Genetic analysis of SARS-CoV-2 and the common golden nucleotides to human gene

Hamed Babaee (✉ hamedbabaee7@gmail.com)

Payame Noor University <https://orcid.org/0000-0001-7093-6400>

Research Article

Keywords: Coronavirus, COVID-19, SARS-CoV-2, RNA sequencing, Human genome

Posted Date: August 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-854956/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

The coronavirus disease pandemic began in 2019 in Wuhan, China, and continues into 2021, as new mutant viruses appear and require new solutions and treatments.

Methods and Results

In this study, by genetic analysis of data from 24 SARS-CoV-2 samples from different countries and their alignment with each other and the Wuhan reference virus as well as the human genome, the disease factor is looked at from another angle. The result is the identification of genetic differences in viruses, and the finding of a unique 17-nucleotide sequence between human genes, viruses, and enzymes that can contribute to the onset and progression of the disease.

Conclusions

The role of this sequence in DNA replication and the production of new proteins and its alignment with the EPPK1 gene that may cause disease and its various symptoms are likely.

Introduction

In 2019, the world was again struck by a new viral contagious disease called COVID-19 or Coronavirus disease. Apparently, this disease originated from a seafood wholesale market in Wuhan, China from an unknown animal source. In the beginning, China had the highest case report and mortality rates which were mitigated by severe quarantine measures. Nearly every single country has had an epidemic of this virus from that initial timepoint up to today. Currently (April/25/2021), about 147 million cases have been reported globally with the highest case reports in order of USA, India, Brazil, France, Russia, Turkey, England, Italy, Spain; and China as the possible origin of the coronavirus disease ranking 95 in the Worldometer website list without any mortality. Certainly, city and country population density and number, safety, epidemic management, virus mutations, and population genetics, etc. affect the variation in epidemic statistics and mortality. It is also certain that this virus has undergone separate mutations in different countries and has probably adapted to different population genetics of different populations based on its nature and is finding new methods for replicating and initiating disease in each country.

Many detection and treatment methods have been tested and evaluated with most detection methods based on clinical imaging. Multiple companies have produced globally available vaccines that have undergone clinical trials. Some of these vaccines only have positive results in preliminary trials and some have big and small side effects on different people. Research on vaccines and drugs is a work in progress.

SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) is an RNA single positive-strand virus. This substance generally enters human cells by connecting to the Angiotensin-converting enzyme 2

(ACE2) [1, 2]. In addition to the connection and entry of the virus to human cells, its replication and pathogenesis are still incompletely understood. Two possible routes might be taken by the virus after entering the body, either direct engagement in the production of pathogenic proteins for specific organs or indirectly, by connecting virus sequence segments to human sequences and genes they use and transcribe the human genes to change cellular functions or produce new proteins for invading cells and tissues. Finally, this virus is completely dependent on a host like humans and is difficult to grow in laboratory conditions, unlike rhinoviruses. Considering that the coronavirus genome directly acts as an mRNA, it does not need splicing or transcription to be translated. Therefore, these viruses can be translated without an intermediate.

When the RNA of the virus is released into the cell it directly moves to the ribosomes on the endoplasmic reticulum and has no need to enter the nucleus to be transcribed therefore it directly produces its antigens [3]. It is also possible that this virus begins its invasive activities after it has been fixed into a secure position and is protected within the organism.

The SARS-CoV-2 genome has 10 open reading frames (ORF). The virus's structural proteins E, N, M, and S along with other accessory proteins are coded in the ORFs 2–10. Proteins E, M, and S shape the virus coating and the N protein helps pack the RNA genome of the virus. The NSP proteins are a group of coronavirus functional proteins involved in host infection, virus replication, RNA transcription, splicing and modification, protein translation, and synthesis. Proteins like PLpro 3, CLpro, RdRp, and Nsp13-helicase have biologic functions and vital enzyme active sites that make them important targets. NSP1, NSP3C, and ORF7a of the coronavirus disrupt the innate immune system of the host and help evasion of the virus immune system [4]. These proteins increase the infiltration speed of the virus by disrupting the defense system of the host. In some studies, several virus proteins attack the 1- β hemoglobin chain and remove the iron from the porphyrin which reduces the oxygen transferring capabilities of hemoglobin [5]. Additionally, some studies show a significant role of hormonal status on higher mortality rates which explain higher mortality rates and disease intensity in men compared to women [6].

