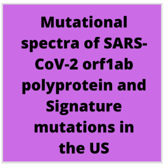
Mutational spectra of SARS-CoV-2 orf1ab polyprotein and Signature mutations in the US

May 29, 2020

Mutational Spectra of SARS-CoV-2 orf1ab polyprotein and Signature mutations in the United States of America

Shuvam Banerjee[1], Sohan Seal[2], Riju Dey[2], Kousik Kr. Mondal[3], Pritha Bhattacharjee[3]

[1]University of Calcutta; UGC DAE CSR Kolkata Centre,

[2]University of Calcutta; Ramakrishna Mission Vidyamandira, Belur Math, Howrah, [3]University of Calcutta

1 Works for me dx.doi.org/10.17504/protocols.io.bgesjtee

Coronavirus Method Development Community

Shuvam Banerjee
University of Calcutta

ABSTRACT

Pandemic COVID-19 outbreak has been caused due to SARS-COV2 pathogen, resulting millions of infection and death worldwide, USA being on top at the present moment. The long, complex orf1ab polyproteins of SARS-COV2 play an important role in viral RNA synthesis. To assess the impact of mutations in this important domain, we analyzed 1134 complete protein sequences of orf1ab polyprotein from NCBI Virus database from affected patients across various states of USA from December 2019 to 25thApril, 2020. Multiple sequence alignment using Clustal Omega followed by statistical significance was calculated. Four significant mutations T265I (nsp 2), P4715L (nsp 12) and P5828L and Y5865C (both at nsp 13) were identified in important non-structural proteins, which function either as replicase or helicase. A comparative analysis shows 265T>I, 5828P>L and 5865Y>C are unique to USA and not reported from Europe or Asia; while one, 4715P>L is predominant in both Europe and USA. Mutational changes in amino acids are predicted to alter structure and function of corresponding proteins, thereby it is imperative to consider the mutational spectra while designing new antiviral therapeutics targeting viral orf1ab.

EXTERNAL LINK

https://www.biorxiv.org/content/10.1101/2020.05.01.071654v1.full

ATTACHMENTS

2020.05.01.071654v1.full
.pdf

MATERIALS TEXT

Protein sequences were retrieved from *NCBI Virus* database
Clustal Omega
Ancestral orf1ab polyprotein sequence of Wuhan (YP_009742608)
UCSF Chimera and *PyMOL*

SAFETY WARNINGS

Please refer to the Safety Data Sheets (SDS) for health and environmental hazards.

BEFORE STARTING

SARS-CoV-2 is the responsible pathogen for pandemic COVID 19. Positive-stranded, RNA genomes of Coronaviruses is around 27 to 32-kb in length, of which about 2/3rdencompasses viral Orf1ab gene and expresses the largest and most complex polyproteins of any RNA viruses. The open reading frame 1 (ORF1), functions as replicase, replicase/transcriptase, or polymerase polyproteins, is translated into ORF1a (approximately 486kDa, major product) and ORF1b (~306KDa) polyproteins in the host cell. Virus-encoded proteinases including Papain like protease (PLPs) and 3C like protease (3CL Pro) cleaves ORF1 into 16 nonstructural proteins (nsps). ORF1a comprising nsps (nsp 1 to nsp 10) play an important role in coping cellular stresses and maintaining functional integrity of the cellular components along with the pivotal roles in viral replication. On the other hand, ORF1b encodes viral RNA-dependent RNA polymerase (nsp12), helicase (nsp13), exonuclease (nsp14), a polyU (Uridylate) specific endonuclease (nsp15), and methyl transferase (nsp16). Hence, majority of these nsps play an important role in viral pathogenesis and promising target for anti-viral drug targeting and vaccine synthesis.

