Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag

Prashant Pradhan[$1,2], Ashutosh Kumar Pandey[$1], Akhilesh Mishra[$1], Parul Gupta[1], Praveen Kumar Tripathi[1], Manoj Balakrishnan Menon[1], James Gomes[1], Perumal Vivekanandan*[1]and Bishwajit Kundu*[1]

[1]Kusuma School of biological sciences, Indian institute of technology, New Delhi-110016, India.

[2]Acharya Narendra Dev College, University of Delhi, New Delhi-110019, India

[$]Equal contribution

* Corresponding authors- email: bkundu@bioschool.iitd.ac.in

vperumal@bioschool.iitd.ac.in

Abstract:

We are currently witnessing a major epidemic caused by the 2019 novel coronavirus (2019-nCoV). The evolution of 2019-nCoV remains elusive. We found 4 insertions in the spike glycoprotein (S) which are unique to the 2019-nCoV and are not present in other coronaviruses. Importantly, amino acid residues in all the 4 inserts have identity or similarity to those in the HIV-1 gp120 or HIV-1 Gag. Interestingly, despite the inserts being discontinuous on the primary amino acid sequence, 3D-modelling of the 2019-nCoV suggests that they converge to constitute the receptor binding site. The finding of 4 unique inserts in the 2019-nCoV, all of which have identity /similarity to amino acid residues in key structural proteins of HIV-1 is unlikely to be fortuitous in nature. This work provides yet unknown insights on 2019-nCoV and sheds light on the evolution and pathogenicity of this virus with important implications for diagnosis of this virus.

Introduction

Coronaviruses (CoV) are single-stranded positive-sense RNA viruses that infect animals and humans. These are classified into 4 genera based on their host specificity: *Alphacoronavirus, Betacoronavirus, Deltacoronavirus and Gammacoronavirus* (Snijder et al., 2006). There are seven known types of CoVs that includes 229E and NL63 (Genus Alphacoronavirus), OC43, HKU1, MERS and SARS (Genus Betacoronavirus). While 229E, NL63, OC43, and HKU1 commonly infect humans, the SARS and MERS outbreak in 2002 and 2012 respectively occurred when the virus crossed-over from animals to humans causing significant mortality (J. Chan et al., n.d.; J. F. W. Chan et al., 2015). In December 2019, another outbreak of coronavirus was reported from Wuhan, China that also transmitted from animals to humans. This new virus has been temporarily termed as 2019-novel Coronavirus (2019-nCoV) by the World Health Organization (WHO) (J. F.-W. Chan et al., 2020; Zhu et al., 2020). While there are several hypotheses about the origin of 2019-nCoV, the source of this ongoing outbreak remains elusive.

The transmission patterns of 2019-nCoV is similar to patterns of transmission documented in the previous outbreaks including by bodily or aerosol contact with persons infected with the virus.

Cases of mild to severe illness, and death from the infection have been reported from Wuhan. This outbreak has spread rapidly distant nations including France, Australia and USA among others. The number of cases within and outside China are increasing steeply. Our current understanding is limited to the virus genome sequences and modest epidemiological and clinical data. Comprehensive analysis of the available 2019- nCoV sequences may provide important clues that may help advance our current understanding to manage the ongoing outbreak.

The spike glycoprotein (S) of cornonavirus is cleaved into two subunits (S1 and S2). The S1 subunit helps in receptor binding and the S2 subunit facilitates membrane fusion (Bosch et al., 2003; Li, 2016). The spike glycoproteins of coronoviruses are important determinants of tissue tropism and host range. In addition the spike glycoproteins are critical targets for vaccine development (Du et al., 2013). For this reason, the spike proteins represent the most extensively studied among coronaviruses. We therefore sought to investigate the spike glycoprotein of the 2019-nCoV to understand its evolution, novel features sequence and structural features using computational tools.

