# Pre-pandemic artificial MERS analog of pathogenic polyfunctional SARS-CoV-2 S1/S2 furin cleavage site domain is unique among betacoronaviruses

*Andreas Martin Lisewski*[1]
*Department of Life Sciences and Chemistry*
*School of Science*
*Constructor University*
*28759 Bremen, Germany*

[1] *Email corresponding author: alisewski@constructor.university*

## Abstract

SARS-CoV-2 spike (S) glycoprotein furin cleavage site is a key determinant of SARS-CoV-2 virulence and COVID-19 pathogencity. Located at the S1/S2 junction, it is unique among sarbecoviruses but frequently found among betacoronaviruses. Recent evidence suggests that this site includes two additional functional motifs: a pat7 nuclear localization signal and two flanking O-glycosites. However, a systematic overview of spikes bearing this polyfunctional sequence domain has been missing. Here we report that among spike sequences of genus Betacoronavirus and outside of the SARS-CoV-2 clade an analogous domain was found in only one other virus: an artificial MERS infectious clone, described already in 2017, which was rationally selected from serial passage in genetically humanized mice. As the evolutionarily closest betacoronaviruses outside of the SARS-CoV-2 clade lack all its three functional motifs, this critical domain becomes an unlikely product of natural evolution alone, which extends the current view on SARS-CoV-2 origins, evolution and pathogenesis.

**Keywords**: *COVID-19, SARS-CoV-2 evolution, COVID-19 pathogenesis, spike glycoprotein, furin cleavage site, nuclear localization*

## Introduction

The furin cleavage site (FCS) at the S1/S2 domain junction of the SARS-CoV-2 spike (S) glycoprotein has been recurrently discussed in the context of SARS-CoV-2 origins, SARS-CoV-2 virulence, and COVID-19 pathogenicity (Holmes et al., 2021)(Hasan et al., 2021). In comparison to bat coronavirus (BatCoV) RaTG13 (GenBank accession MN996532) and BANAL-20-52 (GenBank MZ937000), the closest genomic relatives to SARS-CoV-2, the reference sequence (Wuhan Hu-1 isolate, GenBank NC_045512) features a four amino acid $^{681}$PRRA$^{684}$ insert between two adjacent Ser and Arg residues, resulting in a RXXR minimal FCS. This FCS, which does not fully match the canonical FCS motif RX(K/R)R (Holmes et al., 2021), has not been seen in other sarbecoviruses (Coutard et al., 2020). On the other hand, simple furin and furin-like cleavage sites at S1/S2 domains in other betacoronavirus spike glycoproteins have been known (Wu and Zhao, 2021) and used as evolutionary evidence against arguments that the SARS-CoV-2 FCS might not be of natural origin (Holmes et al., 2021)(Garry, 2022).

In comparison to other local sequence features, the novel FCS has been a main research focus even though it was already predicted in early 2020 (Andersen et al., 2020) and then experimentally confirmed that this FCS is also flanked by two proximal *O*-linked glycosylation sites, Thr678 and Ser686 (Shajahan et al., 2020) (Sanda et al., 2021) (Gao et al., 2020). *O*-linked glycosylation of these two residues demonstrated their functional role as modulators of FCS, membrane fusion, and virus penetration activity (Zhang et al., 2021) (Shajahan et al., 2021) (Gong et al., 2021). In addition to these functions, Hatmal and colleagues (Hatmal et al., 2020) predicted in 2020 that this FCS itself is part of a pat7 nuclear localization signal (NLS) at the S1/S2 domain junction of SARS-CoV-2 spike, $^{681}$PRRARSV$^{687}$. This prediction was consistent with later observations of the spike glycoprotein localizing inside the nucleus during SARS-CoV-2 infection and COVID-19 progression (Eymieux et al, 2021)(Kim et al, 2022)(Chen et al, 2022). More specifically, Sattar *et al*. recently confirmed the $^{681}$PRRARSV$^{687}$ S1/S2 NLS (Sattar et al., 2023) and showed that SARS-CoV-2 spike translocated into the nucleus whereas a pat7 deficient SARS-CoV spike did not. The structural location of the NLS within the S ectodomain might therefore suggest that spike protein export bifurcates at the Golgi apparatus with an additional nuclear pathway that mimics or hijacks cellular fibroblast growth factor receptor 1 (FGFR1) signaling (Stachowiak, 2016), where extraction from membranes is coupled with solubilisation in the cytosol, followed by transport of the solubilised receptor through the nuclear pore. While the functional mechanisms and pathogenic implications of this NLS remain to be further elucidated, the available data already present the possibility that SARS-CoV-2 spike nuclear translocation is linked to pathogenesis in a similar way as the nucleocapsid protein's is linked to it (Gao et al, 2021) (Timani et al., 2005).  As one of the classical NLS (Hicks and Raikhel, 1995), pat7 is defined by seven consecutive residues starting with a Pro, followed by a stretch of four residues that include three basic amino acids (Nakai and Horton, 1999). Together with this new putative function of pat7 NLS mediated S translocation, we hypothesize that the spike SARS-CoV-2 S1/S2 junction domain between residues Thr678 and Val687 is polyfunctional bearing at least the above three functional sequence motifs.

