Two Short Low Complexity Regions (LCRs) are Hallmark Sequences of the Delta SARS-CoV- 2 Variant Spike Protein

Arturo Becerra

Universidad Nacional Autónoma de México

Israel Muñoz-Velasco

Universidad Nacional Autónoma de México

Abelardo Aguilar-Cámara

Universidad Nacional Autónoma de México

Wolfgang Cottom-Salas

Universidad Nacional Autónoma de México

Adrián Cruz-González

Universidad Nacional Autónoma de México

Alberto Vázquez-Salazar

University of California

Ricardo Hernández-Morales

Universidad Nacional Autónoma de México

Rodrigo Jácome

Universidad Nacional Autónoma de México

José Alberto Campillo-Balderas

Universidad Nacional Autónoma de México

Antonio Lazcano (✉ alar@ciencias.unam.mx)

Universidad Nacional Autónoma de México

Research Article

Keywords: Low complexity regions (LCRs), SARS-CoV-2, variants of concern, Delta variant, spike protein

Posted Date: August 6th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-785282/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published at Scientific Reports on January 18th, 2022.
See the published version at <https://doi.org/10.1038/s41598-022-04976-8>.

Abstract

Low complexity regions (LCRs) are protein sequences formed by a set of compositionally biased residues. LCRs are extremely abundant in cellular proteins and have also been reported in viruses, where they may partake in evasion of the host immune system. Here we report four novel extremely conserved LCRs in the spike proteins of several SARS-CoV-2 variants. With the exception of Iota, where it is absent, the Spike LCR-1 is present in the signal peptide of 80.57% of the Delta variant, and in other variants of concern and interest. The Spike LCR-2 is highly prevalent (79.87%) in Iota. Two distinctive LCRs are present in the Delta spike protein. The Delta Spike LCR-3 is present in 99.19% of the analyzed sequences, and the Delta Spike LCR-4 in 98.3% of the same set of proteins. These two LCRs are present in the furin cleavage site and HR1 domain, respectively, and may be considered hallmark traits of the Delta variant. The presence of two medically-important point mutations, P681R and D950N, in these LCRs, combined with the extraordinary conservation of these regions, suggests that they may be a trait associated with the rapid spread of the highly contagious Delta lineage.

Introduction

Protein segments that exhibit a bias in their composition can be formed by (a) a small number of different amino acids, in which case they are called low complexity regions (LCRs); or (b) homopolymers or homorepeats, if they consist of a long repetition of a single amino acid^{1,2}. They tend to be more prevalent in proteins associated with polysaccharide-, ion-, and nucleic acid binding, as well as in phospholipid interaction, transcription, translation, and folding functions³. It is estimated that approximately 0.4% of eukaryotic proteomes are LCRs, which is up to 23 times higher than in prokaryotes³.

LCRs emergence has been associated with replication slippage and the formation of microsatellites during genome replication or recombination events⁴. The regions of the proteins in which the LCRs are located evolve rapidly, but there is an ongoing debate whether they change neutrally or under selective pressures⁵. Given the immunological significance of pathogens' surface proteins in which many LCRs are located^{6,7,9}, it is somewhat surprising that little attention has been given to their presence in viral proteomes. In *sensu stricto*, the presence and location of LCRs has only been reported in the HIV-1⁸ and, more recently, in SARS-CoV-2¹⁰. They are rather abundant in the HIV-1 gp120 protein, and over 30% of them are located in the hypervariable regions of the connecting loops present in the protein, where they probably play a role in immune escape⁸. LCRs are scattered throughout the SARS-CoV-2 proteome, but are conspicuously absent in some proteins of the replication-transcription complex (RdRp, helicase, and NSP14 exonuclease), and in the NSP1, 3CL protease, NSP9-11, NSP15, ORF3a, membrane (M) protein, ORF6, ORF8 and ORF10 proteins¹⁰. This indicates that these proteins are under strong purifying selection, and that any compositional change is rapidly selected against. In SARS-CoV-2 LCRs are more prevalent in the non-structural protein 3, spike protein, and the nucleocapsid protein, where they may simultaneously enhance immune evasion and induce a strong immunogenic response¹⁰.