Methodology

Retrieval and alignment of nucleic acid and protein sequences

We retrieved all the available coronavirus sequences (n=55) from NCBI viral genome database (https://www.ncbi.nlm.nih.gov/) and we used the GISAID (Elbe & Buckland-Merrett, 2017)[https://www.gisaid.org/] to retrieve all available full-length sequences (n=28) of 2019-nCoV as on 27 Jan 2020. Multiple sequence alignment of all coronavirus genomes was performed by using MUSCLE software (Edgar, 2004) based on neighbour joining method. Out of 55 coronavirus genome 32 representative genomes of all category were used for phylogenetic tree development using MEGAX software (Kumar et al., 2018). The closest relative was found to be SARS CoV. The glycoprotein region of SARS CoV and 2019-nCoV were aligned and visualized using Multalin software (Corpet, 1988). The identified amino acid and nucleotide sequence were aligned with whole viral genome database using BLASTp and BLASTn. The conservation of the nucleotide and amino acid motifs in 28 clinical variants of 2019-nCoV genome were presented by performing multiple sequence alignment using MEGAX software. The three dimensional structure of 2019-nCoV glycoprotein was generated by using SWISS-MODEL online server (Biasini et al., 2014) and the structure was marked and visualized by using PyMol (DeLano, 2002).

Results

Uncanny similarity of novel inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag

Our phylogentic tree of full-length coronaviruses suggests that 2019-nCoV is closely related to SARS CoV [Fig1]. In addition, other recent studies have linked the 2019-nCoV to SARS CoV. We therefore compared the spike glycoprotein sequences of the 2019-nCoV to that of the SARS CoV (NCBI Accession number: AY390556.1). On careful examination of the sequence alignment we found that the 2019- nCoV spike glycoprotein contains 4 insertions [Fig.2]. To further investigate if these inserts are present in any other corona virus, we performed a multiple

sequence alignment of the spike glycoprotein amino acid sequences of all available coronaviruses (n=55) [refer Table S.File1] in NCBI refseq (ncbi.nlm.nih.gov) this includes one sequence of 2019-nCoV[Fig.S1]. We found that these 4 insertions [inserts 1, 2, 3 and 4] are unique to 2019-nCoV and are not present in other coronaviruses analyzed. Another group from China had documented three insertions comparing fewer spike glycoprotein sequences of coronaviruses . Another group from China had documented three insertions comparing fewer spike glycoprotein sequences of coronaviruses (Zhou et al., 2020).

