BMC Genomic Data

RESEARCH

Open Access

Probable human origin of the SARS-CoV-2 polybasic furin cleavage motif

Antonio R. Romeu[1*]

Abstract

Background The key evolutionary step leading to the pandemic virus was the acquisition of the PRRA furin cleavage motif at the spike glycoprotein S1/S2 junction by a progenitor of SARS-CoV-2. Two of its features draw attention: (*i*) it is absent in other known lineage B beta-coronaviruses, including the newly discovered coronaviruses in bats from Laos and Vietnam, which are the closest known relatives of the covid virus; and, (*ii*) it introduced the pair of arginine codons (CGG-CGG), whose usage is extremely rare in coronaviruses. With an occurrence rate of only 3%, the arginine CGG codon is considered a minority in SARS CoV-2. On the other hand, Laos and Vietnam bat coronaviruses contain receptor-binding domains that are almost identical to that of SARS-CoV-2 and can therefore infect human cells despite the absence of the furin cleavage motif.

Results Based on these data, the aim of this work is to provide a detailed sequence analysis between the SARS-CoV-2 S gene insert encoding PRRA and the human mRNA transcripts. The result showed a 100% match to several mRNA transcripts. The set of human genes whose mRNAs match this S gene insert are ubiquitous and highly expressed, e.g., the ATPase F1 (ATP5F1) and the ubiquitin specific peptidase 21 (USP21) genes; or specific genes of target organs or tissues of the SARS-CoV-2 infection (e.g., MEMO1, SALL3, TRIM17, CWC15, CCDC187, FAM71E2, GAB4, PRDM13). Results suggest that a recombination between the genome of a SARS-CoV-2 progenitor and human mRNA transcripts could be the origin of the S gene 12-nucleotide insert encoding the S protein PRRA motif.

Conclusions The hypothesis of probable human origin of the SARS-CoV-2 polybasic furin cleavage motif is supported by: (*i*) the nature of human genes whose mRNA sequence 100% match the S gene insert; (*ii*) the synonymous base substitution in the arginine codons (CGG-CGG); and (*iii*) further spike glycoprotein PRRA-like insertions suggesting that the acquisition of PRRA may not have been a single recombination event.

Keywords SARS-CoV-2, Polybasic furin cleavage motif, Human mRNA transcripts, Spike glycoprotein insertions, Bioinformatics

Background

The key evolutionary step leading to the pandemic virus was the acquisition of the furin polybasic motif at the spike glycoprotein S1/S2 junction by a progenitor of SARS-CoV-2. In the first SARS-CoV-2 clinical isolates it was proline (P), arginine (R), arginine and alanine (A) (PRRA) [1–5]. Two of its features draw attention: (*i*) PRRA is absent in other known sarbecoviruses (lineage B beta-coronaviruses), including the newly discovered coronaviruses in bats from Laos and Vietnam, which are the closest known relatives of the covid virus [6, 7]; and, (*ii*) it introduced the pair of arginine codons (CGG-CGG), whose usage is extremely rare in coronaviruses [8, 9]. With an occurrence rate of only 3%, the arginine CGG codon is considered a minority in SARS CoV-2 [10, 11]. On the other hand, Laos and Vietnam bat

*Correspondence:
Antonio R. Romeu
antonioramon.romeu@iubilo.urv.cat
[1] Biochemistry and Molecular Biology, University Rovira i Virgili, Tarragona, Spain

coronaviruses contain receptor-binding domains that are almost identical to that of SARS-CoV-2 and can therefore infect human cells despite the absence of the furin cleavage motif [6, 7]. Based on these data, the aim of this work is to provide a detailed sequence analysis between the SARS-CoV-2 S gene insert encoding PRRA and the human mRNA transcripts [12, 13]. The result showed a 100% match to several mRNA transcripts suggesting the hypothesis of a probable human origin of the PRRA coding sequence acquired through recombination by a progenitor of the SARS-CoV-2.

Results and discussion

SARS-CoV-2 S gene 12-nucleotide insert and human mRNA transcripts

With coordinate based on SARS-CoV-2 reference sequence [14], the S gene 12-nucleotide fragment encoding the PRRA polybasic motif is located within the TCA S680 codon. Figure 1 shows the two possibilities: "CT CCT CGG CGG G" or "T CCT CGG CGG GC" depending on whether the insertion took place between positions 1 and 2 or 2 and 3 of the TCA codon, respectively. Here the reverse complement of the two possible S gene inserts have been included in the similarity analysis with the mRNA transcripts. In the creation of new SARS-CoV-2 particles in infected human cells, both ssRNA(+) and ssRNA(−) SARS-CoV-2 genomes that coexist through the viral RNA-dependent RNA polymerase (RdRp) [16] are equally important with respect to a possible recombination with human mRNAs.

Sequence analysis showed that the SRAS-CoV-2 S gene 12-nucleotide fragments, potentially involved in the PRRA coding, 100% match to several NCBI human mRNA RefSeq transcripts (Tables 1, 2, 3, and 4).

Tables 1, 2, 3, and 4 show the curate RefSeq mRNA protein-coding human transcripts, labelled "NM_" that matched the S gene insert. They group together 44 human genes and their variants. All of the human mRNA fragments matching the 12-nucleotide viral sequences are located in specific exons. No functional trend has

A

	P14	*P13*	*P12*	*P11*	*P10*	*P9*	*P8*	*P7*	*P6*	*P5*	*P4*	*P3*	*P2*	*P1*	*P1'*	*P2'*	*P3'*	*P4'*	*P5'*	*P6'*	
BANAL-103	A	S	Y	Q	T	Q	T	N	S	–	–	–	–	R	S	V	A	S	Q	S	682
BANAL-236	A	S	Y	Q	T	Q	T	N	S	–	–	–	–	R	S	V	A	S	Q	S	682
RaTG13	A	S	Y	Q	T	Q	T	N	S	–	–	–	–	R	S	V	A	S	Q	S	687
BANAL-52	A	S	Y	Q	T	Q	T	N	S	–	–	–	–	R	S	V	A	S	Q	S	687
Rp22DB159	A	S	Y	Q	T	Q	T	N	S	–	–	–	–	R	S	V	A	S	Q	S	687
SARS-CoV-2 (Wuhan-Hu-1)	A	S	Y	Q	T	Q	T	N	S	**P**	**R**	**R**	**A**	R	S	V	A	S	Q	S	691
SARS-CoV-2 (WH04)	A	S	Y	Q	T	Q	T	N	S	**P**	**R**	**R**	**A**	R	S	V	A	S	Q	S	691
	*	*	*	*	*	*	*	*	*					*	*	*	*	*	*	*	