## Methods

The representative set of 2,465 betacoronavirus S protein overlapping homologous superfamily sequences was retreived in fasta format on 4 December 2022 from the InterPro repository at https://www.ebi.ac.uk/interpro/entry/InterPro/IPR042578/. From these sequences were extracted 98,122 furin cleavage site (FCS) motifs using the FindFur algorithm as described by (Gu,

2020) and deposited on 15 December 2020 at the GitHub software repository at https://github.com/chwisteeng/FindFur. These sequences were individually checked for The/Ser *O*-glycosite residue pairs with the standard prediction software NetOGlyc4.0 (Steentoft et al., 2013) as available at https://services.healthtech.dtu.dk/services/NetOGlyc-4.0/. Comprehensive sequence database searches using were performed using the NCBI protein BLAST (BLASTP) algorithm with webservice available at  https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins; the following BLASTP search parameters and settings were used: Word size=2; Expect value=200000; Hitlist size=500; Gapcosts=9,1; Matrix=PAM30; Filter string= F; Genetic Code=1;Window Size=40; Threshold=11; Composition-based stats=0; Database Posted date=Jan 19, 2023 2:59 AM; Number of letters=17,117,563; Number of sequences=10,766; Entrez query: Includes: Betacoronavirus (taxid:694002) Excludes: SARS-CoV-2 (taxid:2697049). The pat7 input query consensus motif sequences were TXXPR(K/H/R)XRSX and TXXPRX(K/H/R)RSX. The resulting text output of this sequence search was compiled and deposited in Data File 1 (Supplementary Material).
Spike protein amino acid substitution tables for pandemic variants of concern Alpha, Beta, Gamma, Delta and Omicron were compiled from from the Expasy:Viral Zone online repository at https://viralzone.expasy.org/9556 (version 4 September 2023 of *SARS-CoV-2 Circulating Variants*). Global prevalence data of SARS-CoV-2 spike substitutions P681H, P681R, A683V, V687I, V687L were retrieved from the SARS-CoV-2 lineage mutation tracker reports through online calls https://outbreak.info/situation-reports?muts=S:P681H (and with corresponding calls for the other substitutions; version 14 November 2023) which were based on extensive sequencing data from the global GISAID Initiative (https://gisaid.org/). Global prevalence

data for variants of concern were retrieved from the online CoVariants repository at https://covariants.org/ (version 14 November 2023 with United States data as reference) which also represented the same GISAID sequence data source.

## Results

To systematically analyze the occurrence of analogous polyfunctional sequence domains across a comprehensive set of relevant virus species, we turned to the curated '*betacoronavirus spike glycoprotein*' collection of overlapping homologous superfamilies (InterPro entry IPR042578). This collection represented the entire genus *Betacoronavirus* through 2,465 evolutionarily diverse spike protein sequences across 34 virus species. We then extracted all 98,122 predicted FCS motifs (see, Methods) within a constant frame of twenty amino acid residues (Gu, 2020) which were subsequently automatically filtered for pat7 NLS motifs. After removing sequence fragments and duplicates, this procedure resulted in a set of twenty representative sequences (Table 1; number 1-19 and 21) to which two related betacoronavirus sequences (Table 1; 20, 22) were manually added as pat7 negatives.  This analysis thus produced only two betacoronavirus positive hits outside of the SARS-CoV-2 clade: one human merbecovirus and one human embecovirus (Table 1; 19 and 21).

To test the sensitivity of this outcome on the size of the sequence search space, the output was also independently verified through NCBI BLASTP searches, across all 10,766 betacoronavirus protein sequences outside of the SARS-CoV-2 clade in that database. This number was an order of magnitude larger than the 1,179 betacoronavirus InterPro/UniProt sequences of that kind that were used above. This test confirmed the non-random and spike S1/S2 specific pat7/FCS motif design (see, Methods and Supplemental Material, Data File 1) as no other spike pat7/FCS sequence motif representatives were found than those already given in Table 1, numbers 19 and 21. In addition, the presence of the adjacent The/Ser *O*-glycosite residue pair was homology inferred within the SARS-CoV-2 clade, and for other sequences individually checked with the standard prediction software NetOGlyc4.0 (Steentoft et al., 2013), where positive hits were recorded (Table 1) only for MERS betacoronavirus first isolate EMC/2012 and for MERS betacoronavirus isolate MA30 (denoted as $\text{MERS}_{\text{MA30}}$).

