International Journal of Antimicrobial Agents 57 (2021) 106272

Contents lists available at ScienceDirect

International Journal of Antimicrobial Agents

journal homepage: www.elsevier.com/locate/ijantimicag

Different mutations in SARS-CoV-2 associate with severe and mild outcome

Ádám Nagy[a,b], Sándor Pongor[c], Balázs Győrffy[a,b,*]

[a] Department of Bioinformatics, Semmelweis University, Budapest, Hungary
[b] TTK Momentum Cancer Biomarker Research Group, Budapest, Hungary
[c] Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary

ARTICLE INFO

Article history:
Received 21 October 2020
Accepted 12 December 2020

Dr Jim Gray

Keywords:
SARS-CoV-2
mutation
next generation sequencing
genome
death
high-risk

ABSTRACT

Introduction: Genomic alterations in a viral genome can lead to either better or worse outcome and identifying these mutations is of utmost importance. Here, we correlated protein-level mutations in the SARS-CoV-2 virus to clinical outcome.

Methods: Mutations in viral sequences from the GISAID virus repository were evaluated by using "hCoV-19/Wuhan/WIV04/2019" as the reference. Patient outcomes were classified as mild disease, hospitalization and severe disease (death or documented treatment in an intensive-care unit). Chi-square test was applied to examine the association between each mutation and patient outcome. False discovery rate was computed to correct for multiple hypothesis testing and results passing FDR cutoff of 5% were accepted as significant.

Results: Mutations were mapped to amino acid changes for 3,733 non-silent mutations. Mutations correlated to mild outcome were located in the ORF8, NSP6, ORF3a, NSP4, and in the nucleocapsid phosphoprotein N. Mutations associated with inferior outcome were located in the surface (S) glycoprotein, in the RNA dependent RNA polymerase, in ORF3a, NSP3, ORF6 and N. Mutations leading to severe outcome with low prevalence were found in the ORF3A and in NSP7 proteins. Four out of 22 of the most significant mutations mapped onto a 10 amino acid long phosphorylated stretch of N indicating that in spite of obvious sampling restrictions the approach can find functionally relevant sites in the viral genome.

Conclusions: We demonstrate that mutations in the viral genes may have a direct correlation to clinical outcome. Our results help to quickly identify SARS-CoV-2 infections harboring mutations related to severe outcome.

1. Introduction

There are seven human coronaviruses including MERS, Human-HKU-1, Human NL63, Human 229E, Human OC43, SARS-CoV, and SARS-CoV-2. The natural host of this latest RNA virus is the Chinese rufous horseshoe bat (*Rhinolophus sinicus*) and the transfer to human initiated the ongoing COVID-19 outbreak at the end of 2019 [1]. Some studies estimated a low mortality rate of SARS-CoV-2 in the overall population[2,3], while other investigators reported mortality percentages up to 26% when the virus strikes a critically ill patient [4]. Overall, based on current data of the WHO (October 2020), the mortality rate is around 2.7%.

The linear genome of the SARS-CoV-2 virus has 29,903 bases and harbors 25 genes [5], the reference sequence is accessible in GeneBank using the accession number MN908947. Phylogenetic analysis of SARS-CoV-2 genomes show three variants termed A, B and C which have different distribution when comparing sequences from Asia, Europe or the Americans [6]. The viral genes encode among others an envelope protein, an RNA dependent RNA polymerase, a surface glycoprotein, an exonuclease, a methyltransferase, and 11 nonstructural proteins. Some of these are within the virus, but others, including the spike glycoprotein, the membrane glycoprotein, and the envelope protein are on the viral surface.

In theory, any functional or structural viral gene can have an effect on the efficiency of a virus and both mutations [7] or alter-

* Correspondence: Balázs Győrffy MD PhD DSc, Department of Bioinformatics, Semmelweis University, Tüzoltó u. 7-9, 1094, Budapest, Hungary. Tel: +3630-514-2822.
E-mail address: gyorffy.balazs@med.semmelweis-univ.hu (B. Győrffy).

https://doi.org/10.1016/j.ijantimicag.2020.106272

ation in the expression [8] can increase pathogenicity. It is important to emphasize that even the untranslated regions of a coronavirus can have important role in the viral replication as has been previously demonstrated for the 3' untranslated region [9]. SARS-CoV-2 is no different compared to other viruses and new mutations continually pop up with its spread [10]. Some mutations uncovered in the SARS-CoV-2 virus lead to a novel RNA-dependent-RNA polymerase variant [11], while other genomic changes drive the evolution and the spread of the virus by resulting in a more transmissible form of the virus [12]. Mutations potentially making the virus more transmissible have a significant evolutionary advantage as has been demonstrated for the SARS-CoV-2 variant with spike G614 which mainly replaced D614 between February and July 2020 [13].