B

```
RaTG13          GCC AGT TAT CAG ACT CAA ACT AAT TCA----------------CGT AGT GTG GCC AGT CAA TCT  23620
                _A__S__Y__Q__T__Q_ _T__N__S__-__-__-__-__R__S__V__A__S__Q__S_                    687
BANAL-52        GCC AGT TAT CAG ACT CAA ACT AAT TCA----------------CGT AGT GTG GCC AGT CAA TCC  23572
                _A__S__Y__Q__T__Q_ _T__N__S__-__-__-__-__R__S__V__A__S__Q__S_                    687
Rp22DB159       GCC AGT TAT CAG ACT CAA ACT AAT TCA----------------CGT AGT GTG GCC AGT CAA TCT  23594
                _A__S__Y__Q__T__Q__T__N__S__-__-__-__-__R__S__V__A__S__Q__S_                     687
Wuhan-Hu-1      GCT AGT TAT CAG ACT CAG ACT AAT TCT CCT CGG CGG GCA CGT AGT GTA GCT AGT CAA TCC  23635
Wuhan-Hu-1      GCT AGT TAT CAG ACT CAG ACT AAT TCT CCT CGG CGG GCA CGT AGT GTA GCT AGT CAA TCC  23635
                _A__S__Y__Q__T__Q_ _T__N__S__P__R__R__A__R__S__V__A__S__Q__S_                    691
```

Fig. 1 SARS-CoV-2 spike glycoprotein polybasic furin cleavage site. Fragment of a multiple sequence alignment covering the SARS-CoV-2 spike glycoprotein polybasic furin cleavage site. The first line at the top indicates the positions, in a P14-P6′ nomenclature, of the canonical structure of a furin site in a given protein. The specific cleavage site is between positions P1 and P1′. The core regions is between positions P6–P2′ and there are two flanking solvent accessible regions: P7–P14 and P3′–P6′ [15]. Part **A**. Fragment of the protein multiple sequence aligment including Laos bat *Rhinolophus* coronaviruses BANAL-52 (GISAID, EPI_ISL_4302644: 21512–25,321), BANAL-103 (GISAID, EPI_ISL_4302645: 21498–25,294), BANAL-236 (GISAID, EPI_ISL_4302647: 21538–25,344), Vietnam bat *Rhinolophus pusillus* Rp22DB159 coronavirus (GenBank: WLJ60537.1 coded by OR233302.1:21533..25342 genome), Bat coronavirus RaTG13 (GenBank: QHR63300.2 coded by MN996532.2: 21560..25369 genome) and the reference SARS-CoV-2 sequences (isolates Wuhan-Hu-1 and WH04) [14]. The SARS-CoV-2 polybasic insert (PRRA) is denoted in bold. Strictly conserved amino acids are denoted by *. The amino acid position is indicated at the numbers on the right. Part **B**. Fragment of the codon alignment. For simplicity, from the Laos coronavirus only the BANAL-52 sequence has been included. The two possibles 12 nucleotide fragment encoding PRRA inserted within the S680 codon are highlighted in yellow and orange, respectively. The S680 TCA codon is denoted in green. The differences at the third codon position are denoted in gray. The protein and genome sequence position is indicated at the numbers on the right

Table 1 Human NCBI NM_RefSeq Transcripts (curated protein coding) that match the CTCCTCGGCGGG SARS-CoV-2 furin cleavage insert

Query and tranacript position	RefSeq GenBank title	Chr	Exon range
CTCCTCGGCGGG-2921	NM_000466.3 Homo sapiens peroxisomal biogenesis factor 1 (PEX1), transcript variant 1, mRNA	7	Exon, 2869..3011
CTCCTCGGCGGG-2750	NM_001282677.2 Homo sapiens peroxisomal biogenesis factor 1 (PEX1), transcript variant 2, mRNA	7	Exon, 2669..2803
CTCCTCGGCGGG-2956	NM_001282678.2 Homo sapiens peroxisomal biogenesis factor 1 (PEX1), transcript variant 3, mRNA	7	Exon, 2869..3011
CTCCTCGGCGGG-1643	NM_001099289.3 Homo sapiens SH3 domain containing ring finger 3 (SH3RF3), mRNA	2	Exon, 1636..1739
CTCCTCGGCGGG-869	NM_001145873.1 Homo sapiens CD8a molecule (CD8A), transcript variant 3, mRNA	2	Exon, 762..1080
CTCCTCGGCGGG-851	NM_001382698.1 Homo sapiens CD8a molecule (CD8A), transcript variant 5, mRNA	2	Exon, 744..1062
CTCCTCGGCGGG-3430	NM_020910.3 Homo sapiens KIAA1549 (KIAA1549), transcript variant 1, mRNA	7	Exon, 3790..3967
CTCCTCGGCGGG-3430	NM_001164665.2 Homo sapiens KIAA1549 (KIAA1549), transcript variant 2, mRNA	7	Exon, 3790..3967
CTCCTCGGCGGG-853	NM_001291291.2 Homo sapiens MISP family member 3 (MISP3), transcript variant 1, mRNA	19	Exon, 1..1092
CTCCTCGGCGGG-853	NM_001393577.1 Homo sapiens MISP family member 3 (MISP3), transcript variant 3, mRNA	19	Exon, 1..1092
CTCCTCGGCGGG-169	NM_004717.3 Homo sapiens diacylglycerol kinase iota (DGKI), transcript variant 1, mRNA	7	Exon, 1..403
CTCCTCGGCGGG-279	NM_001321708.2 Homo sapiens diacylglycerol kinase iota (DGKI), transcript variant 2, mRNA	7	Exon, 1..513
CTCCTCGGCGGG-145	NM_001321710.2 Homo sapiens diacylglycerol kinase iota (DGKI), transcript variant 4, mRNA	7	Exon, 1..379
CTCCTCGGCGGG-279	NM_001388092.1 Homo sapiens diacylglycerol kinase iota (DGKI), transcript variant 5, mRNA	7	Exon, 1..513
CTCCTCGGCGGG-7	NM_004093.4 Homo sapiens ephrin B2 (EFNB2), transcript variant 1, mRNA	13	Exon, 1..820
CTCCTCGGCGGG-7	NM_001372056.1 Homo sapiens ephrin B2 (EFNB2), transcript variant 2, mRNA	13	Exon, 1..820
CTCCTCGGCGGG-7	NM_001372057.1 Homo sapiens ephrin B2 (EFNB2), transcript variant 3, mRNA	13	Exon, 1..820
CTCCTCGGCGGG-7	NM_001372058.1 Homo sapiens ephrin B2 (EFNB2), transcript variant 4, mRNA	13	Exon, 1..820
CTCCTCGGCGGG-118	NM_004637.6 Homo sapiens RAB7A, member RAS oncogene family (RAB7A), mRNA	3	Exon, 1..177
CTCCTCGGCGGG-307	NM_006843.3 Homo sapiens serine dehydratase (SDS), mRNA	12	Exon, 276..315
CTCCTCGGCGGG-187	NM_016085.5 Homo sapiens all-trans retinoic acid induced differentiation factor (ATRAID), transcript variant 1, mRNA	2	Exon, 1..248
CTCCTCGGCGGG-1032	NM_021620.4 Homo sapiens PR/SET domain 13 (PRDM13), mRNA	6	Exon, 600..3129
CTCCTCGGCGGG-36	NM_022831.4 Homo sapiens axin interactor, dorsalization associated (AIDA), mRNA	1	Exon, 1..284
CTCCTCGGCGGG-2215	NM_171999.4 Homo sapiens spalt like transcription factor 3 (SALL3), mRNA	18	CDS*, 458..4360

The rows highlighted in gray denote the transcripts whose genes also match TCCTCGGCGGGC (Table 2)