While most resulting virus consensus sequences (Table 1; 1-18) corresponded to within-clade SARS-CoV-2 variants that tightly preserved the entire polyfunctional TXXPRRXRSX consensus motif, the two additional betacoronavirus sequences detected (Table 1; 19, 21) were from the spike of human embecovirus HKU1 (HCoV-HKU1, with genotype B, GenBank ABD75545); and from a murine adapted MERS merbecovirus isolate MA30 (GenBank MT576585). The HCoV-HKU1 sequence presented the canonical FCS motif RRKR embedded into a complete pat7 motif PSSRRKR; however, there was no Thr/Ser *O*-glycosite pair at the expected flanking positions, and therefore the sequence represented not a full functional analog of the corresponding SARS-CoV-2 domain. Also, unlike SARS-CoV-2 and MERS $\text{CoV}_{\text{MA30}}$, the HKU1 furin cleavage site dependent proteolytic cut was outside of the pat7 NLS due to a double amino acid shift of the FCS sequence location and so the two motifs did not functionally interfere in HKU1 CoV. By contrast, the $\text{MERS}_{\text{MA30}}$ sequence comprised the entire polyfunctional domain, $^{744}$TLTPRRVRSV$^{753}$ , with robust *O*-glycosite predictions for Thr744 and for Ser752 and with the FCS located within the pat7 NLS. These data indicated that within a broad and representative set of betcoronaviruses S sequences MERS$_{MA30}$ spike was the only instance of a complete pat7/FCS/*O*-glycosite pair motif analogous to the S1/S2 polyfunctional sequence domain of SARS-CoV-2 spike (Wuhan Hu-1 genotype).

The corresponding TXXPRRXRSX consensus motif between the spikes of SARS-CoV-2 and MERS$_{MA30}$, which overall share a level of 32.7% in protein sequence identity, was not a product of convergent evolution. This was evident from two distinct observations:

First, the closest parental strain to MERS$_{MA30}$ was the MERS first isolate EMC/2012 (Genbank NC_019843), which lacks the pat7 motif ($^{744}$TLT-PRSVRSV$^{753}$; see Table 1). Only during adaptation and rational selection after serial passage in artificially humanized mice the pat7 precursor sequence PRSVRSV changed into a full pat7 NLS through a non-synonymous mutation (Ser749Arg) in MERS S (Li et al., 2017). MERS$_{MA30}$ S pat7/FCS is therefore a product of directed evolution/adaptation and rational selection in bioengineered host cells (Figure 1A).

Second, the closest genomic relative to SARS-CoV-2, the bat coronavirus RaTG13, is devoid of the polyfunctional domain ($^{678}$TNS----RSV$^{683}$; see Table 1) but already includes the fixed and highly conserved residue positions 678/682 and 681, which are functionally inactive *O*-glycosite and FCS precursors, respectively. If, on this conditional (Thr678/Ser682 and Arg681) background, natural selection first proceeds to a complete FCS and then progresses into a full pat7, then this FCS would produce a selective disadvantage for pat7 because FCS proteolytic cleavage would abrogate the pat7 phenotype. Conversely, a full pat7 motif without FCS would produce a selective disadvantage for FCS, because S protein cellular localization shifts to the nucleus, *i.e.* away from the normal sites of furin expression and activity (see, UniProt entry P09958).

This evolutionary incompatibility argument is supported by three independent lines of evidence: (*a*) in MERS CoV, continued serial passage of $\mathrm{MERS}_{\mathrm{MA30}}$ in the humanized (*hDPP4* knock-in) murine model lead to an abrogation of pat7 NLS through a P747H mutation (Li et al., 2017), which demonstrated that the pat7/FCS MA30 genotype was not fixed during adaptation and ultimately was selected against and, more broadly, through the absolute lack of evidence of any naturally occurring MERS CoV with a pat7/FCS spike (Table 1); (*b*) in the betacoronavirus HKU1 spike sequence, which does not present the corresponding glycosites (Table 1), any cleavage induced by the canonical RRKR↓ motif does not incompatible with pat7 as such proteolytic cut (↓) would occur, in contrast to $\mathrm{MERS}_{\mathrm{MA30}}$ CoV and SARS-CoV-2, after the last residue in its pat7 motif; (*c*) for SARS-CoV-2 and in further analogy with $\mathrm{MERS}_{\mathrm{MA30}}$ CoV,  all globally dominant SARS-CoV-2 variants (in chronological order: Alpha, Delta, Omicron, and their later recombinants) that had emerged after the original 2019 Wuhan Hu-1 isolate included spike P681(H/R) mutations which ablate pat7 NLS (Figure 1B). Specifically, the presence and absence of the P681(H/R) substitution was, among all characteristic spike mutations within the pandemic variants of concern Alpha, Beta, Gamma, Delta, Omicron, the only one which was strictly correlated with global dominance of the Alpha, Delta and Omicron, and with the lack thereof for the marginal Beta and Gamma variants (Figure 1B), respectively. Moreover, based on an extensive analysis of 9,310,717 SARS-CoV-2 genomic sequences collected worldwide between January 2020 and November 2023 (Figure 1C), the earliest time point of dominance of a P681(H/R) spike variant coincided with the earliest time point of emergence of the first globally dominant (with prevalence >50%) variant of concern Alpha, followed by Delta, and Omicron. In a negative control, all other neighbouring mutations inside the pat7/FCS motif remained marginal with prevalence below 0.5% (Figure 1C). Thus the P681(H/R) genotype was an accurate indicator of both pat7 loss and of a selective sweep from globally emerging SARS-CoV-2 variants. Of note, such specific substitution of the characteristic pat7 Pro residue is known to be associated with a functional loss of nuclear translocation in analogous pat7 NLS (Korb et al., 2013). This overall analysis suggests that a combined pat7/FCS spike S1/S2 motif is directly selected against and that its recent emergence as a SARS-CoV-2 Wuhan Hu-1 genotype during the course of natural betacoronavirus evolution alone was unlikely.