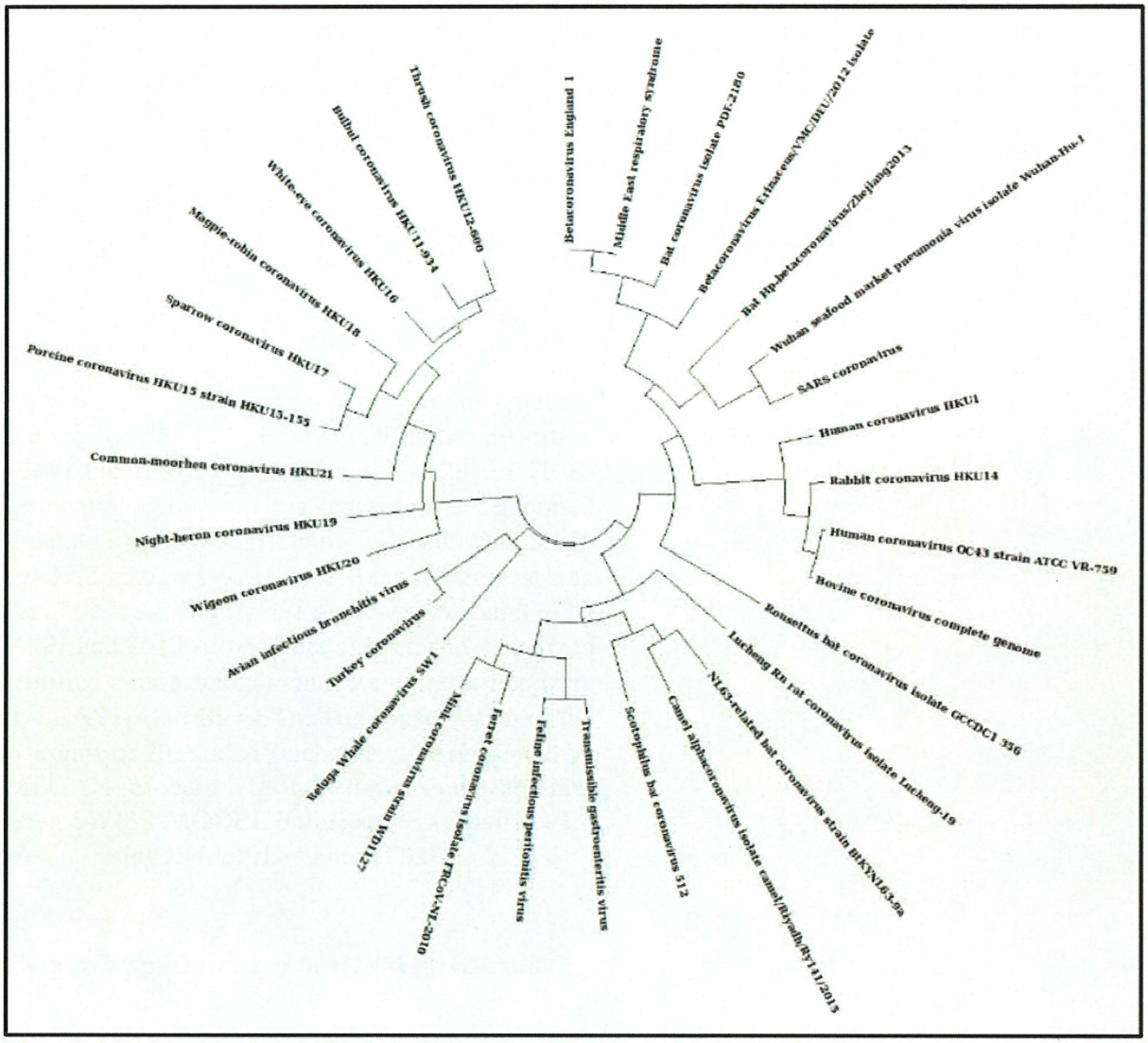

Figure 1: Maximum likelihood genealogy show the evolution of 2019- nCoV: The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model. The tree with the highest log likelihood (12458.88) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood

value. This analysis involved 5 amino acid sequences. There were a total of 1387 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.