*: in gene annotation, coding sequences (CDS)

Table 2 Human NCBI NM_RefSeq Transcripts (curated protein coding) that match the TCCTCGGCGGGC SARS-CoV-2 furin cleavage insert

Query transpt. post.	RefSeq GenBank title	Chr	Exon range
TCCTCGGCGGGC-180	NM_001001937.2 Homo sapiens ATP synthase F1 subunit alpha (ATP5F1A), transcript variant 1, mRNA; nuclear gene for mitochondrial product	18	Exon, 92..199
TCCTCGGCGGGC-106	NM_004046.6 Homo sapiens ATP synthase F1 subunit alpha (ATP5F1A), transcript variant 2, mRNA; nuclear gene for mitochondrial product	18	Exon, 1..125
TCCTCGGCGGGC-106	NM_001257334.2 Homo sapiens ATP synthase F1 subunit alpha (ATP5F1A), transcript variant 3, mRNA; nuclear gene for mitochondrial product	18	Exon, 1..125
TCCTCGGCGGGC-106	NM_001001935.3 Homo sapiens ATP synthase F1 subunit alpha (ATP5F1A), transcript variant 4, mRNA; nuclear gene for mitochondrial product	18	Exon, 1..125
TCCTCGGCGGGC-106	NM_001257335.2 Homo sapiens ATP synthase F1 subunit alpha (ATP5F1A), transcript variant 5, mRNA; nuclear gene for mitochondrial product	18	Exon, 1..461
TCCTCGGCGGGC-253	NM_001303447.2 Homo sapiens thioredoxin domain containing 11 (TXNDC11), transcript variant 1, mRNA	16	Exon, 1..381
TCCTCGGCGGGC-253	NM_015914.7 Homo sapiens thioredoxin domain containing 11 (TXNDC11), transcript variant 2, mRNA	16	Exon, 1..381
TCCTCGGCGGGC-253	NM_001324022.2 Homo sapiens thioredoxin domain containing 11 (TXNDC11), transcript variant 3, mRNA	16	Exon, 1..381
TCCTCGGCGGGC-253	NM_001324024.2 Homo sapiens thioredoxin domain containing 11 (TXNDC11), transcript variant 4, mRNA	16	Exon, 1..381
TCCTCGGCGGGC-253	NM_001324025.2 Homo sapiens thioredoxin domain containing 11 (TXNDC11), transcript variant 5, mRNA	16	Exon, 1..381
TCCTCGGCGGGC-5	NM_001371914.2 Homo sapiens mediator of cell motility 1 (MEMO1), transcript variant 9, mRNA	2	Exon, 1..15
TCCTCGGCGGGC-5	NM_001371916.2 Homo sapiens mediator of cell motility 1 (MEMO1), transcript variant 12, mRNA	2	Exon, 1..15
TCCTCGGCGGGC-308	NM_006843.3 Homo sapiens serine dehydratase (SDS), mRNA	12	Exon, 276..315
TCCTCGGCGGGC-188	NM_016085.5 Homo sapiens all-trans retinoic acid induced differentiation factor (ATRAID), transcript variant 1, mRNA	2	Exon, 1..248
TCCTCGGCGGGC-2216	NM_171999.4 Homo sapiens spalt like transcription factor 3 (SALL3), mRNA	18	CDS*, 458..4360

The rows highlighted in gray denote the transcripts whose genes also match CTCCTCGGCGGG (Table 1)

*: in gene annotation, coding sequences (CDS)

been observed in the gene products. Also, these genes are distributed throughout all chromosomes. This suggests that these human transcripts could be good candidates as donors of mRNA template in a potential recombination link to a SARS-CoV-2 furin cleavage motif. However, the presence of the CGG-CGG in these human mRNA transcripts does not necessarily imply there would be an arginine pair in the gene product. It may not be in the reading frame.

Tables 1S-4S show the results extended to the four series of human mRNA RefSeq transcripts: NM_ (curated mRNA protein-coding), NR_ (RNA non-protein-coding), XM_ (predicted model protein-coding) and XR_ (RNA predicted model non-protein-coding).

Tissue-specificity of genes whose mRNA transcripts 100% match viral sequences

The set of human genes whose mRNAs match the SARS-CoV-2 S gene PRRA coding region (Tables 1, 2, 3, and 4) can be grouped into two categories in terms of tissue-specificity: (*i*) ubiquitous and highly expressed; and (*ii*) specific to target organs or tissues

Table 3 Human NCBI NM_RefSeq Transcripts (curated protein coding) that match the CCCGCCGAGGAG (CTCCTCGGCGGG reverse complement) SARS-CoV-2 furin cleavage insert