In this context, the most important question is to identify viral mutations leading to different patient outcomes. Mutations resulting in a mild disease could facilitate the spread of the virus and thereby maintain the outbreak. Other mutations leading to a more severe disease need immediate attention to prevent detrimental outcomes. Here, our goal was to identify and rank mutations associated with altered patient outcome by simultaneously correlating outcomes to all mutations across a large cohort of patients.

2. Materials and Methods

2.1. Data source

All available SARS-CoV-2 (taxid: 2697049) viral nucleic acid sequences were downloaded from the GISAID virus repository (https://www.gisaid.org/). The sequences were acquired in FASTA format. Those viral sequences were selected where the entire viral nucleic acid sequence was published. A second filtering was executed to include only virus genomes with available patient follow-up status.

2.2. Mapping of mutations to viral genes

The mutations were evaluated using the CoVsurver (https://corona.bii.a-star.edu.sg). To achieve this, the viral sequences in .FASTA format were used as the query and the "hCoV-19/Wuhan/WIV04/2019" was used as the reference. The analysis was run by using batches of 1000 samples in one run. Protein mutations do not have overlaps, and the genomic boundaries of the various proteins in the WIV04 reference genome are displayed in **Table 1**.

2.3. Clinical classification

As the patient samples were annotated with all together more than sixty different outcome classification, we had to coerce these into three major categories.

Patients who were "asymptomatic", were "not hospitalized", had a "mild" disease, were at "home" were all assigned to have a "mild" disease. Also patients who were treated at outpatient departments, were quarantined or were treated by the physician network were classified as "mild".

Patients who definitely needed medical care were assigned to the "hospitalized" group. These include those "hospitalized", "inpatient", "discharged", "released", and "recovered". In addition, combinations of the annotations which included any of these were also assigned into this cohort (e.g. "initially hospitalized" or "to be hospitalized").

Finally, patients with detrimental outcome were allocated to the "severe" cohort. These include those "deceased", those with a "severe" disease, those who entered "intensive care units". Also any combination of these with other annotations (e.g. "hospitalized / ICU") were also added to this category.

Table 1
Genomic boundaries of the SARS-CoV-2 proteins in the WIV04 reference genome.

Protein name	Code	Genomic positions
Envelope (E) protein	E	26245-26472
Membrane glycoprotein	M	26523-27191
Nucleocapsid phosphoprotein	N	28274-29533
ORF3a protein	NS3	25393-26220
ORF6 protein	NS6	27202-27387
ORF7a protein	NS7a	27394-27759
ORF7b	NS7b	27756-27887
ORF8 protein	NS8	27894-28259
NSP1	NSP1	266-805
NSP10	NSP10	13025-13441
NSP11	NSP11	13442-13480
RNA dependent RNA polymerase	NSP12	13442-13468|13468-16236
helicase	NSP13	16237-18039
3′-to5′ exonuclease	NSP14	18040-19620
endoRNAse	NSP15	19621-20658
2′O'ribose methyltransferase	NSP16	20659-21552
NSP2	NSP2	806-2719
NSP3	NSP3	2720-8554
NSP4	NSP4	8555-10054
NSP5	NSP5	10055-10972
NSP6	NSP6	10973-11842
NSP7	NSP7	11843-12091
NSP8	NSP8	12092-12685
NSP9	NSP9	12686-13024
Surface (S) glycoprotein	Spike	21563-25384

2.4. Statistical computation

All data processing and statistical analysis steps were performed in the R statistical environment v 3.6.3. Data processing was performed on 18th October 2020. Chi-square test was applied to examine the association between each mutation and patient status data. False discovery rate using the Benjamini-Hochberg method was computed to correct for multiple hypothesis testing and only results passing a FDR cutoff of 5% were accepted as significant.

3. Results

3.1. Dataset

All together 149,061 SARS-CoV-2 viral nucleic acid sequences were available, and 147,960 of these included the entire viral nucleic acid sequence. Clinical data was available for 7,702 patients, and 4,566 of these had also follow-up data. This is a small fraction of the total data which implies that our findings could contain a sampling bias.