In addition to LCRs, the presence of microsatellites, or nucleotide simple sequence repeats (SSR), have also been reported in viral genomes. Like LCRs, SSRs have also been associated with increased adaptability, as well as with enhanced recombination rates and indel generation in both cells and in viruses^{11–14}. A strong correlation has been found between viral genome size and the number of microsatellites present¹⁵. Viral SSRs are present in both RNA- and DNA viruses, including DNA mycobacteriophages¹³, economically relevant plant viruses such as potyviruses¹⁶, tobamoviruses¹⁷ and geminiviruses¹⁸, as well as in medically important viruses like herpesviruses¹⁹, HIV-1²⁰ and filoviruses²¹. More recently, several SSRs rich in hexameric repeats have been identified in the SARS-CoV-2 genome^{22,23}. They appear to be more prevalent in the ORF1ab, S, ORF3a, N and ORF7a of the SARS-CoV-2 genic regions²³, which opens the possibility that these regions may exhibit a trend towards greater genetic variability^{10,22,23}. In this work, we report the conservation and variability of LCRs in several SARS-CoV-2 variants of concern (VOC) and interest (VOI) using comparative proteomics and protein structure analyses. We have identified three previously unreported LCRs that are present only in some VOIs and VOCs (Fig. 1). Quite significantly, these LCRs do not exhibit a random distribution. Two of them are located in highly conserved positions of the spike S1 and S2 subunits of the extremely contagious Delta variant. Our results demonstrate that these two conserved (99 – 98%) short LCRs are hallmark sequences of the highly transmissible Delta SARS-CoV-2 variant, which suggest that they may play a significant role in the viral adaptation and rapid spread of this VOC.

Results

In this work, we have analyzed a total of 28,231 SARS-CoV-2 whole proteome sequences to search for LCRs. As summarized in Figs. 1, 2 and Figure S1, our results indicate that most of the LCRs are present in the viral reference genome and its variants. However, we detected conserved important differences in the prevalence of these LCRs between the SARS-CoV-2 VOCs and VOIs viral proteomes.

As shown in Figs. 1 and S1, the LCR formed by the sequence FVFLVLLPLV is present between residues 2 and 11 of all the spike protein variants¹⁰ except for the Iota variant. Here we report three previously undescribed, highly prevalent, short, specific LCRs in the spike proteins of the Delta-, Iota-, and Kappa variants (Figs. 1, 2 and Figure S1). In this work we have named each LCR according to the following rules: the first word of the name corresponds to the protein in which the LCR is located, and the number corresponds to its position in each of the SARS-CoV-2 proteins. Figure 3a displays the actual location of these LCRs in a spike protein 3D structure (PDB ID: 7BNM). The LCR, which we have named Spike LCR-2 (Fig. 3), is located between the residues 252 and 264 of the N-terminal domain (NTD) of the Iota variant spike protein (Fig. 3c). The sequence of this LCR is GGSSSGWTAGAAA (Fig. 3b and Table 1), and it is present in 79.87% of the Iota variants. In contrast, this LCR is completely absent in the Eta-, and Kappa variants (Figure S1), and its prevalence in other VOCs, VOIs and other SARS-CoV-2 samples is below 3% (Figure S1).

Table 1

Spike proteins LCRs reported in this work for the Delta, Iota and Kappa variants. The position of each LCR in the spike sequence is indicated. The first letter of the name corresponds to the protein in which the LCR is located, and the number corresponds to its position in each of the SARS-CoV-2 proteins.

Protein	Sequence	~ Start position	~ End position	Variants	Name
Spike	FVFLVLLPLV	2	11	All (except for Iota)	Spike LCR-1
Spike	GGSSSGWTAGAAA	252	264	Iota	Spike LCR-2
Spike	SRRRARSVASQSIIA	680	694	Delta, Kappa	Spike LCR-3
Spike	LQNVVNQNAQALN	948	960	Delta	Spike LCR-4

The Spike LCR-3 (Delta-Kappa prevalent) is positioned between residues 680 and 694 in the Delta- and Kappa spike protein variants (Fig. 3b). Its sequence is the polybasic, conserved 15 amino acid segment SRRRARSVASQSIIA (Table 1), and it is located precisely in the furin cleavage site, which is present in the S1 C-terminus, whose tertiary structure has not been visualized due to its inherent flexibility²⁶ (Fig. 3c). This LCR is found in 99.19% of the Delta variants (Figure S1). As shown in Fig. 2, the complexity value of this region in the Delta- and Kappa variants is significantly lower in comparison with the rest of the protein; however, a small number of the Delta sequences (39/4830) do not surpass our cutoff value due to the presence of amino acid substitutions that raise the complexity value of the regions.