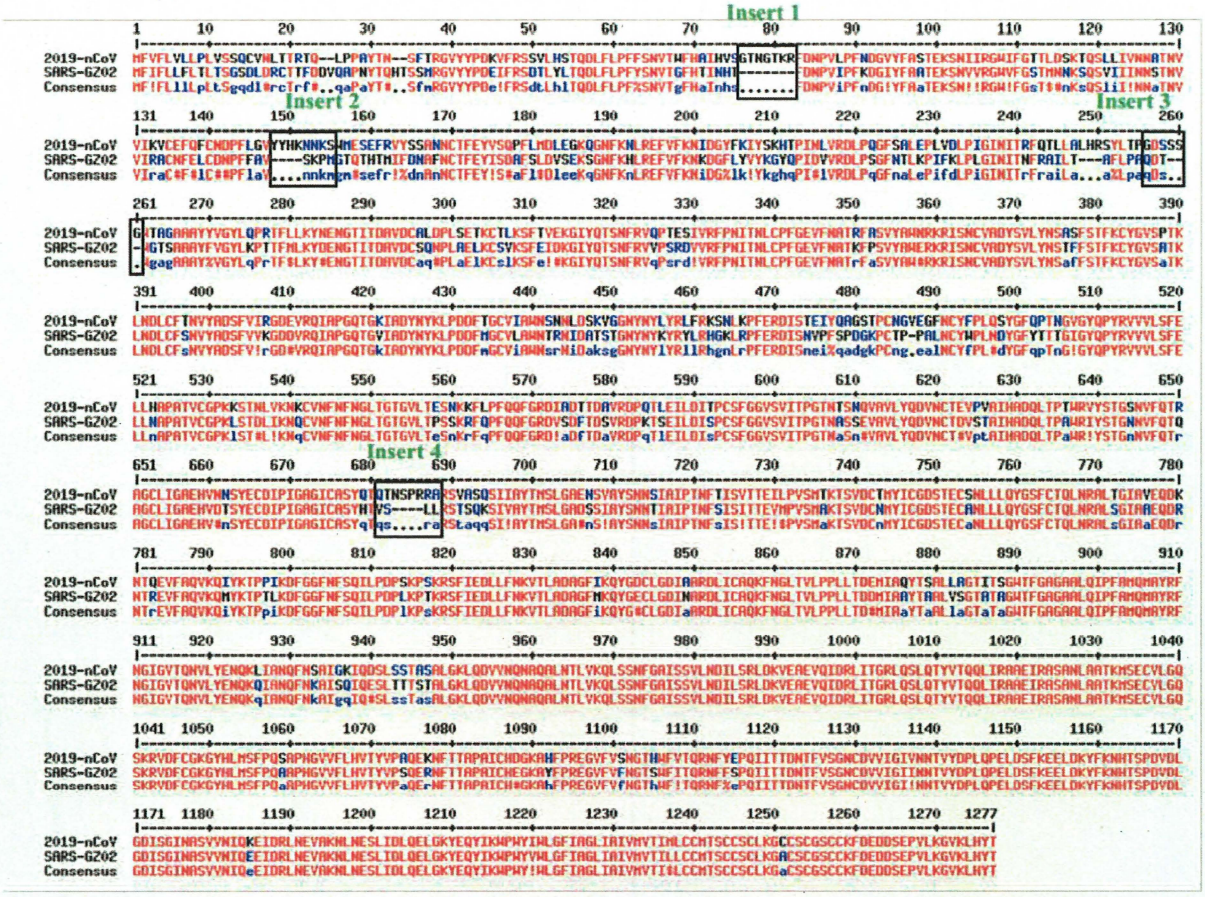

Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS. The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.

We then analyzed all available full-length sequences (n=28) of 2019-nCoV in GISAID (Elbe & Buckland-Merrett, 2017) as on January 27, 2020 for the presence of these inserts. As most of these sequences are not annotated, we compared the nucleotide sequences of the spike glycoprotein of all available 2019-nCoV sequences using BLASTp. Interestingly, all the 4 insertions were absolutely (100%) conserved in all the available 2019- nCoV sequences analyzed [Fig.S2, Fig.S3].

We then translated the aligned genome and found that these inserts are present in all Wuhan 2019-nCoV viruses except the 2019-nCoV virus of Bat as a host [Fig.S4]. Intrigued by the 4 highly conserved inserts unique to 2019-nCoV we wanted to understand their origin. For this purpose, we used the 2019-nCoV local alignment with each insert as query against all virus genomes and considered hits with 100% sequence coverage. Surprisingly, each of the four inserts aligned with short segments of the Human immunodeficiency Virus-1 (HIV-1) proteins. The amino acid positions of the inserts in 2019-nCoV and the corresponding residues in HIV-1 gp120 and HIV-1 Gag are shown in Table 1. The first 3 inserts (insert 1,2 and 3) aligned to short segments of amino acid residues in HIV-1 gp120. The insert 4 aligned to HIV-1 Gag. The insert 1 (6 amino acid residues) and insert 2 (6 amino acid residues) in the spike glycoprotein of 2019-nCoV are 100% identical to the residues mapped to HIV-1 gp120. The insert 3 (12 amino acid residues) in 2019-nCoV maps to HIV-1 gp120 with gaps [see Table 1]. The insert 4 (8 amino acid residues) maps to HIV-1 Gag with gaps.

Although, the 4 inserts represent discontiguous short stretches of amino acids in spike glycoprotein of 2019-nCoV, the fact that all three of them share amino acid identity or similarity with HIV-1 gp120 and HIV-1 Gag (among all annotated virus proteins) suggests that this is not a random fortuitous finding. In other words, one may sporadically expect a fortuitous match for a stretch of 6-12 contiguous amino acid residues in an unrelated protein. However, it is unlikely that all 4 inserts in the 2019-nCoV spike glycoprotein fortuitously match with 2 key structural proteins of an unrelated virus (HIV-1).