When looking on the clinical parameters of these patients, 58.6% were male and 36.5% were female (remaining samples did not had this information). The geographical origin of the samples covers the entire globe: 4.2% were from Africa, 46% from Asia, 26.8% from Europe, 12.4% from North America and 10.2% from South America. Collection of the samples happened between 30.12.2019 and 14.9.2020. Of all patients with a follow-up 708 had a mild disease, 3,306 had to be hospitalized and 552 patients had a severe disease.

3.2. Mutation rate

All together 3,733 different mutations affecting the protein amino acid sequence were identified, and 937 of these mutations were not present in samples with clinical follow-up. When looking

Table 2
SARS-CoV-2 mutations correlated to mild outcome in 4,566 patients with available genomic and follow-up information were found in five distinct genes.

Protein name	Protein mutation	N wild type + mild outcome	N wild type + hospitalized	N wild type + severe outcome	N mutant + mild outcome	N mutant + hospitalized	N mutant + severe outcome	Chi squared test p value
ORF8 protein	L84S	477	3085	536	**231**	221	16	2.3E-101
NSP6	L37F	564	2816	537	**144**	490	15	1.1E-18
ORF3a protein	G196V	534	3247	545	**174**	59	7	3.7E-137
NSP4	F308Y	533	3241	545	**175**	65	7	2.2E-133
Nucleocapsid phosphoprotein	S197L	533	3241	545	**175**	65	7	2.2E-133

on all mutations, we have identified on average 4.7 mutations in each sample.

As an internal control to validate any potential bias in the mutation prevalence related to patient proportions we computed the average numbers of mutation in each clinical outcome cohort and found similar values (mean in those with mild, hospitalized, and severe outcome were 4.8, 4.6, and 5.1, respectively).

When analyzing the correlation to clinical outcome across all mutations, 79 mutations reached statistical significance at FDR<5%. The complete list of these mutations with sample numbers in each cohort is displayed in **Supplemental Table 1** and mutation data for each investigated patient is provided in **Supplemental Table 2.**

3.3. Mutations related to mild disease

In order to concentrate only on mutations with a clinical relevance, we selected only those mutations which were present in at least 2% of the samples (this corresponds to a cutoff of at least 91 patient samples with a mutation). When looking at mutation related to mild outcome, only five mutations passed all filtering criteria - L84S in the ORF8 protein, L37F in the NSP6 protein, G196V in the ORF3a protein, F308Y in the NSP4 protein, and the S197L mutation in the nucleocapsid phosphoprotein. The complete list as well as distribution among patient samples is provided in **Table 2**.

3.4. Mutations associated with severe disease

When searching for mutations related to hospitalization or to severe outcome, we used the above filter of including only mutations present in at least 2% of the samples. All together 15 mutations passed these criteria. These originated in seven genes: L54F, D614G and V1176F in the surface (S) glycoprotein, A97V and P323L in the RNA dependent RNA polymerase, Q57H and G251V in the ORF3a protein, P13L, S194L, R203K, G204R and I292T in the nucleocapsid phosphoprotein, I33T in the ORF6 protein, S1197R and T1198K mutations in the NSP3 protein.

In order not to miss mutations leading to deadly outcome we also included all mutations which were present in at least 10 patients with severe outcome. This additional analysis delivered two further mutations, the L71F in the NSP7 protein and the S253P mutation in the ORF3A gene. These were linked to 53 and 11 severe outcomes after being spotted in 60 (L71F) and 11 (NSP7) patients, respectively.

Interestingly, the overall prevalence of mutations leading to mild outcome (n=1,851) was smaller than the prevalence of those leading to worse outcome (n=11,725), but at the same time the proportion of patients with mild outcome in the entire cohort was also smaller (18.3%). Nevertheless, a significant proportion of the mutations (n=7,875) were not significantly correlated to any clinical outcome.

The complete list of all mutations correlated to severe disease is presented in **Table 3**.

4. Discussion

We have simultaneously analyzed the correlation between patient outcome and all identified mutations resulting in amino acid sequence changes in the viral proteins. Strikingly, we have not only found a significant number of mutations, but some of these were correlated to mild diseases while other had a significant correlation to severe outcome.

