. Valkenburg, Gaya K. Amarasinghe, Daisy W. Leung. Characterization of SARS-CoV-2 N protein reveals multiple functional consequences of the C-terminal domain. *BioRxiv*, doi:https://doi.org/10.1101/2020.11.30.404905 (2020).
36 Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459-+, doi:10.1038/s41586-020-2286-9 (2020).
37 Lowrey, A. J., Cramblet, W. & Bentz, G. L. Viral manipulation of the cellular sumoylation machinery. *Cell Commun Signal* **15**, doi:ARTN 27
10.1186/s12964-017-0183-0 (2017).
38 Wu, C. H., Chen, P. J. & Yeh, S. H. Nucleocapsid Phosphorylation and RNA Helicase DDX1 Recruitment Enables Coronavirus Transition from Discontinuous to Continuous Transcription. *Cell Host Microbe* **16**, 462-472, doi:10.1016/j.chom.2014.09.009 (2014).
39 Wu, C. H. *et al.* Glycogen Synthase Kinase-3 Regulates the Phosphorylation of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein and Viral Replication. *J Biol Chem* **284**, 5229-5239, doi:10.1074/jbc.M805747200 (2009).
40 Carlson, C. R. *et al.* Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions. *Mol Cell* **80**, 1092-+, doi:10.1016/j.molcel.2020.11.025 (2020).
41 Savastano, A., de Opakua, A. I., Rankovic, M. & Zweckstetter, M. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nat Commun* **11**, doi:ARTN 6041
10.1038/s41467-020-19843-1 (2020).
42 Lu, S. *et al.* The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nature communications* **12**, 502, doi:10.1038/s41467-020-20768-y (2021).
43 Gill, S. E. *et al.* Transcriptional profiling of leukocytes in critically ill COVID19 patients: implications for interferon response and coagulation. *Intensive Care Med Exp* **8**, 75, doi:10.1186/s40635-020-00361-9 (2020).
44 Jain, R. *et al.* Host transcriptomic profiling of COVID-19 patients with mild, moderate, and severe clinical outcomes. *Comput Struct Biotechnol J* **19**, 153-160, doi:10.1016/j.csbj.2020.12.016 (2021).
45 Nienhold, R. *et al.* Two distinct immunopathological profiles in autopsy lungs of COVID-19. *Nature communications* **11**, 5086, doi:10.1038/s41467-020-18854-2 (2020).
46 Lieberman, N. A. P. *et al.* In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. *PLoS Biol* **18**, e3000849, doi:10.1371/journal.pbio.3000849 (2020).
47 Sposito, B. *et al.* Severity of SARS-CoV-2 infection as a function of the interferon landscape across the respiratory tract of COVID-19 patients. *bioRxiv*, 2021.2003.2030.437173, doi:10.1101/2021.03.30.437173 (2021).
48 Kishimoto, M. *et al.* TMPRSS11D and TMPRSS13 Activate the SARS-CoV-2 Spike Protein. *Viruses* **13**, doi:10.3390/v13030384 (2021).

49 Fajnzylber, J. *et al.* SARS-CoV-2 viral load is associated with increased disease severity and mortality. *Nature communications* **11**, 5493, doi:10.1038/s41467-020-19057-5 (2020).
50 Pujadas, E. *et al.* SARS-CoV-2 viral load predicts COVID-19 mortality. *Lancet Respir Med* **8**, e70, doi:10.1016/S2213-2600(20)30354-4 (2020).
51 Chang, C. K., Hou, M. H., Chang, C. F., Hsiao, C. D. & Huang, T. H. The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res* **103**, 39-50, doi:10.1016/j.antiviral.2013.12.009 (2014).
52 Lal, M. S. a. S. K. in *Molecular Biology of the SARS-Coronavirus* (ed Sunil K. Lal) 129-151 (2009).
53 Wegener, M. & Muller-McNicoll, M. View from an mRNP: The Roles of SR Proteins in Assembly, Maturation and Turnover. *Adv Exp Med Biol* **1203**, 83-112, doi:10.1007/978-3-030-31434-7_3 (2019).