Query and tranacript position	RefSeq GenBank title	Chr	Exon range
CCCGCCGAGGAG-1518	NM_001003699.4 Homo sapiens ras responsive element binding protein 1 (RREB1), transcript variant 1, mRNA	6	Exon, 1282..4192
CCCGCCGAGGAG-1444	NM_001168344.2 Homo sapiens ras responsive element binding protein 1 (RREB1), transcript variant 2, mRNA	6	Exon, 1208..4118
CCCGCCGAGGAG-1518	NM_001003698.4 Homo sapiens ras responsive element binding protein 1 (RREB1), transcript variant 3, mRNA	6	Exon, 1282..4192
CCCGCCGAGGAG-1518	NM_001003700.2 Homo sapiens ras responsive element binding protein 1 (RREB1), transcript variant 4, mRNA	6	Exon, 1282..4192
CCCGCCGAGGAG-1209	NM_012475.5 Homo sapiens ubiquitin specific peptidase 21 (USP21), transcript variant 1, mRNA	1	Exon, 1094..1249
CCCGCCGAGGAG-1350	NM_001014443.3 Homo sapiens ubiquitin specific peptidase 21 (USP21), transcript variant 3, mRNA	1	Exon, 1235..1390
CCCGCCGAGGAG-1297	NM_001319847.2 Homo sapiens ubiquitin specific peptidase 21 (USP21), transcript variant 4, mRNA	1	Exon, 1182..1337
CCCGCCGAGGAG-1209	NM_001319848.2 Homo sapiens ubiquitin specific peptidase 21 (USP21), transcript variant 5, mRNA	1	Exon, 1094..1249
CCCGCCGAGGAG-763	NM_016102.4 Homo sapiens tripartite motif containing 17 (TRIM17), transcript variant 1, mRNA	1	Exon, 324..793
CCCGCCGAGGAG-727	NM_001024940.3 Homo sapiens tripartite motif containing 17 (TRIM17), transcript variant 2, mRNA	1	Exon, 476..909
CCCGCCGAGGAG-879	NM_001134855.2 Homo sapiens tripartite motif containing 17 (TRIM17), transcript variant 4, mRNA	1	Exon, 476..909
CCCGCCGAGGAG-1631	NM_001037814.1 Homo sapiens GRB2 associated binding protein family member 4 (GAB4), transcript variant 1, mRNA	22	Exon, 1585..1689
CCCGCCGAGGAG-1172	NM_001366857.1 Homo sapiens GRB2 associated binding protein family member 4 (GAB4), transcript variant 2, mRNA	22	Exon, 1126..1230
CCCGCCGAGGAG-1649	NM_001080461.3 Homo sapiens UNC homeobox (UNCX), mRNA	7	Exon, 597..2091
CCCGCCGAGGAG-2090	NM_020773.3 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 1, mRNA	4	Exon, 1901..2159
CCCGCCGAGGAG-2004	NM_001113361.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 2, mRNA	4	Exon, 1815..2073
CCCGCCGAGGAG-1245	NM_001113363.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 3, mRNA	4	Exon, 1056..1314
CCCGCCGAGGAG-1161	NM_001286805.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 4, mRNA	4	Exon, 972..1230
CCCGCCGAGGAG-1398	NM_001330638.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 5, mRNA	4	Exon, 1209..1467
CCCGCCGAGGAG-429	NM_013314.4 Homo sapiens B cell linker (BLNK), transcript variant 1, mRNA	10	Exon, 376..532
CCCGCCGAGGAG-429	NM_001114094.2 Homo sapiens B cell linker (BLNK), transcript variant 2, mRNA	10	Exon, 376..532
CCCGCCGAGGAG-429	NM_001258440.2 Homo sapiens B cell linker (BLNK), transcript variant 3, mRNA	10	Exon, 376..532
CCCGCCGAGGAG-429	NM_001258441.2 Homo sapiens B cell linker (BLNK), transcript variant 4, mRNA	10	Exon, 376..532
CCCGCCGAGGAG-429	NM_001258442.2 Homo sapiens B cell linker (BLNK), transcript variant 5, mRNA	10	Exon, 376..532
CCCGCCGAGGAG-61	NM_001199563.2 Homo sapiens blood vessel epicardial substance (BVES), transcript variant C, mRNA	6	Exon, 1..189
CCCGCCGAGGAG-547	NM_012464.5 Homo sapiens tolloid like 1 (TLL1), transcript variant 1, mRNA	4	Exon, 1..837
CCCGCCGAGGAG-547	NM_001204760.2 Homo sapiens tolloid like 1 (TLL1), transcript variant 2, mRNA	4	Exon, 1..837
CCCGCCGAGGAG-1037	NM_012292.5 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 1, mRNA	19	Exon, 1030..1167
CCCGCCGAGGAG-913	NM_001258328.4 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 2, mRNA	19	Exon, 906..1043
CCCGCCGAGGAG-526	NM_001282335.3 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 4, mRNA	19	Exon, 519..656
CCCGCCGAGGAG-885	NM_001321232.2 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 5, mRNA	19	Exon, 878..1015
CCCGCCGAGGAG-189	NM_001287135.2 Homo sapiens cyclin dependent kinase 14 (CDK14), transcript variant 1, mRNA	7	Exon, 1..398
CCCGCCGAGGAG-2603	NM_015149.6 Homo sapiens ral guanine nucleotide dissociation stimulator like 1 (RGL1), transcript variant 1, mRNA	1	Exon, 2535..2649
CCCGCCGAGGAG-2602	NM_001297669.3 Homo sapiens ral guanine nucleotide dissociation stimulator like 1 (RGL1), transcript variant 2, mRNA	1	Exon, 2534..2648
CCCGCCGAGGAG-2492	NM_001297670.3 Homo sapiens ral guanine nucleotide dissociation stimulator like 1 (RGL1), transcript variant 3, mRNA	1	Exon, 2424..2538
CCCGCCGAGGAG-2251	NM_001297671.3 Homo sapiens ral guanine nucleotide dissociation stimulator like 1 (RGL1), transcript variant 4, mRNA	1	Exon, 2183..2297
CCCGCCGAGGAG-2164	NM_001297672.3 Homo sapiens ral guanine nucleotide dissociation stimulator like 1 (RGL1), transcript variant 5, mRNA	1	Exon, 2096..2210
CCCGCCGAGGAG-226	NM_001363371.2 Homo sapiens CWC15 spliceosome associated protein homolog (CWC15), transcript variant 1, mRNA	11	Exon, 1..284
CCCGCCGAGGAG-226	NM_001363372.2 Homo sapiens CWC15 spliceosome associated protein homolog (CWC15), transcript variant 2, mRNA	11	Exon, 1..252
CCCGCCGAGGAG-4557	NM_001378188.1 Homo sapiens coiled-coil domain containing 187 (CCDC187), transcript variant 3, mRNA	9	Exon, 4475..4560
CCCGCCGAGGAG-1755	NM_032382.5 Homo sapiens component of oligomeric golgi complex 8 (COG8), transcript variant 1, mRNA	16	Exon, 1595..1877
CCCGCCGAGGAG-1896	NM_001379261.1 Homo sapiens component of oligomeric golgi complex 8 (COG8), transcript variant 3, mRNA	16	Exon, 1736..2018
CCCGCCGAGGAG-1755	NM_001379262.1 Homo sapiens component of oligomeric golgi complex 8 (COG8), transcript variant 4, mRNA	16	Exon, 1595..1771
CCCGCCGAGGAG-1794	NM_001379263.1 Homo sapiens component of oligomeric golgi complex 8 (COG8), transcript variant 5, mRNA	16	Exon, 1595..1916
CCCGCCGAGGAG-1755	NM_001379264.1 Homo sapiens component of oligomeric golgi complex 8 (COG8), transcript variant 8, mRNA	16	Exon, 1595..1874
CCCGCCGAGGAG-667	NM_001394.7 Homo sapiens dual specificity phosphatase 4 (DUSP4), transcript variant 1, mRNA	8	Exon, 1..839
CCCGCCGAGGAG-1873	NM_003249.5 Homo sapiens thimet oligopeptidase 1 (THOP1), mRNA	19	Exon, 1803..1931
CCCGCCGAGGAG-776	NM_004327.4 Homo sapiens BCR activator of RhoGEF and GTPase (BCR), transcript variant 1, mRNA	22	Exon, 1..1731
CCCGCCGAGGAG-776	NM_021574.3 Homo sapiens BCR activator of RhoGEF and GTPase (BCR), transcript variant 2, mRNA	22	Exon, 1..1731
CCCGCCGAGGAG-2797	NM_005157.6 Homo sapiens ABL proto-oncogene 1, non-receptor tyrosine kinase (ABL1), transcript variant a, mRNA	9	Exon, 1872..5578
CCCGCCGAGGAG-3938	NM_007313.3 Homo sapiens ABL proto-oncogene 1, non-receptor tyrosine kinase (ABL1), transcript variant b, mRNA	9	Exon, 3013..6719
CCCGCCGAGGAG-821	NM_013962.3 Homo sapiens neuregulin 1 (NRG1), transcript variant GGF2, mRNA	1	Exon, 1..1185
CCCGCCGAGGAG-4884	NM_015254.4 Homo sapiens kinesin family member 13B (KIF13B), mRNA	8	Exon, 4558..5254
CCCGCCGAGGAG-43	NM_031866.3 Homo sapiens frizzled class receptor 8 (FZD8), mRNA	10	Exon, 1..4050
CCCGCCGAGGAG-427	NM_080825.4 Homo sapiens chromosome 20 open reading frame 144 (C20orf144), mRNA	20	Exon, 146..522
CCCGCCGAGGAG-776	NM_001348758.2 Homo sapiens chromosome 2 open reading frame 42 (C2orf42), transcript variant 1, mRNA	2	Exon,1..836

The rows highlighted in gray denote the transcripts whose genes also match GCCCGCCGAGGA (Table 4)

of the SARS-CoV-2 infection. As an example, Fig. 2 shows some examples of tissue specificity. In the first group, the alpha subunit of the ATPase F1 (ATP5F1) gene encodes the catalytic core of the mitochondrial ATP synthase; and the ubiquitin specific peptidase 21 (USP21) gene-encoding protein cleaves ubiquitin for recycling in intracellular protein degradation. In the second group there are genes with a tissue specificity in the brain, kidney, prostate and testis, which are targets in virus infection [17] (e.g., MEMO1, SALL3, TRIM17, CWC15). However, some genes must be highlighted as unique and highly expressed in the testis (e.g., CCDC187, FAM71E2, GAB4, PRDM13).