The third LCR, or Spike LCR-4, has the conserved 13-aa polar-rich sequence LQNVVNQNAQALN, and is located between residues 946 and 958 of the spike protein of the highly transmissible Delta variant. It is found in an alpha-helix rich domain (HR1) (Fig. 3b) that is part of the spike protein S2 stalk region (Fig. 3a). This low complexity region is present in every Delta variant; only 1.7% of them do not surpass the LCR cutoff value defined here (Fig. 1, Figure S1). In the Beta-, Eta-, and Kappa variants analyzed here, this LCR is completely absent (Figure S1), whereas in the VOCs, VOIs, and "Others" SARS-CoV-2 categories, its prevalence is below 2% (Figure S1).

In the Alpha variant sequences analyzed here, the NSP3 LCR-3 (Figure S1) is missing in 98.95% of the proteomes. The Lambda NSP3 LCR-4 (Figure S1) is absent in 98.84% of the analyzed proteomes.

Discussion And Conclusions

The presence of LCRs is a trait of a broad spectrum of proteins, and appears to contribute to the antigenic variability in both viral and cellular pathogen populations. It is generally agreed that their origin is associated with errors due to polymerase slippage during genome replication and/or unequal crossover recombination events^{27,28}. The processes that lead to their preservation in highly streamlined genomes, such as those of most RNA viruses, are not well understood, and their tempo and mode of evolution remain open issues. However, the extraordinary conservation of the two small LCRs reported here (Spike LCR-3,

Spike LCR-4) in the rapidly-spreading Delta variant suggests that they are part of its hallmark traits. Accordingly, a detailed analysis of their frequency and phenotypic significance may contribute to the understanding of the origin of this variant's increased transmissibility. A possible new variant of Delta, known as Delta Plus, was recently reported by the Public Health England service in a June 11, 2021 bulletin. This variant has a K417N mutation in the receptor-binding domain (RBD) of the spike protein^{30,31}. Although the Delta Plus variant is considered a VOC, its properties are still being investigated. Our analysis of the spike protein of this putative variant shows that it has the same LCRs present in the SARS-CoV-2 Delta spike protein itself.

The Spike LCR-1 (FVFLVLLPLV) is a highly hydrophobic region that consists of helix-forming residues, including phenylalanine, valine and leucine, and it is the major component of the signal peptide (amino acids 1–13) located upstream of the N-terminus domain³² (Fig. 3). In the lumen of the endoplasmic reticulum this signal peptide plays a key role in guiding the spike protein to its membrane location by cellular signal peptidases³³.

As noted above, the Kappa/Delta Spike LCR-3 and the Delta Spike LCR-4 regions are located in the spike S1 and S2 subunits, respectively. The mutation P681R detected in the Spike LCR-3 (SRRRARSVASQSIIA) (Fig. 3) at the furin cleavage site increases the polybasic nature of this region, which could augment its affinity with the furin protease³⁴. *In vitro* experiments and SARS-CoV-2 infections in animal models have demonstrated that the P681R mutation enhances both the fusogenicity and pathogenicity of the virus^{35,36}. The phylogenetic affinities between the Kappa- and Delta variants, both of which are part of the lineage B.1.617^{37,38} may explain in part the presence of these two mutations in the Delta- and the Kappa Spike LCR-3 (Figs. 1 and S1).