Nucleocapsid phosphoprotein was the protein with most significant mutations linked to both mild and severe patient outcome. All these changes are at a close genomic positions, G196V and S197L resulting in mild outcome and R203K, G204R, and S194L resulting in inferior outcome. Interestingly, when comparing the S197L (71% of mild outcome) to the S194L (1% chance of a mild outcome) variants, the relative risk was extremely high. Interestingly, the majority of the nucleocapsid phosphoprotein mutations were mapped to a small stretch of amino acids from position 194 to 204. This region coincides with the phosphorylated “RS-motif” [14] which maps onto the intrinsically unstructured serine rich region 181-213 of the protein [15]. Phosphorylation of this site is known to play important roles such as recruitment of host RNA helicase DDX1 which facilitates template readthrough and enables longer subgenomic mRNA synthesis (https://www.uniprot.org/uniprot/P59595). This observation needs further follow-up – especially because the nucleocapsid phosphoprotein is one of the potential drug targets against SARS-CoV-2 [16].

Overall, we have observed more mutations in the structural proteins (spike and nucleocapsid phosphoprotein) than in non-structural proteins. Of note, destabilization mutations in nonstructural proteins were suggested to represent a potential mechanism differentiating SARS-CoV-2 from SARS-CoV [17]. It remains an open question how our results will relate to the recent observation of different molecular architectures of the first and second waves of infection [18].

Researchers from the University of Washington compared two dominant clades of virus in circulation and have observed no difference in outcome when comparing these in patients sufficiently ill to warrant testing for virus [19]. Previously, a 382-nucleotide deletion (Δ382) in the open reading frame 8 was associated with a milder infection [20]. In another recent study, a set of common deletions were identified in the spike protein of SARS-CoV-2 [21]. Other deletions were also validated by RT-PCR [22]. However, due to missing data about insertions and deletions in GISAID we could not evaluate a potential link between deletions and patient outcome.

Importantly, our findings might contain a sampling bias, since only a fraction of the available genomes had patient outcome data. On the other hand, four out of 22 potentially significant mutations (listed in Tables 2 and 3) map to an about 10 amino acid long, functionally important region of the nucleocapsid phosphoprotein which leads us to believe that the current statistical approach can reveal functionally important sites within the COVID 19 genome.

The main limitation of our study results from the database used. Information was retrieved from GISAID, a repository that

Table 3
SARS-CoV-2 mutations correlated to hospitalization and severe outcome in 4,566 patients with available genomic and follow-up information were found in seven distinct genes.

Protein name	Protein mutation	N wild type + mild outcome	N wild type + hospitalized	N wild type + severe outcome	N mutant + mild outcome	N mutant + hospitalized	N mutant + severe outcome	Chi squared test p value
Surface (S) glycoprotein	D614G	382	980	83	326	2326	**469**	1.02E-52
RNA dependent RNA polymerase	P323L	394	1028	43	314	2278	**509**	1.07E-72
ORF3a protein	Q57H	602	2448	358	106	**858**	194	1.09E-15
Nucleocapsid phosphoprotein	R203K	590	2580	309	118	**726**	**243**	2.12E-33
Nucleocapsid phosphoprotein	G204R	594	2586	311	114	720	**241**	1.21E-33
RNA dependent RNA polymerase	A97V	650	2996	546	58	**310**	6	4.1E-10
Nucleocapsid phosphoprotein	P13L	648	3009	547	60	**297**	5	5.54E-10
NSP3	T1198K	654	3010	547	54	**296**	5	5.21E-10
Nucleocapsid phosphoprotein	S194L	705	3067	496	3	239	**56**	2.89E-13
NSP3	S1197R	701	3151	552	7	**155**	0	8.31E-11
ORF3a protein	G251V	700	3182	546	8	**124**	6	1.95E-05
Surface (S) glycoprotein	V1176F	708	3298	443	0	8	**109**	5.2E-162
Nucleocapsid phosphoprotein	I292T	703	3245	515	5	**61**	37	1.06E-13
Surface (S) glycoprotein	L54F	706	3214	543	2	**92**	9	0.000147
ORF6 protein	I33T	703	3247	516	5	**59**	36	2.34E-13
NSP7	L71F	708	3299	499	0	7	**53**	5.53E-73
ORF3a protein	S253P	708	3306	541	0	0	**11**	3.88E-18

contains only general information about patient outcome. The patient treatment protocols resulting in designation into “mild”, “hospitalized” and “severe” cohorts may significantly depend on the country and even on the region where patients were managed. We could also not include potential confounding factors including age, comorbidities and treatment against COVID-19 in our analysis.

Coronaviruses have generally a stable genome which changes very little over time [23]. A fundamental question of SARS-CoV-2 research is whether or not the virus can get weaker or stronger with time. Our findings suggest that there are mutations that can support either of these changes so the theoretical possibility is there that in the future the viral effect will shift towards milder or more severe patient outcomes.