## Discussion

The collected evidence presented here suggests that, when the pat7 nuclear localization signal (NLS) motif is additionally taken into consideration, the resulting S1/S2 pat7/FCS polyfunctional domain may support—in analogy with the non-natural virus isolate $\text{MERS}_{\text{MA30}}$ and in addition to its natural evolution—an artificial adaptation and selection process in the original lineage that lead to SARS-CoV-2. As a cautionary remark, it should be noted that a more comprehensive sequence sampling of betacoronaviruses in the future might still reject this claim by identifying other natural betacoronaviruses spikes with a combined pat7/FCS S1/S2 motif. Nevertheless, the broader perspective around pat7 NLS contrasts earlier viewpoints (Wu and Zhao, 2021) (Holmes et al., 2021) (Garry, 2022), which claimed that simple SARS-CoV-2 spike FCS, along with its observed co-occurrence in other betacoronaviruses spikes, is an indicator of a natural evolutionary origin of this key pathogenic feature of SARS-CoV-2.

More specifically, our results and observations suggest that among betacoronavirus spike glycoprotein S1/S2 domains there is a natural selection against a simultaneous emergence of combined pat7/FCS motifs. Therefore it can be argued that a spike SARS-CoV-2 FCS/pat7 was acquired not through convergent evolution but more likely through an *all-in-one* step, for example through natural recombination, or through genetic engineering and/or passage in an artificial host (Figure 1B). However, potential recombination with other positive-strand RNA viruses remains elusive (Wang et al., 2022), as neither a specific recombination event following host co-infection or a natural pat7/FCS nucleotide matching (betacorona-)virus RNA sequence template outside of the SARS-CoV-2 clade have been identified. Also, a SARS-CoV-2 progenitor's host, which could have served as a recombination donor for this critical domain, has not been identified but a genetically engineered human gene (Ambati et al, 2022) and, more evidently, artificial yeast cells (Lisewski, 2022) have been discussed.

The humanized murine animal model adapted $\text{MERS}_{\text{MA30}}$ CoV may be characterized as an artificial virus because its primary host is a bioengineered, transgenic organism that does not naturally exist. Specifically, MERS$_{\text{MA30}}$ originated from thirty passages in humanized (transgenic) mice whose MERS infection insensitive endogenous dipeptidyl peptidase 4 (*DPP4*) gene was—at key exons—replaced with the corresponding human *hDPP4* entry receptor gene regions (Li et al., 2017). After thirty passages, a genomically stable and highly pathogenic genotype was rationally selected to represent the coronavirus quasispecies, by minimizing the number of genomic deletions, and then bioactively preserved for future applications (Gutierrez-Alvarez et al., 2021) (Heise et al., 2021) through reverse genetics assembly into an infectious cDNA clone (realized as a bacterial artificial chromosome). The resulting infectious clone of MERS$_{\text{MA30}}$, which differentially carried the pat7 NLS conferring S749R mutation within the S gene, presented both increased infectivity and higher pathogenicity than both MERS S pat7-negative (P747H and R749S) variants. In this animal model the combined sequence motif S1/S2 pat7 NLS/FCS, but not S1/S2 FCS alone, was differentially and significantly associated with increased infectivity and with stronger disease phenotype (Li et al., 2017). Due to the exact motif analogy between MERS$_{\text{MA30}}$ and SARS-CoV-2 at the critical S1/S2 junction domain, a similar pathogenic role of SARS-CoV-2 spike pat7 NLS might also exist during COVID-19 progression. This hypothesis warrants further experimental verification.

In summary, a systematic analysis shows that, within genus *Betacoronavirus,* the S1/S2 domain of MERS$_{\text{MA30}}$ spike constitutes the only SARS-CoV-2 spike functional analog of the polyfunctional pat7/FCS S1/S2 domain. This unique analogy with artificial MERS$_{\text{MA30}}$ betacoronavirus suggests that the original SARS-CoV-2 lineage was not entirely natural.

## References

Ambati B.K., Varshney A., Lundstrom K., Palú G., Uhal B.D., Uversky V.N., Brufsky A.M., 2022. MSH3 Homology and Potential Recombination Link to SARS-CoV-2 Furin Cleavage Site. Front. Virol. 2:834808.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat Med 26, 450–452.