The SARS-CoV-2 spike protein is endowed with two heptad repeat motifs (HR1 and HR2) in its ectodomain that are involved in cell fusion, which is a key step in viral entry^{39,40}. The Spike LCR-4 (LQNVVNQNAQALN) includes charged-neutral, polar (asparagine and glutamine) and hydrophobic amino acids (leucine, valine, and alanine), which are typical of heptad repeat motifs. The interaction of HR1 and HR2 leads to the formation of a six-helical bundle that mediates cell fusion³⁹. Accordingly, it is possible that the asparagine (N) of the mutation D950N (Fig. 3) of the Spike LCR-4 may enhance the stabilization of the post-fusion hairpin conformation, since the conservation of the N and Q residues of HR1 is known to play an important role in the arrangement of hydrogen-bonding zippers that force HR2 to adopt its final conformation in SARS-CoV⁴⁰. The structural relevance of this region has been demonstrated by studies with other RNA viruses, in which the use of fusion inhibitors that disrupt HR1-HR2 conformational changes, are known to limit viral entry^{41,42}.

Although there may be minuscule variations in the LCRs length and/or amino acid composition, the segments described in this work fall well into the low complexity category, suggesting that their biased composition may confer adaptive advantages to the Delta variant. For instance, the polybasic Spike LCR-3, which includes several arginines in its N-terminus, is a highly conserved sequence located precisely in the

furin cleavage site at spike S1/S2, which is essential for membrane fusion, and plays a key role in viral infection and transmission^{43,44}.

The use of the stringent cut-off value used here shows that, except for a limited number of sequences of the Spike LCR-3 and the Spike LCR-4 (39 and 82, respectively), these two LCRs are extremely prevalent (99.19% and 98.3% of all sequences) in our Delta variant sample (n = 4830). Although they display the biological traits of typical low complexity regions (Fig. 2), the multiple sequence alignments (Supplementary file 1 and 2) of the sequences that escape our cutoff values show single point mutations within these LCRs. These one-amino acid substitutions increase the complexity of the fragments, which prevents their detection by the methodology employed here. It is important to note that most of these sequences (34/39 and 82/82) lack some of the characteristic mutations (Supplementary file 1 and 2) that define the Delta variant²⁹, including the prevalent P681R. It is thus possible they may be in fact sub-lineages of this VOC.

The SARS-CoV-2 Delta variant was detected in the late 2020³⁷, and the proteomics traits described here may contribute to explain in part its rapid worldwide expansion. The role of LCRs in enhancing sequence variability in surface proteins of viral and cellular pathogens has been postulated^{8,10}. The extremely high conservation of the position and sequence of two LCRs (Spike LCR-3 and Spike LCR-4) in the Delta variant we have described is a strong indication of the importance of newly acquired raw material by slippage or recombination that may have undergone adaptative fine tuning, leading to the evolution and development of new functions or the improvement of existing ones.

LCRs can lead to variations in genome size in both viral and cellular systems⁴⁵ and may have participated in some of the processes that led to an increase in the size of primitive cellular RNA genomes^{8,46}. However, although compositionally biased sequences in SARS-CoV-2 are quite ubiquitous in most of the coronaviral proteins (Fig. 1 and S1), they do not contribute significantly to the increase of its genome size. In contrast, we hypothesize that the high conservation of the two LCRs in the Delta spike protein suggests that, together with the seven mutations present in this variant, they are part of the phenotypic traits associated with its high infectivity. Laboratory studies are required to confirm the possibility that the presence of compositionally biased segments in the Delta variant spike protein may be related to increased transmission, which is part of the defining features of VOCs and VOIs^{47–49}.

Methods

SARS-CoV-2 genome sequences

To retrieve a list of proteomes meeting the requirements to be considered as input to the pipeline (https://github.com/abelardoacm/SARS-COV2_LCRs), we downloaded metadata of all the sequences available on the China National Center for Bioinformation web portal (<https://ngdc.cncb.ac.cn/news/85>) on July 17, 2021. The entries were filtered, keeping only those that corresponded to complete proteomes (Nuc. Completeness = Complete), with high sequence quality (Sequence Quality = High) available in NCBI