Funding

The research was financed by the 2018-2.1.17-TET-KR-00001 and KH-129581 grants and by the Higher Education Institutional Excellence Programme of the Ministry for Innovation and Technology (MIT) in Hungary, within the framework of the Bionic thematic programme of the Semmelweis University as well as by OTKA grant 12065 provided by MIT, Hungary to Pázmány University.

Ethical Approval

Not required

Declaration of Competing Interests

None

Acknowledgements

The authors wish to acknowledge the support of ELIXIR Hungary (www.elixir-hungary.org) as well as the advice of Drs Sebastian Maurer-Stroh (Bioinformatics Institute, A*STAR, Singapore) and Balázs Ligeti (Pázmány University, Budapest).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijantimicag.2020.106272.

References

[1] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26:450–2. doi:10.1038/s41591-020-0820-9.
[2] Roussel Y, et al. SARS-CoV-2: fear versus data. Int J Antimicrob Agents 2020;55:105947. doi:10.1016/j.ijantimicag.2020.105947.
[3] Giraud-Gatineau A, et al. Comparison of mortality associated with respiratory viral infections between December 2019 and March 2020 with that of the previous year in Southeastern France. Int J Infect Dis 2020;96:154–6. doi:10.1016/j.ijid.2020.05.001.
[4] Grasselli G, et al. Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. JAMA 2020. doi:10.1001/jama.2020.5394.
[5] Wu F, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265–9. doi:10.1038/s41586-020-2008-3.
[6] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A 2020;117:9241–3. doi:10.1073/pnas.2004999117.
[7] Chapman S, Hills G, Watts J, Baulcombe D. Mutational analysis of the coat protein gene of potato virus X: effects on virion morphology and viral pathogenicity. Virology 1992;191:223–30. doi:10.1016/0042-6822(92)90183-p.
[8] Tober C, et al. Expression of measles virus V protein is associated with pathogenicity and control of viral RNA synthesis. J Virol 1998;72:8124–32. doi:10.1128/JVI.72.10.8124-8132.1998.
[9] Williams GD, Chang RY, Brian DA. A phylogenetically conserved hairpin-type 3′ untranslated region pseudoknot functions in coronavirus RNA replication. J Virol 1999;73:8349–55. doi:10.1128/JVI.73.10.8349-8355.1999.
[10] van Dorp L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020;83:104351. doi:10.1016/j.meegid.2020.104351.
[11] Pachetti M, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020;18:179. doi:10.1186/s12967-020-02344-6.
[12] Yurkovetskiy, L. et al. SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. *bioRxiv: the preprint server for biology*, doi:10.1101/2020.07.04.187757 (2020).
[13] Korber B, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 2020. doi:10.1016/j.cell.2020.06.043.

[14] Peng TY, Lee KR, Tarn WY. 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 2008;275:4152–63. doi:10.1111/j.1742-4658.2008.06564.x.

[15] Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–4. doi:10.1093/bioinformatics/bti541.

[16] Yadav R, Imran M, Dhamija P, Suchal K, Handu S. Virtual screening and dynamics of potential inhibitors targeting RNA binding domain of nucleocapsid phosphoprotein from SARS-CoV-2. Journal of biomolecular structure & dynamics 2020:1–16. doi:10.1080/07391102.2020.1778536.

[17] Angeletti S, et al. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. J Med Virol 2020;92:584–8. doi:10.1002/jmv.25719.

[18] Long, S. W. et al. Molecular Architecture of Early Dissemination and Massive Second Wave of the SARS-CoV-2 Virus in a Major Metropolitan Area. *medRxiv*, doi:10.1101/2020.09.22.20199125 (2020).

[19] Nakamichi, K. et al. Outcomes associated with SARS-CoV-2 viral clades in COVID-19. *medRxiv*, doi:10.1101/2020.09.24.20201228 (2020).

[20] Young BE, et al. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet 2020;396:603–11. doi:10.1016/S0140-6736(20)31757-8.

[21] Liu Z, et al. Identification of common deletions in the spike protein of SARS-CoV-2. J Virol 2020. doi:10.1128/JVI.00790-20.

[22] Holland LA, et al. An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). J Virol 2020;94. doi:10.1128/JVI.00711-20.

[23] Vijgen L, Lemey P, Keyaerts E, Van Ranst M. Genetic variability of human respiratory coronavirus OC43. J Virol 2005;79:3223–4 author reply 3224-3225. doi:10.1128/JVI.79.5.3223-3225.2005.