Citation: Shuvam Banerjee, Sohan Seal, Riju Dey, Kousik Kr. Mondal, Pritha Bhattacharjee (05/29/2020). Mutational Spectra of SARS-CoV-2 orf1ab polyprotein and Signature mutations in the United States of America. https://dx.doi.org/10.17504/protocols.io.bgesjtee

At present, USA is one of the worst affected countries globally in terms of affected individuals and the number of death, is concerned. Till 25th April, USA is reported to have 860,772 positive cases and 44,053 deaths. This adverse condition led to investigate the sequence of viral whole genome reported to NCBI virus database. As on 25th April 2020 (till 12 noon, IST), around 1134 Orf1ab polyprotein sequence have been submitted from USA alone. Different states of USA like Washington D.C, New York, Connecticut, Idaho, Georgia etc have uploaded sequence of Orf1ab polyprotein in the database. Since COVID-19 originated in Wuhan and found to be extended to different parts of the globe with variation in its virulence, it is imperative to identify the mutations occurred in Orf1ab polyprotein and consequent impact in protein structure and interaction with the host body. Hence the present study aims (i) to identify the mutations observed in orf1ab polyprotein, (ii) to predict the conformational changes of SARS-CoV-2 polyprotein due to the mutations and (iii) to identify the signature pattern, if any for USA.

ORF1ab protein sequence retrieval from the database

1 Retrieve the protein sequences through the *NCBI Virus* database.

> Specific input was "SARS-CoV-2"

2 Refine the output with a sequence length of 7100 from the initial 7050.

> The length of the target orf1ab polyprotein is 7096.

3 In total, 1307 sequences will be retrieved, among them 125 sequences were from Asia, 42 from Europe and 1134 solely from USA.

Screening of submitted sequences and selection of study sample

4 Eliminate all incomplete sequences or sequences with undermined residues (mentioned as X).

5 In total, 867 orf1ab polyprotein complete sequences deposited from 31 different states within the US were considered for the study (Group A).

> All the above-mentioned sequences from Asia (Group B) and Europe (Group C) were also selected.

Multiple Sequence Alignment (MSA) and analysis of mutational spectra

6 Align multiple sequences of each of the above-mentioned groups using *Clustal Omega*.

7 Subdivide sequences in Group A into 31 subgroups depending on the regional source from which the sequence was originated.

Citation: Shuvam Banerjee, Sohan Seal, Riju Dey, Kousik Kr. Mondal, Pritha Bhattacharjee (05/29/2020). Mutational Spectra of SARS-CoV-2 orf1ab polyprotein and Signature mutations in the United States of America. https://dx.doi.org/10.17504/protocols.io.bgesjtee

8 MSA was conducted encompassing all subgroups using ancestral orf1ab polyprotein sequences of Wuhan (YP_009742608, comprising of 7096 residues) as reference sequence.

9 Set gap opening penalty and gap extension penalty at 12 and 2, respectively, to ensure that unnecessary gaps are not created during alignment and alterations are visualised easily.

10 Thoroughly screen alignment results to determine the exact locaitons of the mutations and alteration linked to that position.

Calculation of statistical significance to detect signature mutations of USA

11 Calculate the number of occurance of each mutated variant in Group A, B, and C.

12 Divide by the total number of sequence submitted under that group.

> The proportion of each variant in Group A would be a/x and likewise for the other groups.

13 Caluclate the p-value through Z-score using the proportion of the mutant and sample size to establish whether the occurance of that mutant variant in USA (Group A) is significantly higher in comparison to Asia and Europe (Group B and C).

> Two-tailed p values were calculated using 0.05 as the significance level. Thus attempt has been made to identify the signature mutations in the region of orf1ab polyprotein.

Homology Modelling and simulation of protein structure

14 Report structures of the associated non-structural proteins for wild type at *I-Tasser* server.

> The structures were not available for the varied amino acid alteration.

15 To identify the alteration, generate secondary structure by using the Homology Modelling method by *I-Tasser*.

16 Superimpose the alteration with the wild type using *UCSF Chimera* and *PyMOL* for easy visualization and comparison.

Citation: Shuvam Banerjee, Sohan Seal, Riju Dey, Kousik Kr. Mondal, Pritha Bhattacharjee (05/29/2020). Mutational Spectra of SARS-CoV-2 orf1ab polyprotein and Signature mutations in the United States of America. https://dx.doi.org/10.17504/protocols.io.bgesjtee