GenBank (Data Source = GenBank). The sample size per variant was limited to a maximum of 4,000 sequences, a figure comparable to the numbers of the Alpha- and Delta samples analyzed here, and included multiple geographical regions (217 locations from 64 countries) that were sampled between January 20, 2020 and July 17, 2021. A subset was made for each variant classified either as a VOC (Alpha n= 3,903; Beta n= 384; Gamma n= 4,000; and Delta n= 4,830) or as a VOI (Eta n= 363; Iota n= 4,000; Kappa n= 115; and Lambda n= 259). We have also included proteomes from a random sampling using the R *sample{base}* function, of 10,377 non-VOC/VOI that met the same quality criteria and were classified as "Others SARS-CoV-2" (Others SARS-CoV-2 n= 10,377). Proteomes were downloaded using NCBI batch entrez. Accessions with empty fields in their metadata were discarded, leaving a total of 28,231 proteome files (Supplementary file 3).

Detection of low complexity regions (LCRs)

To search for the LCRs in the sample, the SEG algorithm was used with $W = 12$, $K1 = 1.9$, $K2 = 2.1$ parameters, which are slightly stricter than the default values. The pipeline "SARS-COV2_LCRs" was built to couple annotation data from genomic GenBank files with SEG output files and locate and identify LCRs within each genome.

A "genomic features" csv-file containing coordinates for both genes and proteins was prepared, which served as a template to create a proteomic fasta enriched with location information. All the PERL and R scripts we have employed are available at https://github.com/abelardoacm/SARS-COV2_LCRs.git.

Once all LCRs were identified within all proteomes in our sample, their frequency was calculated using an R script (Figure S1). From this analysis, LCRs of interest were selected based on their high prevalence in each variant proteome dataset (Table S1). Subsequently, a LCRs of interest presence matrix was calculated by an R script, and used as input to plot the total counts per variant and number of versions per low complexity region (Figure 2). The amino acid composition of the 4,830 Delta spike sequences was analyzed with a multiple sequence alignment built with MUSCLE²⁴ v3.8.1551, followed by an amino acid Logo representation (Figure S2) made with the WebLogo 3 program (<http://weblogo.threeplusone.com/create.cgi>)²⁵.

Declarations

Acknowledgments

WC-S is a PhD student from the Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship CVU-815057 from CONACyT. AA-C is a MSc student from the Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship CVU-1034340 from CONACyT. AC-G received a fellowship from CONACyT CVU-1002377. Support from DGAPA-PAPIIT (IN214421), DGAPA-PAPIME (PE204921) and SRE-AMEXCID (CH.06.UNAM) is gratefully acknowledged.

Author contributions statement

All authors contributed equally to the results and analyses presented here. All authors reviewed the manuscript.

Additional information

Competing interests

The authors declare no competing interests.

References

1. Haerty, W. & Golding, G. B. Low-complexity sequences and single amino acid repeats: not just “junk” peptide sequences. *Genome* 53, 753–762, [https://doi.org/10.1139/g10-063\(2010\)](https://doi.org/10.1139/g10-063(2010)).
2. Mier, P. *et al.* Disentangling the complexity of low complexity proteins. *Briefings in Bioinformatics* 21, 458–472, [https://doi.org/10.1093/bib/bbz007\(2020\)](https://doi.org/10.1093/bib/bbz007(2020)).
3. Ntountoumi, C. *et al.* Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. *Nucleic Acids Research* 47, 9998–10009, [https://doi.org/10.1093/nar/gkz730\(2019\)](https://doi.org/10.1093/nar/gkz730(2019)).
4. Jorda, J. & Kajava, A. V. Protein homorepeats sequences, structures, evolution, and functions. *Advances in protein chemistry and structural biology* 79, 59–88, [https://doi.org/10.1016/S1876-1623\(10\)79002-7\(2010\)](https://doi.org/10.1016/S1876-1623(10)79002-7(2010)).
5. Haerty, W. & Golding, G. B. Genome-wide evidence for selection acting on single amino acid repeats. *Genome Research* 20, 755–760, [https://doi.org/10.1101/gr.101246.109\(2010\)](https://doi.org/10.1101/gr.101246.109(2010)).
6. Fankhauser, N., Nguyen-Ha, T. M., Adler, J. & Mäser, P. Surface antigens and potential virulence factors from parasites detected by comparative genomics of perfect amino acid repeats. *Proteome Science* 5, 20, [https://doi.org/10.1186/1477-5956-5-20\(2007\)](https://doi.org/10.1186/1477-5956-5-20(2007)).
7. Mendes, T. A. O. *et al.* Repeat-Enriched Proteins Are Related to Host Cell Invasion and Immune Evasion in Parasitic Protozoa. *Molecular Biology and Evolution* 30, 951–963, [https://doi.org/10.1093/molbev/mst001\(2013\)](https://doi.org/10.1093/molbev/mst001(2013)).
8. Velasco, M. A. *et al.* Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 gp120 protein. *Journal of Theoretical Biology* 338, 80–86, [https://doi.org/10.1016/j.jtbi.2013.08.039\(2013\)](https://doi.org/10.1016/j.jtbi.2013.08.039(2013)).
9. Mier, P. & Andrade-Navarro, M. A. The Conservation of Low Complexity Regions in Bacterial Proteins Depends on the Pathogenicity of the Strain and Subcellular Location of the Protein. *Genes*, 12(3), 451, [https://doi.org/10.3390/genes12030451\(2021\)](https://doi.org/10.3390/genes12030451(2021)).