Table 4 Human NCBI NM_RefSeq Transcripts (curated protein coding) that match the GCCCGCCGAGGA (TCCTCGGCGGGC reverse complement) SARS-CoV-2 furin cleavage insert

Query and tranacript position	RefSeq GenBank title	Chr	Exon range
GCCCGCCGAGGA-762	NM_016102.4 Homo sapiens tripartite motif containing 17 (TRIM17), transcript variant 1, mRNA	1	Exon, 324..793
GCCCGCCGAGGA-726	NM_001024940.3 Homo sapiens tripartite motif containing 17 (TRIM17), transcript variant 2, mRNA	1	Exon, 324..757
GCCCGCCGAGGA-878	NM_001134855.2 Homo sapiens tripartite motif containing 17 (TRIM17), transcript variant 4, mRNA	11	Exon, 476..909
GCCCGCCGAGGA-217	NM_025112.5 Homo sapiens ZXD family zinc finger C (ZXDC), transcript variant 1, mRNA	3	Exon, 1..933
GCCCGCCGAGGA-217	NM_001040653.4 Homo sapiens ZXD family zinc finger C (ZXDC), transcript variant 2, mRNA	3	Exon, 1..933
GCCCGCCGAGGA-123	NM_001042424.3 Homo sapiens nuclear receptor binding SET domain protein 2 (NSD2), transcript variant 10, mRNA	4	Exon, 1..150
GCCCGCCGAGGA-2089	NM_020773.3 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 1, mRNA	4	Exon, 1901..2159
GCCCGCCGAGGA-2003	NM_001113361.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 2, mRNA	4	Exon, 1815..2073
GCCCGCCGAGGA-1244	NM_001113363.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 3, mRNA	4	Exon, 1056..1314
GCCCGCCGAGGA-1160	NM_001286805.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 4, mRNA	4	Exon, 972..1230
GCCCGCCGAGGA-1397	NM_001330638.2 Homo sapiens TBC1 domain family member 14 (TBC1D14), transcript variant 5, mRNA	4	Exon, 1209..1467
GCCCGCCGAGGA-428	NM_013314.4 Homo sapiens B cell linker (BLNK), transcript variant 1, mRNA	10	Exon, 376..532
GCCCGCCGAGGA-428	NM_001114094.2 Homo sapiens B cell linker (BLNK), transcript variant 2, mRNA	10	Exon, 376..532
GCCCGCCGAGGA-428	NM_001258440.2 Homo sapiens B cell linker (BLNK), transcript variant 3, mRNA	10	Exon, 376..532
GCCCGCCGAGGA-428	NM_001258441.2 Homo sapiens B cell linker (BLNK), transcript variant 4, mRNA	10	Exon, 376..532
GCCCGCCGAGGA-428	NM_001258442.2 Homo sapiens B cell linker (BLNK), transcript variant 5, mRNA	10	Exon, 376..532
GCCCGCCGAGGA-457	NM_001145402.2 Homo sapiens family with sequence similarity 71 member E2 (FAM71E2), mRNA	19	CDS, 195..2963
GCCCGCCGAGGA-1036	NM_012292.5 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 1, mRNA	19	Exon, 1030..1167
GCCCGCCGAGGA-912	NM_001258328.4 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 2, mRNA	19	Exon, 906..1043
GCCCGCCGAGGA-525	NM_001282335.3 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 4, mRNA	19	Exon, 519..656
GCCCGCCGAGGA-884	NM_001321232.2 Homo sapiens Rho GTPase activating protein 45 (ARHGAP45), transcript variant 5, mRNA	19	Exon, 878..1015
GCCCGCCGAGGA-188	NM_001287135.2 Homo sapiens cyclin dependent kinase 14 (CDK14), transcript variant 1, mRNA	7	Exon, 1..398
GCCCGCCGAGGA-225	NM_001363371.2 Homo sapiens CWC15 spliceosome associated protein homolog (CWC15), transcript variant 1, mRNA	11	Exon, 1..284
GCCCGCCGAGGA-225	NM_001363372.2 Homo sapiens CWC15 spliceosome associated protein homolog (CWC15), transcript variant 2, mRNA	11	Exon, 1..252
GCCCGCCGAGGA-666	NM_001394.7 Homo sapiens dual specificity phosphatase 4 (DUSP4), transcript variant 1, mRNA	8	Exon, 1..839
GCCCGCCGAGGA-1495	NM_004170.6 Homo sapiens solute carrier family 1 member 1 (SLC1A1), mRNA	9	Exon, 1406..1540
GCCCGCCGAGGA-775	NM_004327.4 Homo sapiens BCR activator of RhoGEF and GTPase (BCR), transcript variant 1, mRNA	22	Exon, 1..1731
GCCCGCCGAGGA-775	NM_021574.3 Homo sapiens BCR activator of RhoGEF and GTPase (BCR), transcript variant 2, mRNA	22	Exon, 1..1731
GCCCGCCGAGGA-820	NM_013962.3 Homo sapiens neuregulin 1 (NRG1), transcript variant GGF2, mRNA	1	Exon, 1..1185

The rows highlighted in gray denote the transcripts whose genes also match CCCGCCGAGGAG (Table 3)