Materials And Methods

In this study 24 FASTA sequences of the SARS-CoV-2 virus were analyzed and evaluated from 21 countries. USA, Iran, and China each have two sequence samples. Sequence samples were downloaded through specific accession numbers within the NCBI Nucleotide database and had deposit dates from April/2020 up to March/2021.

The CLC Genomics Workbench software was used in this experiment to align sample sequences and the human genome for analysis and assessment of the differences and commonalities and finally draw the phylogenetic tree. Sequences length were 29003–29903bp long. All sequences were aligned and evaluated based on the Wuhan reference virus genome (NC_045512.2) which was the longest viral genome. Therefore, to facilitate the identification of nucleotide mutation locations, alignment position order counting was used to obtain uniformity and a single identification method, although the

identification region of each virus mutation is readily observable from its sequence. Usually, differences in sequences length are about 70bp in the 5' and 3' end and the ORFs 1–10 are perfectly aligned except for rare mutations.

Results

Samples have from 2 (France and USA1 variants) up to 853 (China2 variant) mutations and all samples except Wuhan and the USA1 variants have unique and specific mutations. In the following mutations are named such as “T 241 C/T”. The first letter T or D shows the mutation type, i.e., Translocation or Deletion. The following digits show the base number of the mutation within the alignment position order and C/T shows the substitution of cytosine for thymine (Table 1).

The South African and Brazil (MT324062.1, MT350282.1) samples are like the Wuhan variant in length (29903bp) but the South African variant has three translocation mutations that separate it from the Wuhan variant, the Brazil variant has ten translocations with four of them similar to the Wuhan variant. Three C/T mutations are at positions 241, 3037, and 14408 of the alignment positioning (T 241 C/T, T 3037 C/T, T 14408 C/T) and a single A/G substitution (T 23403 A/G). Between all of these variants, only France, South Africa, and Turkey have unique mutations.

Four translocation mutations (T 241 C/T, T 3037 C/T, T 23403 A/G, and T 14408 C/T) of the Wuhan virus exist in ten countries, i.e., Iran, Colombia, Italy, Nepal, Vietnam, India, England, South Korea, Brazil. This probably shows the source of these viruses is from Wuhan while the others are missing these mutations.

The sequences of variants from England, Belarus, the Philippines, the USA 2, and China 2, in addition to translocation mutations, also have deletion mutations. Interestingly the viruses containing deletions are seen in 2021.

The English variant has a 24 nucleotide deletion (D 23598–23621) in gene S. The Belarus virus has two mutations including a 9 nucleotide deletion (D 686–694) in ORF1ab and a 15 nucleotide deletion (D 27764–27778) in ORF7b. The Philippine virus has three mutations including a 9 nucleotide deletion (D 11288–11296) in ORF1ab, and a 6 nucleotide deletion in (D 21766–21771) in ORF7b, and a 3 nucleotide deletion (D 21994–21996) in gene S. The USA variant 2 has a 10 nucleotide deletion (D 80–89) in the 5' and the China variant 2 has five long deletions; a 26 nucleotide deletion (D 27375–27400) in ORF6, a 130 nucleotide deletion (D 27416–27545) in ORF7a, a 332 nucleotide deletion (D 27555–27886) in ORF7a and ORF7b, a 104 nucleotide deletion (D 27908–28011) in ORF8, and a 233 nucleotide deletion (D 28023–28255) in ORF8. An interesting feature of the China variant 2 is the high density of deletions that almost completely remove ORF7b, ORF7a, ORF8. No sample had a deletion in ORF10, gene N, gene M, gene E, or ORF3a. More than 50% of mutations in 2020 are in ORF1ab, while the 2021 samples have less than 50% of the mutations in ORF1ab. Additionally, the number of mutations in 2021 is 2 to 3 times 2020's average. Despite all these mutations, the virus continues to function and has not diminished its aggressiveness.