Chen M., Ma Y., Chang W., 2022. SARS-CoV-2 and the Nucleus. Int J Biol Sci. 18(12):4731-4743.

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N.G., Decroly, E., 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 176, 104742.

Cubuk, J., Alston, J.J., Incicco, J.J., Singh, S., Stuchell-Brereton, M.D., Ward, M.D., Zimmerman, M.I., Vithani, N., Griffith, D., Wagoner, J.A., Bowman, G.R., Hall, K.B., Soranno, A., Holehouse, A.S., 2021. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. Nat Commun 12, 1936.

Eymieux S., Rouillé Y., Terrier O., Seron K., Blanchard E., Rosa-Calatrava M., Dubuisson J., Belouzard S., Roingeard P. Ultrastructural modifications induced by SARS-CoV-2 in Vero cells: a kinetic analysis of viral factory formation, viral particle morphogenesis and virion release. Cell Mol Life Sci. 78(7):3565-3576.

Gao T., Gao Y., Liu X., Nie Z., Sun H., Lin K., Peng H., Wang S. Identification and functional analysis of the SARS-COV-2 nucleocapsid protein. BMC Microbiol. 2021 Feb 22;21(1):58.

Gao, C., Zeng, J., Jia, N., Stavenhagen, K., Matsumoto, Y., Zhang, H., Li, J., Hume, A.J., Mühlberger, E., van Die, I., Kwan, J., Tantisira, K., Emili, A., Cummings, R.D., 2020. SARS-CoV-2 Spike Protein Interacts with Multiple Innate Immune Receptors. bioRxiv 2020.07.29.227462

Garry, R.F., 2022. SARS-CoV-2 furin cleavage site was not engineered. Proceedings of the National Academy of Sciences 119, e2211107119.

Gong, Y., Qin, S., Dai, L., Tian, Z., 2021. The glycosylation in SARS-CoV-2 and its receptor ACE2. Sig Transduct Target Ther 6, 1–24.

Gu, C., 2020. FindFur: A Tool for Predicting Furin Cleavage Sites of Viral Envelope Substrates. Master's Thesis, San Jose State University, CA, USA.

Gutierrez-Alvarez, J., Wang, L., Fernandez-Delgado, R., Li, K., McCray, P.B., Perlman, S., Sola, I., Zuñiga, S., Enjuanes, L., 2021. Middle East Respiratory Syndrome Coronavirus Gene 5 Modulates Pathogenesis in Mice. J Virol 95, e01172-20.

Hasan, A., Paray, B.A., Hussain, A., Qadir, F.A., Attar, F., Aziz, F.M., Sharifi, M., Derakhshankhah, H., Rasti, B., Mehrabi, M., Shahpasand, K., Saboury, A.A., Falahati, M., 2021. A review on the cleavage priming of the spike protein on coronavirus by angiotensin-converting enzyme-2 and furin. J Biomol Struct Dyn 39, 3025–3033.

Hatmal M.M., Alshaer W, Al-Hatamleh M.A.I., Hatmal M., Smadi O., Taha M.O., Oweida A.J., Boer J.C., Mohamud R, Plebanski M. Comprehensive Structural and Molecular Comparison of Spike Proteins of

SARS-CoV-2, SARS-CoV and MERS-CoV, and Their Interactions with ACE2. Cells. 9(12):2638.

Heise, M., Dermody, T.S., Casadevall, A., Sandri-Goldin, RM, Schloss, P.D., 2021. The Decision To Publish Gutierrez-Alvarez et al., "Middle East Respiratory Syndrome Coronavirus Gene 5 Modulates Pathogenesis in Mice." Journal of Virology 95, e02118-20

Hicks, G.R., Raikhel, N.V., 1995. Protein import into the nucleus: an integrated view. Annu Rev Cell Dev Biol 11, 155–188.

Holmes, E.C., Goldstein, S.A., Rasmussen, A.L., Robertson, D.L., Crits-Christoph, A., Wertheim, J.O., Anthony, S.J., Barclay, W.S., Boni, M.F., Doherty, P.C., Farrar, J., Geoghegan, J.L., Jiang, X., Leibowitz, J.L., Neil, S.J.D., Skern, T., Weiss, S.R., Worobey, M., Andersen, K.G., Garry, R.F., Rambaut, A., 2021. The origins of SARS-CoV-2: A critical review. Cell 184, 4848–4856.

Kim ES, Jeon MT, Kim KS, Lee S, Kim S, Kim DG. Spike Proteins of SARS-CoV-2 Induce Pathological Changes in Molecular Delivery and Metabolic Function in the Brain Endothelial Cells. Viruses. 2021 Oct 8;13(10):2021.

Korb E, Wilkinson CL, Delgado RN, Lovero KL, Finkbeiner S. Arc in the nucleus regulates PML-dependent GluA1 transcription and homeostatic plasticity. Nat Neurosci. 2013 Jul;16(7):874-83.