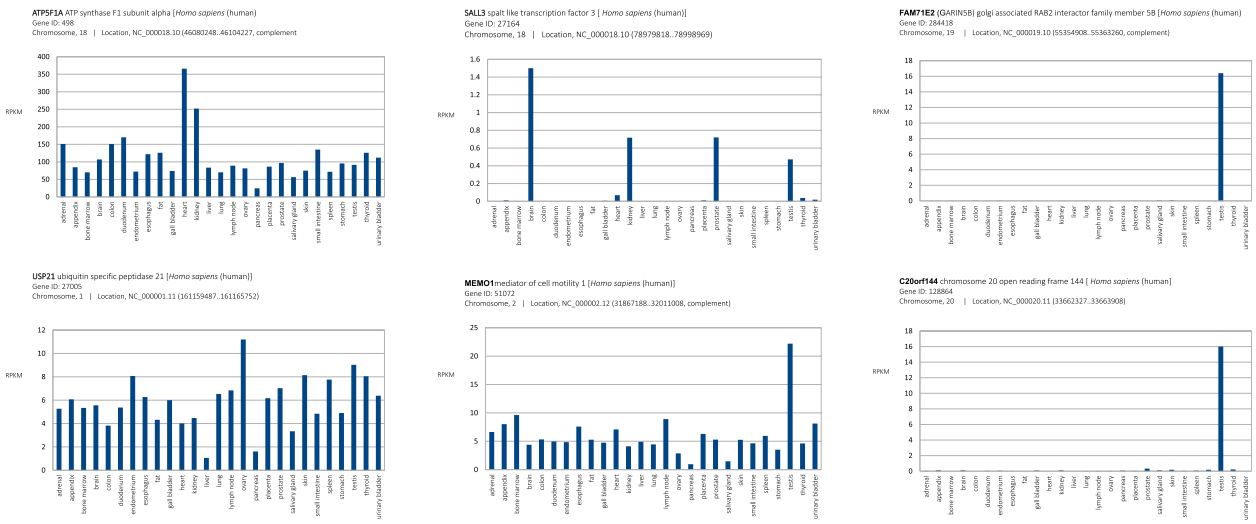

Fig. 2 Tissue specificity of human genes matching SARS-CoV-2 S gene coding PRRA insert. Examples of tissue specificity of human genes (shown in Tables 1, 2, 3, and 4) whose transcripts match 100% with SARS-CoV-2 S gene insert encoded the PRRA polybasic furin motif. Expression pattern based on the Human Protein Atlas RNA-seq normal tissues NCBI BioProject Accession: PRJEB4337 ID: 231263. Data were download from NCBI Human Genome Resources, assembly GRCh38.p14. Units of transcript expression are normalized reads per kilobase of transcript, per million mapped reads (RPKM)

Hypothesis of probable human origin of the SARS-CoV-2 polybasic furin cleavage site

The 100% match between the S gene insert encoding the furin polybasic motif and some human mRNA transcripts suggests that a recombination between the viral and human RNA could be the origin of that S gene insert. This is the proposal of the hypothesis on the probable human origin of the SARS-CoV-2 polybasic furin

cleavage mptif, which agrees with a possibility already put forward by R. F. Garry and co-workers [1] that a progenitor of SARS-CoV-2 passed to humans, acquiring the PRRA during undetected human-to-human transmission.

Recombination is the common method by which viruses acquire new skills [18]. In this case the skill was the ability to interact with the human furin serin protease, which further aids the entry of SARS-CoV-2 into human cells.

Evidence supporting the hypothesis

Evidence 1. Synonymous base substitution at the SARS-CoV-2 S gene arginine codons CGG-CGG

The arginine pair of the furin polybasic motif is essential for the SARS-CoV-2. It is evolutionarily strictly conserved. Although mutations of the variants also extend to the PRRA, the RR pair remains e.g., Delta P/R (RRRA), Omicron P/histidine, H (HRRA). In contrast, this key arginine pair is encoded by the "extremely rare" CGG-CGG codons in coronaviruses. So, everything suggests an evolutionary pressure in that point of the SARS-CoV-2 S gene.

Sequence analysis shows that a synonymous base substitution in the furin arginine pair code actually occurs. Based on the NCBI Virus database, a large sample of spike glycoprotein squences (with release date from January 1, 2020 to September 30, 2023) 8459 out of 3,494,735 (0.2420%) protein sequences showed arginine codon usage bias in one of the CGG-CGG codons (Table 5S). Also, based on data from the GISAID database, 155 out of 78,085 (0.1985%) showed the same arginine codon usage bias (Table 6S). The sequences that have a synonymous base substitution in one of the CGG arginine codons at the furin arginine pair cover several SARS-CoV-2 lineages and geographic regions.

When the analyses focused on specific SARS-CoV-2 lineages, results were significant. Based on a GISAID sample of isolates from the XBB.1.16.20 lineage 75 out of 354 (21.19%) showed a synonymous base substitution in the second arginine codon of the furin polybasic motif, the CGG-CGG code has mutated to CGG-CGT. Isolates from the EE.2 lineage, 533 out of 1021 (52.20%) the CGG-CGG was replaced by CGA-CGG. Furthermore, isolates from the CQ.2, CQ.1 and CQ.1.1 lineages the 100% of the analysed spike glycoprotein sequences showed arginine codon usage bias in one of the CGG pair. The results were as follows: CQ.2 lineage, 286 out of 287 (99.65%); CQ.1516 out of 516 (100%); CQ.1.1117 out of 177 (100%). Table 7S shows basic information of these SARS-CoV-2 lineage isolates.

These results suggest a SARS-CoV-2 S gene trend towards an arginine codon usage bias encoding the spike glycoprotein furin polybasic motif.

Evidence 2. PRRA-like insertions in the SARS-CoV-2 spike glycoprotein sequence

Once SARS-CoV-2 emerged, the question is whether there have been further spike glycoprotein insertions similar to that of the furin polybasic motif in a SARS-CoV-2 progenitor. Based on the NCBI Virus database, the analysis of 2,315,308 spike glycoprotein sequences with no ambiguous characters showed many other PRRA-like insertions throughout the sequence, in the different spike glycoprotein structural domains (Fig. 3). Table 5 shows

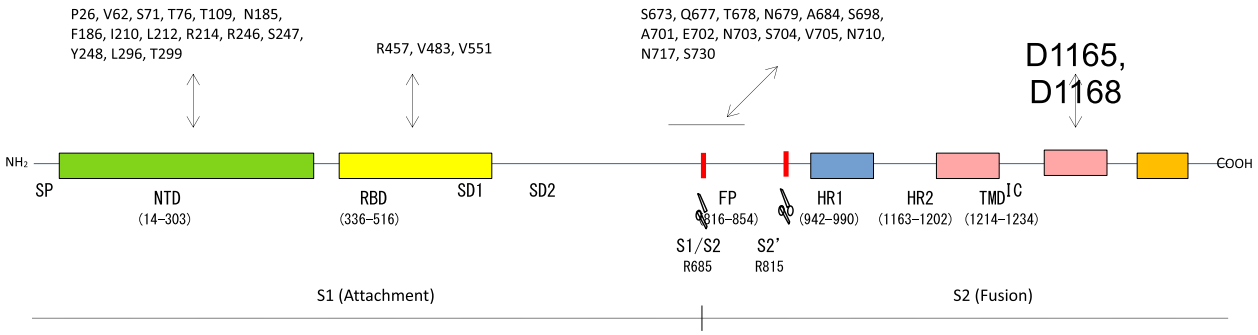

Fig. 3 PRRA-like insertion along the SARS-CoV-2 spike glycoprotein. PRRA-like insertion positions in the SARS-CoV-2 spike glycoprotein. With coordinate based on SARS-CoV reference sequence [14], boxes on the top show the sequence positions of the PRRA-like insertions. The SARS-CoV-2 spike glycoprotein structural domains and cleavage sites are indicated. Protein length: 1273. At the bottom the S1 subunit and S2 subunit in each spike glycoprotein monomer are indicated. The coordinates of some of the domains are in parentheses. Acronyms: S1, subunit 1; S2, subunit 2; SP, signal peptide; NTD. N-terminal domain; RBD, receptor binding domain; SD1, subdomain 1; SD2, subdomain 2; FP, fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; TM, transmembrane region; IC, inracellular domain. Cleavage sites: S1/S2, R685/S686, by host (human) furin, paired basic amino acid cleaving enzyme (FURIN); S2', R815/S816, by host transmembrane serine protease 2 (TMPRSS2) orcathepsin L (CTSL). The figure has been created based on data from [19–21]