10. Gruca, A. *et al.* Common low complexity regions for SARS-CoV-2 and human proteomes as potential multidirectional risk factor in vaccine development. *BMC Bioinformatics* 22,182, <https://doi.org/10.1186/s12859-021-04017-7>(2021).
11. Li, Y. C., Korol, A. B. & Fahima, T. & E., Nevo. Microsatellites Within Genes: Structure, Function, and Evolution. *Molecular Biology and Evolution* 21,991–1007, <https://doi.org/10.1093/molbev/msh073>(2004).
12. Lin, W. H. & Kussell, E. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic acids research*, 40(6), 2399–2413, <https://doi.org/10.1093/nar/gkr1078>(2012).
13. Alam, C. M., Iqbal, A., Sharma, A., Schulman, A. H. & Ali, S. Microsatellite diversity, complexity, and host range of mycobacteriophage genomes of the siphoviridae family. *Frontiers in genetics*, **10**, 207 (2019). <https://doi.org/10.3389/fgene.2019.00207>
14. Laskar, R., Jilani, M. G. & Ali, S. Implications of genome simple sequence repeats signature in 98 Polyomaviridae species. *Biotech* 11,35, <https://doi.org/10.1007/s13205-020-02583-w>(2021).
15. Zhao, X. *et al.* Coevolution between simple sequence repeats (SSRs) and virus genome size. *BMC Genomics* 13,435, <https://doi.org/10.1186/1471-2164-13-435>(2012).
16. Zhao, X. *et al.* Microsatellites in different Potyvirus genomes: Survey and analysis. *Gene* 488,52–56, <https://doi.org/10.1016/j.gene.2011.08.016>(2011).
17. Alam, C. M., Singh, A. K., Sharfuddin, C. & Ali, S. In-silico analysis of simple and imperfect microsatellites in diverse tobamovirus genomes. *Gene* 530,193–200, <https://doi.org/10.1016/j.gene.2013.08.046>(2013).
18. George, B., Mashhood Alam, C., Jain, S. K., Sharfuddin, C. & Chakraborty, S. Differential distribution and occurrence of simple sequence repeats in diverse geminivirus genomes. *Virus Genes* 45,556–566, <https://doi.org/10.1007/s11262-012-0802-1>(2012).
19. Wu, X., Zhou, L., Zhao, X. & Tan, Z. The analysis of microsatellites and compound microsatellites in 56 complete genomes of Herpesvirales. *Gene* 551,103–109, <https://doi.org/10.1016/j.gene.2014.08.054>(2014).
20. Chen, M. *et al.* Similar distribution of simple sequence repeats in diverse completed Human Immunodeficiency Virus Type 1 genomes. *FEBS Letters* 583,2959–2963, <https://doi.org/10.1016/j.febslet.2009.08.004>(2009).
21. Alam, C., Sharfuddin, C. & Ali, S. Analysis of Simple and Imperfect Microsatellites in Ebolavirus Species and Other Genomes of Filoviridae Family. *Gene Cell Tissue*, **2** (2), e26204 (2015). <https://doi.org/10.17795/gct-26204>
22. Satyam, R. *et al.* Deciphering the SSR incidences across viral members of Coronaviridae family. *Chemico-Biological Interactions* 331,109226, <https://doi.org/10.1016/j.cbi.2020.109226>(2020).
23. Siddiqe, R. & Ghosh, A. Genome-wide in silico identification and characterization of Simple Sequence Repeats in diverse completed SARS-CoV-2 genomes. *Gene Reports* 23,101020, <https://doi.org/10.1016/j.genrep.2021.101020>(2021).

24. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797 (2004). <https://doi.org/10.1093/nar/gkh340>
25. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research*, **14** (6), 1188–1190 (2004). <https://doi.org/10.1101/gr.849004>
26. Benton, D. J. *et al.* The effect of the D614G substitution on the structure of the spike glycoprotein of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **118**, e2022586118, 10.1073/pnas.2022586118 (2021).
27. Hancock, J. The contribution of slippage-like processes to genome evolution. *Journal of Molecular Evolution* **41**, <https://doi.org/10.1007/BF00173185> (1995).
28. DePristo, M. A., Zilversmit, M. M. & Hartl, D. L. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* **378**, 19–30, <https://doi.org/10.1016/j.gene.2006.03.023> (2006).
29. Public Health England. SARS-CoV-2 variants of concern and variants under investigation in England. Sage (2021).
30. Roy, B. & Roy, H. The Delta Plus variant of COVID-19: Will it be the worst nightmare in the SARS-CoV-2 pandemic? *Journal of Biomedical Sciences* **8**, 1–2 <https://doi.org/10.3126/jbs.v8i1.38449> (2021).
31. Alexandar, S., Ravinsakar, M., Senthil Kumar, R. & Jakkan, K. A Comprehensive Review on Covid-19 Delta variant. *International Journal of Pharmacology and Clinical Research (IJPCR)* **5**, 83–85 (2021).
32. Xia, X. Domains and Functions of Spike Protein in SARS-Cov-2 in the Context of Vaccine Design. *Viruses* **13**, 109 <https://doi.org/10.3390/v13010109> (2021).
33. Duan, L. *et al.* The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens. *Frontiers in Immunology* **11**, <https://doi.org/10.3389/fimmu.2020.576622> (2020).
34. Peacock, T. P. *et al.* The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nat Microbiol* **6**, 899–909. <https://doi.org/10.1038/s41564-021-00908-w> (2021).
35. Frazier, L. E. *et al.* Spike protein cleavage-activation mediated by the SARS-CoV-2 P681R mutation: a case study from its first appearance in a variant of interest (VOI) A.23.1 identified in Uganda. *bioRxiv* 2021.06.30.450632 <https://doi.org/10.1101/2021.06.30.450632> (2021).
36. Saito, A. *et al.* SARS-CoV-2 spike P681R mutation, a hallmark of the Delta variant, enhances viral fusogenicity and pathogenicity. *bioRxiv* 2021.06.17.448820 <https://doi.org/10.1101/2021.06.17.448820> (2021).
37. GISAID. Tracking of Variants. Retrieved on July 14, from <https://www.gisaid.org/hcov19-variants/> (2021).
38. Centers for Disease Control and Prevention. COVID Data Tracker. Retrieved on July 14, from <https://covid.cdc.gov/covid-data-tracker/#variant-proportions> (2021).
39. Xia, S. *et al.* Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. *Cell Mol Immunol* **17**, 765–767 <https://doi.org/10.1038/s41423-020-0374-2> (2020).