The amino acid residues of inserts 1, 2 and 3 of 2019-nCoV spike glycoprotein that mapped to HIV-1 were a part of the V4, V5 and V1 domains respectively in gp120 [Table 1]. Since the 2019-nCoV inserts mapped to variable regions of HIV-1, they were not ubiquitous in HIV-1 gp120, but were limited to selected sequences of HIV-1 [refer S.File1] primarily from Asia and Africa.

The HIV-1 Gag protein enables interaction of virus with negatively charged host surface (Murakami, 2008) and a high positive charge on the Gag protein is a key feature for the host-virus interaction. On analyzing the pI values for each of the 4 inserts in 2019-nCoV and the corresponding stretches of amino acid residues from HIV-1 proteins we found that a) the pI values were very similar for each pair analyzed b) most of these pI values were 10±2 [Refer Table 1] . Of note, despite the gaps in inserts 3 and 4 the pI values were comparable. This uniformity in the pI values for all the 4 inserts merits further investigation.

As none of these 4 inserts are present in any other coronavirus, the genomic region encoding these inserts represent ideal candidates for designing primers that can distinguish 2019-nCoV from other coronaviruses.

Motifs	Virus Glycoprotein	Motif Alignment	HIV protein and Variable region	HIV Genome Source Country/ subtype	Number of Polar Residues	Total Charge	pI Value
Insert 1	2019- nCoV (GP) HIV1(GP120)	**71 76** **TNGTKR** **TNGTKR** **404 409**	gp120-V4	Thailand */ CRF01_AE	5 5	2 2	11 11
Insert 2	2019- nCoV (GP) HIV1(GP120)	**145 150** **HKNNKS** **HKNNKS** **462 467**	gp120-V5	Kenya*/ G	6 6	2 2	10 10
Insert 3	2019- nCoV (GP) HIV1(GP120)	**245 256** **RSYL----TPGDSSSG** **RTYLFNETRGNSSSG** **136 150**	gp120-V1	India*/C	8 10	2 1	10.84 8.75
Insert 4	2019- nCoV (Poly P) HIV1(gag)	**676 684** **QTNS----------PRRA** **QTNSSILMQRSNFKG PRRA** **366 384**	Gag	India*/C	6 12	2 4	12.00 12.30

Table 1: Aligned sequences of 2019-nCoV and gp120 protein of HIV-1 with their positions in primary sequence of protein. All the inserts have a high density of positively charged residues. The deleted fragments in insert 3 and 4 increase the positive charge to surface area ratio. *please see Supp. Table 1 for accession numbers

The novel inserts are part of the receptor binding site of 2019-nCoV

To get structural insights and to understand the role of these insertions in 2019-nCoV glycoprotein, we modelled its structure based on available structure of SARS spike glycoprotein (PDB: 6ACD.1.A). The comparison of the modelled structure reveals that although inserts 1,2 and 3 are at non-contiguous locations in the protein primary sequence, they fold to constitute the part of glycoprotein binding site that recognizes the host receptor (Kirchdoerfer et al., 2016) (Figure 4). The insert 1 corresponds to the NTD (N-terminal domain) and the inserts 2 and 3 correspond to the CTD (C-terminal domain) of the S1 subunit in the 2019-nCoV spike glycoprotein. The insert 4 is at the junction of the SD1 (sub domain 1) and SD2 (sub domain 2) of the S1 subunit (Ou et al., 2017). We speculate, that these insertions provide additional flexibility to the glycoprotein binding site by forming a hydrophilic loop in the protein structure that may facilitate or enhance virus-host interactions.