Li, K., Wohlford-Lenane, C.L., Channappanavar, R., Park, J.-E., Earnest, J.T., Bair, T.B., Bates, A.M., Brogden, K.A., Flaherty, H.A., Gallagher, T., Meyerholz, D.K., Perlman, S., McCray, P.B., 2017. Mouse-adapted MERS coronavirus causes lethal lung disease in human DPP4 knockin mice. Proc Natl Acad Sci U S A 114, E3119–E3128.

Lisewski A.M., 2022. Evidence for yeast artificial synthesis in SARS-CoV-2 and SARS-CoV-1 genomic sequences. F1000Research 10:912.

Nakai, K., Horton, P., 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24, 34–36.

Pavan, M., Bassani, D., Sturlese, M., Moro, S., 2022. From the Wuhan-Hu-1 strain to the XD and XE variants: is targeting the SARS-CoV-2 spike protein still a pharmaceutically relevant option against COVID-19? Journal of Enzyme Inhibition and Medicinal Chemistry 37, 1704–1714.

Sanda, M., Morrison, L., Goldman, R., 2021. N- and O-Glycosylation of the SARS-CoV-2 Spike Protein. Anal Chem 93, 2003–2009.

Sattar S., Kabat J., Jerome K., Feldmann F., Bailey K., Mehedi M. Nuclear translocation of spike mRNA and protein is a novel feature of SARS-CoV-2. Front Microbiol. 2023 Jan 26;14:1073789.

Shajahan, A., Pepi, L.E., Rouhani, D.S., Heiss, C., Azadi, P., 2021. Glycosylation of SARS-CoV-2: structural and functional insights. Anal Bioanal Chem 413, 7179–7193.

Shajahan, A., Supekar, N.T., Gleinich, A.S., Azadi, P., 2020. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology 30, 981–988.

Stachowiak M.K., Stachowiak E.K.. Evidence-Based Theory for Integrated Genome Regulation of Ontogeny--An Unprecedented Role of Nuclear FGFR1 Signaling. J Cell Physiol. 231(6):1199-218.

Steentoft, C., Vakhrushev, S.Y., Joshi, H.J., Kong, Y., Vester-Christensen, M.B., Schjoldager, K.T.-B.G., Lavrsen, K., Dabelsteen, S., Pedersen, N.B., Marcos-Silva, L., Gupta, R., Bennett, E.P., Mandel, U., Brunak, S., Wandall, H.H., Levery, S.B., Clausen, H., 2013. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J 32, 1478–1488.

Timani, K.A., Liao, Q., Ye, Linbai, Zeng, Y., Liu, J., Zheng, Y., Ye, Li, Yang, X., Lingbao, K., Gao, J., Zhu, Y., 2005. Nuclear/nucleolar localization properties of C-terminal nucleocapsid protein of SARS coronavirus. Virus Res 114, 23–34.

Wang, H., Cui, X., Cai, X., An, T., 2022. Recombination in Positive-Strand RNA Viruses. Frontiers in Microbiology 13:870759.

Wu, Y., Zhao, S., 2021. Furin cleavage sites naturally occur in coronaviruses. Stem Cell Research 50, 102115.

Zhang, L., Mann, M., Syed, Z.A., Reynolds, H.M., Tian, E., Samara, N.L., Zeldin, D.C., Tabak, L.A., Ten Hagen, K.G., 2021. Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation. Proc Natl Acad Sci U S A 118, e2109905118.

## Declarations

No competing interests.

No specific funding for this work.