Table 5 Spike glycoprotein PRRA-like insertions from SARS-CoV-2 isolates

Position	Insert	Genbank Accession	Number
		N-terminal domain	
P26	IEKN	UFA95489	1
V62	TSGNN	QZQ09947	1
S71	HLKD	UEP74721	1
T76	AVMSL	UEV78566	1
T109	ILWW	UHT93004	1
N185	MQAVS	UWL01218	1
	SARW	UYR39941	1
F186	TPAGG	WCF88733	1
I210	KKGA	UWJ32324	1
	RHAVLSG	WDD77904	1
L212	FMAE	UPH02236, UPH02715, UPL60987, UPL79201, UPL83322, UPL97779, UPS43182, UPW23099, UPW23870, UZZ69309, URP81062, URP81074, URP81794, URG88035, UQE96806, UQE28102	16
	TVGG	UPQ93706, UPQ95992, UYL90137, UYL90173, UPN04511, UQR90517, UQR92623, UQS02680, UQS06834, UQS05613, UQE37288	11
	NLTI	WED35345	1
	REPEDR	UJS73581	1
R214	ASPN	UIS34449	1
	DQAF	UAB38160	1
	EPEDN	UIG18418	1
R246	SEIE	QTQ48337	1
	TLRA	WCL14692, WBE99745, WBG79084	3
S247	LRAG	WBG79084, WCC34642, WBY86968, WBT65812	4
	SKWL	WCP01579, WDG34661, OQ439587, OQ673691, OQ445194, OQ458267, OQ193613	7
	SRWM	ZK92183, WAD73656, WAD78937, WAD79819,…. (see Table S1)	331
	SVGS	WCZ87602, OQ431694	1
	YGHT	ULC27716	1
	YHSD	UXM33786	1
	YRSCCIQ	ULP78342	1
Y248	AGTG	UQW64543	1
	HSDR	UXV06186	1
	KWLD	WGI99683	1
	RWMD	WGL43195	1
L296	HGHTF	UHG56576	1
T299	AVPY	UHJ33999	1
		Receptor binding domain	
R457	HYKYF	QVM41426	1
V483	EVQF	UKF78016	1
V551	EIPTS	UEZ97007, UFA03190, UEF49231	3
		Furin cleavage site (S1/S2 junction)	
S673	YSLS	QVL88657, QVL88693	2
T678	TQRA	WCJ55381	1
N679	GIAL	QSX93802	1
	KAVR	QTC70411	1
		S1/S2 region	
S698	LHHV	UVT41921	1
A701	GTNA	UPA68498	1
E702	CGPKKST	UNH71219	1
	LSSTE	UUH48256	1
	SLSSTA	UTV73082	1

Table 5 (continued)

Position	Insert	Genbank Accession	Number
N703	YSLSS	UWQ86862	1
S704	WCWL	URN60119	1
V705	GNICYT	UXB43176	1
N710	KPCNGVAG	UTH52964	1
N717	SHVV	UJT69381	1
S730	TNVS	UNW19168	1
		Heptad repeat 2	
D1165	WLSR	URF73848	1
D1168	ISGIDLGD	UHY88495	1

the inserted fragment, its position and the identification of the involved sequence. These insertions are PRRA-like because they satisfy the following requirements: (*i*) the S gene insert encoding a given S protein insert has to 100% match to human mRNA transcripts; and (*ii*) the related genes have to be ubiquitous and highly expressed genes or specific genes of target organs or tissues of virus infection.

As an example, the S247 serine S, arginine R, tryptophan W, methionine M (SRWM) insert at the N-terminal domain is discussed (Fig. 4). Unlike PRRA, the acquisition of the SRWM insert was probably not associated with a known gain-of-function. Regarding the SRWM insertion, any sequencing errors could be ruled out. In the study sample, the insert has been identified in 331 SARS-CoV-2 isolates, having 15 different submitters from 15 different organizations. All isolates were from USA, but from 26 different states (collection dates November 2022–March 2030). Table 8S shows detailed information of the SRWM related virus isolates. Like the furin PRRA insertion, the SRWM coding region and its reverse complement 100% to several human mRNA transcripts. Table 6 summarizes the related genes which are also ubiquitous and highly expressed or specific of target organs or tissues of virus infection. Figure 5 shows examples of tissue-specificity of these genes.

Table 9S and Fig. 1S show details of further PRRA-like insertions in the N-terminal domain (T109, insert ILWW; T299, insert AVPY), in the ACE2 receptor binding domain (RBD) (V483, insert EVQF) and in the furin cleavage site itself (N679, insert GIAL).

The spike glycoprotein PRRA-like insertions strongly suggest a recombination between the SARS-CoV-2 genome and human mRNA transcripts within an infected human cell. In this sense, the acquisition of PRRA may not have been a single event of recombination. As well as the S gene insert encoding the PRRA motif, the S gene inserts encoding the S protein PRRA-like inserts do not appear to be of viral origin.

Conclusions

The 12-nucleotide fragment of the SARS-CoV-2 that encoded the first identified spike glycoprotein furin polybasic motif (PRRA) 100% matches to several human mRNA transcripts. This is the basis for the hypothesis of the probable human origin of that fragment that was

```
QHD43416.1                  LLALHRS----YLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   286
QHR63260.2                  LLALHRS----YLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   286
OQ431559.1:21504..25325     LLALHRSSRWMDLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   286
OQ508467.1:21529..25347     LLALHRSSRWMDLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   285
OP998412.1:21529..25347     LLALHRSSRWMDLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   285
OQ508544.1:21504..25322     LLALHRSSRWMDLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   285
OQ408054.1:21529..25347     LLALHRSSRWMDLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   285
OQ817102.1:21523..25341     LLALHRSSRWMDLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT   285
.....

Total: 331 sequences
```

Fig. 4 SARS-CoV-2 spike glycoprotein SRWM insert at the N-terminal domain. Header of a fragment the SARS-CoV-2 spike glycoprotein multiple sequence alignment including the reference sequences [22] and sequences identified with the S247 SRWM insert (coordinate based on reference sequence) at the N-terminal domain. The total number of sequences in the multiple alignment is 331, however, for simplicity the figure only shows a part. Sequence are identified by SARS-CoV-2 GenBank genome accession and S gene coordinates. The insert is highlighted in yellow. At the top, the reference sequences are shaded in grey. Dashes denote gaps. The numbers on the right indicate sequence position

Table 6 Human genes whose mRNA 100% match the SARS-CoV-2 S gene insert encoding the spike glycoprotein N-terminal domain SRWM motif

S gene insert	Ubiquitous gene	Ubiquitous and highly expressed[a] gene	Virus target organ or tissue specific gene	Human SARS-CoV-2 target organ or tissue
Coding TCAAGATGGATG	PANX1, SLC9A8, ESPL1, ARHGAP22, ERCC6	HNRNPM, CYTH3	ODF4 PEX5 UPF3A ZBTB21 FLRT2 GRIA1	Testis Testis Testis Adrenal Ovary Brain
Reverse complement CATCCATCTTGA	LSM8		IQCE ST6GAL2 SPIRE1	Testis Thyroid Brain