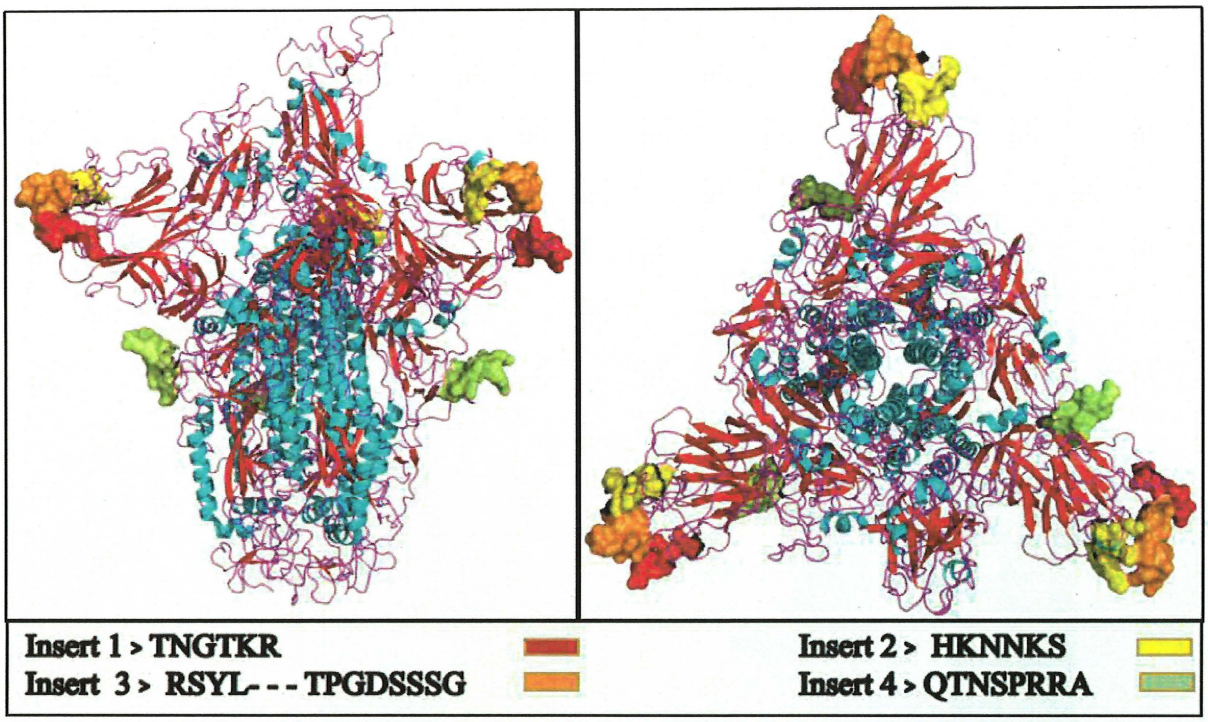

Figure 3. Modelled homo-trimer spike glycoprotein of 2019-nCoV virus. The inserts from HIV envelop protein are shown with colored beads, present at the binding site of the protein.

Evolutionary Analysis of 2019-nCoV

It has been speculated that 2019-nCoV is a variant of Coronavirus derived from an animal source which got transmitted to humans. Considering the change of specificity for host, we decided to study the sequences of spike glycoprotein (S protein) of the virus. S proteins are surface proteins that help the virus in host recognition and attachment. Thus, a change in these proteins can be reflected as a change of host specificity of the virus. To know the alterations in S protein gene of 2019-nCoV and its consequences in structural re-arrangements we performed *in-sillico* analysis of 2019-nCoV with respect to all other viruses. A multiple sequence alignment between the S protein amino acid sequences of 2019-nCoV, Bat-SARS-Like, SARS-GZ02 and MERS revealed that S protein has evolved with closest significant diversity from the SARS-GZ02 (Figure 1).

Insertions in Spike protein region of 2019-nCoV

Since the S protein of 2019-nCoV shares closest ancestry with SARS GZ02, the sequence coding for spike proteins of these two viruses were compared using MultiAlin software. We found four new insertions in the protein of 2019-nCoV- "GTNGTKR" (IS1), "HKNNKS" (IS2), "GDSSSG" (IS3) and "QTNSPRRA" (IS4) (Figure 2). To our surprise, these sequence insertions were not only absent in S protein of SARS but were also not observed in any other member of the *Coronaviridae* family (Supplementary figure). This is startling as it is quite unlikely for a virus to have acquired such unique insertions naturally in a short duration of time.