## Data Availability Statement

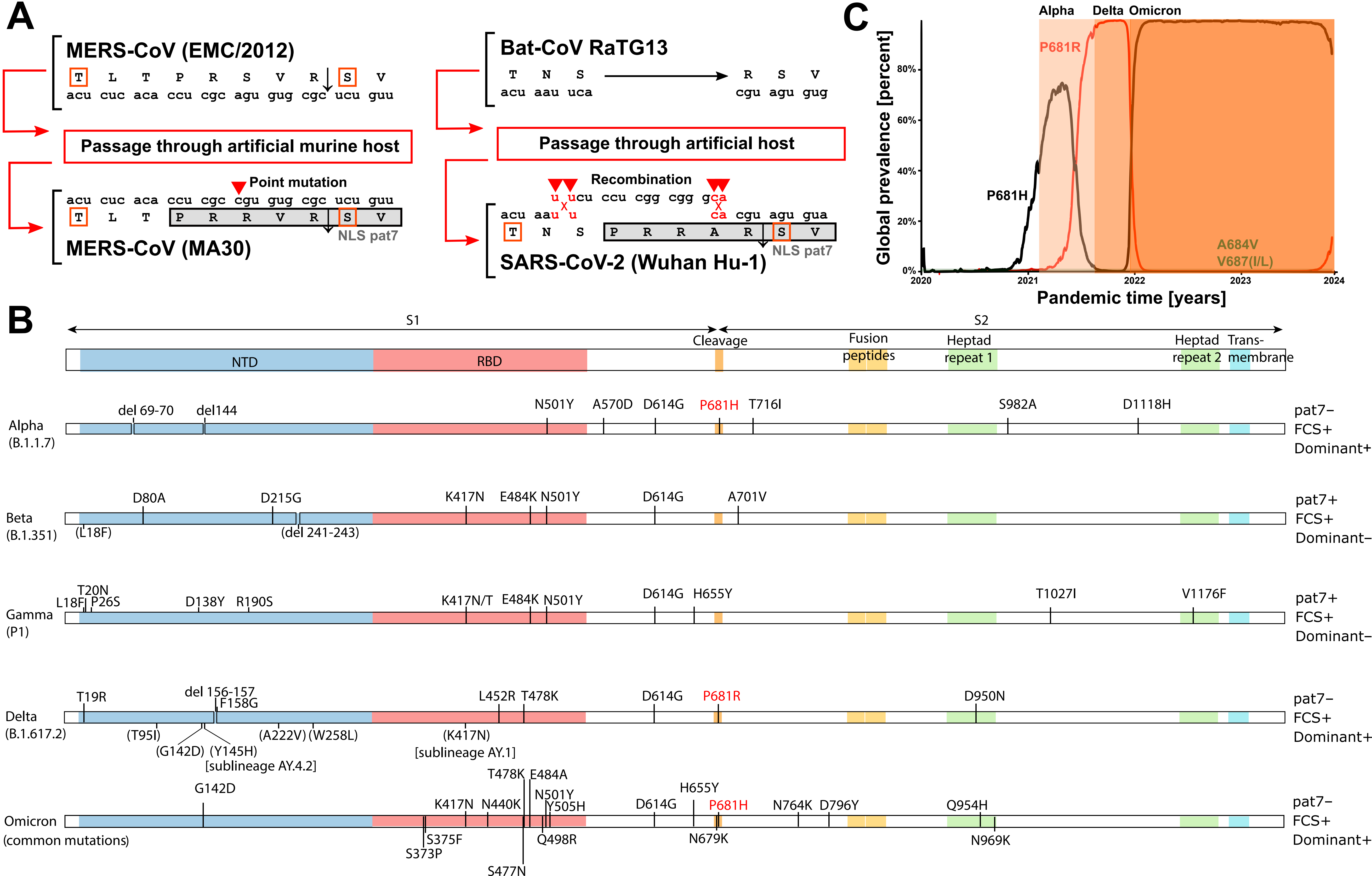
All data is included in the manuscript and/or in Supplementary Material (SI) dataset

## Tables

Table 1. Betacoronavirus spike (S) protein sequence S1/S2 domains from the InterPro/Uniprot IPR042578 collection. Furin cleavage sites of length 20 amino acids (AA) were predicted with FindFur algorithm (Gu, 2020), and filtered for nuclear localization signal (NLS) pat7. *O*-glycosilation sites were predicted with NetOGlyc4.0 (Steentoft et al., 2013). Asterisks (*) symbols denote *O*-glycosite residues, circumflex (^) symbol for furin cleavage site (FCS) minimal consensus motif RXXR, and horizontal (red) bars for pat7 NLS with a length of 7 amino acid residues.

## Figures

Figure 1. (A) Analogy between $\mathrm{MERS_{MA30}}$ and SARS-CoV-2 spike S1/S2 glycoprotein domains. $\mathrm{MERS_{MA30}}$ genotype and pat7/FCS combined phenotype after point mutation (a > c) during passage and artificial selection in humanized mice (left panel). Recombination origin model of the pat7/FCS domain in SARS-CoV-2 (right panel). Boxed amino acid letters indicate flanking Ser/Thr glycosites. Down arrow symbol (↓) indicates cleavage site. (B) Spike glycoprotein sequence domain representation and pandemic variants of concern in chronological order along with their  characteristic substitutions. (C) Global prevalence of the P681/(H/R) genotype during the course of the SARS-CoV-2 pandemic; shaded boxes indicate time intervals of dominant variants. A684V and V687(I/L) as negative controls below 0.5%.