54 Bouhaddou, M. *et al.* The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* **182**, 685-712 e619, doi:10.1016/j.cell.2020.06.034 (2020).
55 Nathan, K. G. & Lal, S. K. The Multifarious Role of 14-3-3 Family of Proteins in Viral Replication. *Viruses* **12**, doi:10.3390/v12040436 (2020).
56 Verheije, M. H. *et al.* The Coronavirus Nucleocapsid Protein Is Dynamically Associated with the Replication-Transcription Complexes. *J Virol* **84**, 11575-11579, doi:10.1128/Jvi.00569-10 (2010).
57 Chen, H. Y. *et al.* Mass spectroscopic characterization of the coronavirus infectious bronchitis virus nucleoprotein and elucidation of the role of phosphorylation in RNA binding by using surface plasmon resonance. *J Virol* **79**, 1164-1179, doi:10.1128/Jvi.79.2.1164-1179.2005 (2005).
58 Peng, T. Y., Lee, K. R. & Tarn, W. Y. Phosphorylation of the arginine/serine dipeptide-rich motif of the severe acute respiratory syndrome coronavirus nucleocapsid protein modulates its multimerization, translation inhibitory activity and cellular localization. *Febs J* **275**, 4152-4163, doi:10.1111/j.1742-4658.2008.06564.x (2008).
59 V'kovski, P. *et al.* Determination of host proteins composing the microenvironment of coronavirus replicase complexes by proximity-labeling. *Elife* **8**, doi:ARTN e42037 10.7554/eLife.42037 (2019).
60 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).
61 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
62 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
63 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).
64 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).
65 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042, doi:10.1093/ve/vex042 (2018).

66 Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution* **10**, 512-526, doi:10.1093/oxfordjournals.molbev.a040023 (1993).
67 Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-174, doi:10.1007/BF02101694 (1985).
68 Volz, E. M. & Frost, S. D. W. Scalable relaxed clock phylogenetic dating. *Virus Evolution* **3**, doi:10.1093/ve/vex025 (2017).
69 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).
70 Hu, Z. *et al.* Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Sci China Life Sci* **63**, 706-711, doi:10.1007/s11427-020-1661-4 (2020).
71 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
72 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org. (2017).
73 Zhang, H. M. *et al.* Arabidopsis proteome and the mass spectral assay library. *Sci Data* **6**, doi:ARTN 278
10.1038/s41597-019-0294-0 (2019).
74 Liu, P., Shuaib, M., Zhang, H. M., Nadeef, S. & Orlando, V. Ubiquitin ligases HUWE1 and NEDD4 cooperatively control signal-dependent PRC2-Ezh1 alpha/beta-mediated adaptive stress response pathway in skeletal muscle cells. *Epigenet Chromatin* **12**, doi:10.1186/s13072-019-0322-5 (2019).
75 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).
76 Wu, D. *et al.* Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer. *Nature* **559**, 637-+, doi:10.1038/s41586-018-0350-5 (2018).
77 Shah, A. D., Goode, R. J. A., Huang, C., Powell, D. R. & Schittenhelm, R. B. LFQ-Analyst: An Easy-To-Use Interactive Web Platform To Analyze and Visualize Label-Free Proteomics Data Preprocessed with MaxQuant. *J Proteome Res* **19**, 204-211, doi:10.1021/acs.jproteome.9b00496 (2020).
78 Zeng, W. H. *et al.* Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem Bioph Res Co* **527**, 618-623, doi:10.1016/j.bbrc.2020.04.136 (2020).
79 Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford)* **2016**, doi:10.1093/database/baw093 (2016).
80 Yates, A. D. *et al.* Ensembl 2020. *Nucleic acids research* **48**, D682-D688, doi:10.1093/nar/gkz966 (2020).
81 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).
82 Zhou, G. *et al.* NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic acids research* **47**, W234-W241, doi:10.1093/nar/gkz240 (2019).