40. Duquerroy, S., Vigouroux, A., Rottier, P. J. M. & Rey, F. A. & Jan Bosch, B. Central ions and lateral asparagine/glutamine zippers stabilize the post-fusion hairpin conformation of the SARS coronavirus spike glycoprotein. *Virology* 335, 276–285, [10.1016/j.virol.2005.02.022](https://doi.org/10.1016/j.virol.2005.02.022) (2005).
41. Feng, M., Bell, D. R., Kang, H., Shao, Q. & Zhou, R. Exploration of HIV-1 fusion peptide–antibody VRC34.01 binding reveals fundamental neutralization sites. *Physical Chemistry Chemical Physics* 21, 18569–18576 <https://doi.org/10.1039/C9CP02909E> (2019).
42. Ispas, G. *et al.* Antiviral Activity of TMC353121, a Respiratory Syncytial Virus (RSV) Fusion Inhibitor, in a Non-Human Primate Model. *PLOS ONE* 10, e0126959 <https://doi.org/10.1371/journal.pone.0126959> (2015).
43. Wu, Y. & Zhao, S. Furin cleavage sites naturally occur in coronaviruses. *Stem Cell Research*, **50**, 102115 <https://doi.org/10.1016/j.scr.2020.102115> (2021).
44. Scudellari, M. How the coronavirus infects cells—and why Delta is so dangerous. *Nature* **595**, 640–644 (2021). <https://doi.org/10.1038/d41586-021-02039-y> (2021).
45. Hancock, J. M. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica* 115, 93–103, <https://doi.org/10.1023/A:1016028332006> (2002).
46. Cisneros-Martínez, A. M., Becerra, A. & Lazcano, A. Ancient gene duplications in RNA viruses revealed by protein tertiary structure comparisons. *Virus evolution*, 7(1), veab019, <https://doi.org/10.1093/ve/veab019> (2021).
47. Weisblum, Y. *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* 9, <https://doi.org/10.7554/eLife.61312> (2020).
48. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 372, eabg3055, <https://doi.org/10.1126/science.abg3055> (2021).
49. Zhou, W. & Wang, W. Fast-spreading SARS-CoV-2 variants: challenges to and new design strategies of COVID-19 vaccines. *Signal Transduction and Targeted Therapy* 6, 226, <https://doi.org/10.1038/s41392-021-00644-x> (2021).

Figures

Figure 1

LCRs in VOCs (Alpha, Beta, Gamma and Delta variants), VOIs (Epsilon, Eta, Iota, Kappa and Lambda variants) and Other SARS-CoV-2 proteomes (Others). a. LCRs present in ORF1ab, which includes nsp1-4, 3CL protease (3CL), nsp6-nsp11, RdRp polymerase (RdRp), Helicase (Hel) and nsp 14-16. b. LCRs along spike, ORF3a (3a), envelope (E), membrane (M), ORFs 6, 7a, 7b, 8, nucleocapsid (N) and ORF10. The LCRs

spike sequences reported here in the Delta, Iota and Kappa variants are represented with red lines. The width of the lines is not proportional to the number of sequences in each variant.

Spike protein complexity

Figure 2

Low complexity regions (LCRs) in the spike proteins of VOIs, VOCs and Other SARS-CoV-2, Delta, Iota and Kappa variants. The LCRs prone regions are shown in colors. In the x axis number of amino acids residues and in the y axis the complexity levels.

Figure 3

a. The SARS-CoV-2 spike protein three-dimensional structure (by cryo-electron microscopy²⁶, PDB code: 7BNM). The structure corresponds to a trimer, where each monomer is represented with a different color. The two subunits that make up each monomer (Subunit 1, also known as Head region, and Subunit 2, or the Stalk region) are indicated. b. Domain organization of the spike protein. The position of each of the LCRs found in this work, together with the mutations present in each variant spike protein are shown. The position of the signal peptide (SP) and Spike LCR-1 are indicated. The green arrow in the Spike LCR-3 box indicates the furin cleavage site. c. Monomer of the spike protein. Close ups of each of the structural regions corresponding to the different LCRs are shown in colored boxes. The sequences of each LCR are represented, with the mutations indicated with a red letter. Abbreviations: SP, signal peptide; NTD, N-terminal domain; RBD, receptor-binding domain; RBM, receptor-binding motif; SD1, subdomain 1; SD2, subdomain 2; FP, fusion peptide; HR1, heptad repeat 1; CH, central helix; CD, connector domain; HR2, heptad repeat 2; TM, transmembrane domain.

Supplementary Files

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

- [FigureS1.png](#)
- [FigureS2.png](#)
- [supplementaryfile1.fasta](#)
- [supplementaryfile2.fasta](#)
- [supplementaryfile3.txt](#)