In the phylogenetic tree (Fig. 1), the differences between the China virus 2 and the Wuhan virus at the beginning and end of the graph from 2019–2021 are obvious. According to the graph, the Wuhan variant is the origin and is shown as a subordinate of several viruses. Additionally, the China virus 2 is a completely new variant compared to all other viruses in other branches of the graph with the Philippine virus on a corresponding branch with different mutations. According to the recent mortality rates in China and the high mutation rates seen in China variant 2 with the deletions of ORF7b, ORF7a, and ORF8 these genes can be considered very important.

Mutations in the functional protein PLpro also exist within the South African variant alignment position sequence (T 5572 T/G), the Columbia variant (T 5298 N/C), the Philippine variant (T 4964 G/A and T 5388 A/C), and the China 2 variant (T 5653 C/T).

The RdRp protein is mutated in the Indian variant (T 14408 C/T, T 16176 C/T) Vietnam, Nepal, Italy, Columbia, Iran 1, Brazil, Iran 2, Wuhan, English, South Korea variants (T 14805 T/C), the Belarus variant (T 15372 T/G) the Philippine variant (T 14676 T/C, T 15279 T/C) and the China 2 variant (T 15165 G/A).

The Nsp13-helicase functional protein is mutated in the Columbia variant (T 17470 T/C), Brazil variant (T 17247 C/T), Philippine variant (T 17615 G/A), and USA 2 variant (T 17014 T/G).

Nsp1 is also mutated in the Spain variant (T 313 T/C), France variant (T 618 G/A). ORF7a (27759 – 27394) is also mutated in the Belarus variant (T 27670 T/G), China 2 variant (T 27412, T 27410 C/T, T 27407 C/T, T 27405 A/T, D 2755–27759, T 27553–27554 C/T, D 27416–27545, T 27413 A/T).

Considering the increase in mutation and contagion we searched for further effective interacting elements. Because the virus genome must interact with other parts of the human genome for replication and pathogenesis and use the human genome for invasive purposes.

In the next part of the analysis, we used RNA sequencing to align 24 virus sequences with the HG38 human genome and we observed that in the genome sequence of all viruses a small sequence is similar to the human genome and aligns to the EPPK1 gene in chromosome 8. This 17 nucleotide sequence (TCCTGCTGCAGATTTGG) containing the PstI restriction site is the unique sequence between all SARS-CoV-2 sample genomes and the hg38 human genome. The sequence between all 24 sample genomes from different countries in gene N of the SARS-CoV-2 virus is almost in the 28500–29500 location and 29474 – 29458 location of the virus sequence alignment position. This sequence is also located in the EPPK1 gene at positions 3961–3977 (Fig. 2).

Discussion

PstI is a type II endonuclease restriction enzyme, Its structure has two parts: A restriction enzyme that cuts external DNA and a methyltransferase that protects endogenous DNA by histone methylation. Combines both defense mechanisms against the invading virus [7]. One can hypothesize that this virus has reverse-engineered this endonuclease's abilities to protect itself. PstI is useful for replicating DNA

because it creates a selective system for producing combinatorial DNA molecules [8]. Perhaps the coronavirus cooperatively functions with the EPPK1 and PstI genes to produce different novel proteins which cause some people to have no or slight symptoms. While others have popular Covid-19 symptoms like fever, dry cough, fatigue, different pains, diarrhea, Conjunctivitis, loss of smell and taste, skin rash, acute respiratory symptoms, and loss of speech and movement.

As seen in the picture of protein expression and the EPPK1 gene RNA a significant number of the coronavirus symptoms are also seen in the protein expression and RNA of EPPK1 (Fig. 3). As demonstrated, the expression of the protein and RNA is high in the digestive tract, muscle tissues, lungs, kidneys, bladder, endocrine system, skin, bone marrow, and lymph tissues. This shows a significant relationship between coronavirus disease and the expression of the EPPK1 gene in the human body. Additionally, the expression of EPPK1 is higher in men's tissues compared to women which follows the trend of higher mortality in men.

EPPK1 is a known human epidermis autoantigen that cooperates with *Plakin* genes in the attachment and anchoring of skeletal fibers with the plasma membrane. The *Plakin* family is a multipurpose organizer of cytoskeleton architecture. It has also been found that they play a role in the uniformity of muscle cells. It was recently discovered that they also exist in the nervous system.