A
MERS-CoV (EMC/2012)
T L T P R S V R S V
acu cuc aca ccu cgc agu gug cgc ucu guu
Passage through artificial murine host
Point mutation
acu cuc aca ccu cgc cgu gug cgc ucu guu
T L T P R R V R S V
NLS pat7
MERS-CoV (MA30)
Bat-CoV RaTG13
T N S R S V
acu aau uca cgu agu gug
Passage through artificial host
Recombination
u ucu ccu cgg cgg gca
acu aau u ca cgu agu gua
T N S P R R A R S V
NLS pat7
SARS-CoV-2 (Wuhan Hu-1)
B
S1
S2
Cleavage
Fusion peptides
Heptad repeat 1
Heptad repeat 2
Trans-membrane
NTD
RBD
Alpha (B.1.1.7)
del 69-70
del144
N501Y
A570D
D614G
P681H
T716I
S982A
D1118H
pat7−
FCS+
Dominant+
Beta (B.1.351)
D80A
D215G
K417N
E484K
N501Y
D614G
A701V
(L18F)
(del 241-243)
pat7+
FCS+
Dominant−
Gamma (P1)
T20N
L18F
P26S
D138Y
R190S
K417N/T
E484K
N501Y
D614G
H655Y
T1027I
V1176F
pat7+
FCS+
Dominant−
Delta (B.1.617.2)
T19R
del 156-157
F158G
L452R
T478K
D614G
P681R
D950N
(T95I)
(G142D) (Y145H)
[sublineage AY.4.2]
(A222V) (W258L)
(K417N)
[sublineage AY.1]
pat7−
FCS+
Dominant+
Omicron (common mutations)
G142D
T478K
E484A
N501Y
Y505H
K417N
N440K
H655Y
D614G
P681H
N764K
D796Y
Q954H
S375F
S373P
S477N
Q498R
N679K
N969K
pat7−
FCS+
Dominant+
C
Alpha
Delta
Omicron
P681R
P681H
A684V
V687(I/L)
Global prevalence [percent]
80%
60%
40%
20%
0%
2020
2021
2022
2023
2024
Pandemic time [years]

| | BetaCoV Spike Protein Domain<br>(20 AA segment at S1/S2 junction) | InterPro/UniProt Identifier<br>(Representative) | Betacoronavirus<br>(InterPro/UniProt description) | FCS<br>(RXXR) | NLS<br>(pat7) | *O*-Glycosite Pair<br>([S/T]XXPXXXX[S/T]) |
|---|---|---|---|---|---|---|
| 1 | ASYQTQTNSPRRARSVASQS<br>*   ^  ^* | A0A8A6U9B3 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 2 | ASYHTQTNSPRRARSVASQS<br>*   ^  ^* | A0A8B1JLN3 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 3 | ASYQTHTNSPRRARSVASQS<br>*   ^  ^* | A0A8B1JBP8 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 4 | ASYQTPTNSPRRARSVASQS<br>*   ^  ^* | A0A8B1J7Y2 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 5 | ASYQTQTKSPRRARSVASQS<br>*   ^  ^* | A0A8B1J577 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 6 | ASYQTQTNSPRRARSIASQS<br>*   ^  ^* | A0A8B6RWV4 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 7 | ASYQTQTNSPRRARSLASQS<br>*   ^  ^* | A0A7U3EFX5 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 8 | ASYQTQTNSPRRARSVAIQS<br>*   ^  ^* | A0A8B1JNU6 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 9 | ASYQTQTNSPRRVRSVASQS<br>*   ^  ^* | A0A8B1KFT9 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 10 | ASYHTQTNSPRRARSVASQS<br>*   ^  ^* | A0A8B6R9J1 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 11 | ASYQTHTNSPRRARSVASQS<br>*   ^  ^* | A0A8B1JSL0 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 12 | ASYQTQTKSPRRARSVASQS<br>*   ^  ^* | A0A8B6RKN5 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 13 | ASYQTQTNSPRRARSLASQS<br>*   ^  ^* | A0A7U3EFX5 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 14 | ASYQTQTNSPRRARSVASQS<br>*   ^  ^* | A0A6G5ZVU5 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 15 | GAGICASYSPRRARSVASQS<br>*   ^  ^* | A0A8A1NDM2 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 16 | ASYQTHTNSPRRARSVASQS<br>*   ^  ^* | A0A8B1J5Y9 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 17 | ASYQTQTNSPRRARSLASQS<br>*   ^  ^* | A0A7U3EFK5 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 18 | ASYQTQTNSPRRARSVVSQS<br>*   ^  ^* | A0A8B0GA47 | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) | + | + | + |
| 19 | LPDTPSTLTPRRVRSVPGEM<br>*   ^  ^* | A0A7D5BTZ9 | Middle East respiratory syndrome-related coronavirus (MERS-CoV)[1] | + | + | + |
| 20 | LPDTPSTLTPRSVRSVPGEM<br>*   ^  ^* | K0BRG7 | Middle East respiratory syndrome-related coronavirus (MERS-CoV)[2] | + | – | + |
| 21 | SGFCIDYALPSSRRKRRGIS<br>^  ^ | Q0ZJI1 | Human coronavirus HKU1 (HCoV-HKU1)[3] | + | + | – |
| 22 | ASYQTQTNS----RSVASQS | A0A6B9WHD3 | Bat coronavirus RaTG13[4] | – | – | – |

[1] Murine adapted (MA30), synthetic MERS clone from 2017

[2] Human betacoronavirus 2c EMC/2012 (first MERS isolate)

[3] Genotype B (not pat7-negative first 2005 isolate N1 with genotype A)

[4] Closest evolutionary BatCoV relative of SARS-CoV-2 genomic sequence (as of 2023); identical data for BatCoV BANAL-20-52

FCS: Furin Cleavage Site
NLS: Nulcear Localization Signal