The alignment of the PstI restriction enzyme and the 17-nucleotide region of EPPK1 and the viral sequence may be indicative of a tool or a dangerous weapon based on DNA cloning to use the virus for disrupting cellular skeletal system and disintegration of muscle cells and manipulation of the nervous system that leads to the coronavirus symptoms.

Conclusions

Many treatment methods have been used for Covid-19. Some of the methods are centered around the virus functions and its proteins while others have focused on host receptors and proteins to prevent virus attachment, replication, and pathogenicity. All these methods have come with side effects or do not protect against novel variants. It seems that the human immune system has engineered methods to combat the new variant's epidemics. In this study similarities between the human genome and the virus have been assessed to find possible explanations for the causes of pathogenesis. Certainly, the expression of any phenotype whether in viruses or humans as the superior living creature stems from their genetic sequences and the interaction of their gene products with other organisms. Every issue must first be looked at from a genetics point of view. Because the design and engineering of all life are summed up in the genetic material which also gives rise to the brain, personality, and behavior of beings and their responses to different stimuli.

Declarations

Acknowledgements

Thanks to the Creator of the universe

Conflicts of interest

The author declare that no conflict of interest.

Funding: No Funding/ 'Not applicable'

Availability of data and material (data transparency)

Data analyzed in this study are available in the NCBI's Nucleotide database:

MT263074.1 MT050493.1 MT192772.1 MT072688.1 MT324062.1 MT066156.1 MT256924.2
MT320891.2 MT350282.1 MT359866.1 MT320538.2 MT447177.1 NC_045512.2 MT344948.1
MT304475.1 MW059036.1 MW306668.1 MW321432.1 MW633517.1 MW674675.1
MW735441.1 MW796991.1 MW691153.1 MW822593.1

Code availability (software application)

The CLC Genomics Workbench software (Evaluation version)

Ethics approval (include appropriate approvals or waivers): 'Not applicable'

Consent to participate (include appropriate statements): 'Not applicable'

Consent for publication (include appropriate statements): 'Not applicable'

References

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W et al (February 2020) "A pneumonia outbreak associated with a new coronavirus of probable bat origin". Nature. 579 (7798): 270–273.
Bibcode:2020Natur.579..270Z.doi:10.1038/s41586-020-2012-7. PMC 7095418. PMID 32015507
2. Letko M, Marzi A, Munster V (February 2020) "Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses". Nature Microbiology 5(4):562–569.
doi:10.1038/s41564-020-0688-y. PMC 7095430. PMID 32094589

3. Zhang Y general (2008) Encyclopedia of global health. Los Angeles: Sage Publications. pp. 433–435. ISBN 978-1-4129-4186-0
4. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y et al (February 2020) "Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods". *Acta Pharmaceutica Sinica B* 10(5):766–788 doi:10.1016/j.apsb.2020.02.008. PMC 7102550. PMID 32292689
5. Shirani K, Sheikhabaei E, Torkpour Z et al. (July 2020). "A Narrative Review of COVID-19: The New Pandemic Disease". 10.30476/IJMS.2020.85869.1549
6. Cascella M, Rajnik M, Aleem A et al (April 2021) "Features, Evaluation, and Treatment of Coronavirus (COVID-19) ". In: NCBI Bookshelf. StatPearls [Internet]. StatPearls Publishing, Treasure Island (FL) 7. 10.1073/pnas.78.3.1503. ISSN 0027-8424. PMC 319159. PMID 6262807
8. Bolivar F, Rodriguez RL, Greene PJ, Betlach MC, Heyneker HL, Boyer HW, Crosa JH, Falkow S (1977) "Construction and characterization of new cloning vehicles. II. A multipurpose cloning system". *Gene* 2(2):95–113. doi:10.1016/0378-1119(77)90000-2. PMID 344137

Table

Due to technological limitations, Table 1 is only available as a download in the supplementary files section.

Figures

Unique sequence of 17 nucleotides (TCCTGCTGCAGATTTGG) in the SARS-Cov-2 virus 24 gene of different countries and EPPK1 gene

Figure 3

protein and RNA expression of EPPK1 gene. <https://www.proteinatlas.org/ENSG00000261150-EPPK1/tissue>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table.1.xlsx](#)