[a] hight expressed, RPKM > 10 (transcript expression units are normalized reads per kilobase of transcript, per million mapped reads, RPKM)

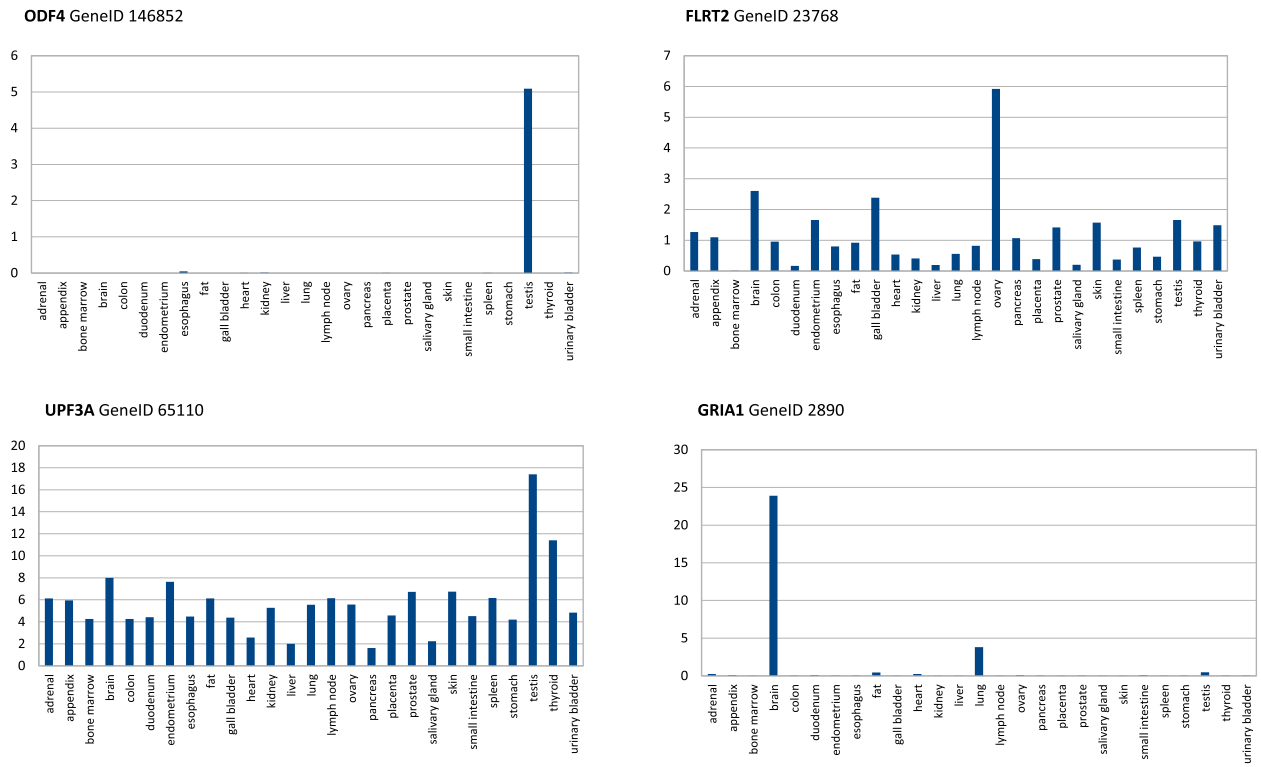

Fig. 5 Tissue specificity of human genes matching SARS-CoV-2 S gene coding SRWM insert at the N-terminal domain. Example of human gene tissue-specificity of genes (referenced in Table 6) whose mRNA transcripts 100% match the SARS-CoV-2 S gene insert encoding the S protein N-terminal domain SRWM insert. Data were download from NCBI Human Genome Resources, assembly GRCh38.p14. Transcript expression units are normalized reads per kilobase of transcript, per million mapped reads (RPKM)

imprinted into the viral S gene. The hypothesis fits with the possibility that a progenitor of SARS-CoV-2 passed to humans, acquiring the PRRA during undetected human-to-human transmission.

The hypothesis is supported by:

- The knowledge that Laos and Vietnam bats host the closest known relatives to SARS-CoV-2 can infect human cells despite the absence of the furin cleavage motif [6, 7].
- The nature of human genes whose transcripts 100% match the furin S gene insert. They are ubiquitous and highly expressed genes or specific of target genes of tissues virus infection.
- Synonymous base substitution at the SARS-CoV-2 furin arginine pair CGG-CGG codon, suggesting

a SARS-CoV-2 evolution to adapting the arginine codon usage.
- PRRA-like insertions at the spike glycoprotein strongly suggest that the fragments inserted in the S gene that encode them do not have a viral origin and were acquired by recombination. The PRRA acquisition may not have been a single event of recombination.

Methods

The source of information was: (*i*) National Center for Biotechnology Information (NCBI) Virus database, SARS-CoV-2 Data Hub [23] and (*ii*) Global Initiative on Sharing Avian Influenza Data (GISAID) database [22, 24]. The reference SARS-CoV-2 S Gene and spike glycoprotein sequences were retrieved from the SARS-CoV-2 reference genomes: (*i*) Wuhan-Hu-1 isolate, GenBank: QHD43416.1 coded by MN908947.3:21563–25,384; and (*ii*) GISAID, EPI_ISL_406801 isolate, genome hCoV-19/Wuhan/WH04/2020: 21551–25,370 [14]. A pipeline of scripts in Perl for data management has been created. The rationale of this work was based on the following tasks:

Task 1. Getting sequences

NCBI SARS-CoV-2 spike glyprotein sequences and coding regions were retrieved from the NCBI Virus database. GISAID SARS-CoV-2 spike glyprotein sequences and coding regions required data parse by executing several chained programs. Briefly:

- Download the complete genomes of the virus isolates
- To retrieve the genome regions covering the S gene (coordinates between 20,000–26,000).
- Using NCBI BLASTn [25], to identify the spike glycoprotein coding region: start and end coordinates. Query: flanking regions of the reference NCBI S gene; subject: the set of genomic regions downloaded from the GISAID database covering the S gene.
- Based on the start- and end-coordinates, to retrieve the S gene region from the downloaded GISAID genomic regions.
- To translate forward three frames of the retrieved S gene regions (coding region).
- To identify the proper translation reading frame (no ambiguous characters, no stop signals). As a result, the spike glycoprotein sequences have been obtained.
- Based on the proper reading frame, to adjust the spike glycoprotein coding regions. As a result, the S gene sequences have been obtained.

Task 2. Similarity analysis between the SARS-CoV-2 S gene insert encoding PRRA and the human mRNA transcripts

The human transcripts database was download from NCBI Human Genome Resources (RefSeq Transcripts, GRch38, download date 05/08/2023). The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins [12]. The downloaded version had 184,489 mRNA sequences grouped in four accession series, denoting the following: "NM_", curated mRNA protein-coding trancripts (66,826 sequences); "NR_", RNA non-protein-coding trascripts (20,584); "XM_", mRNA predicted model protein-coding transcript (69,354); and "XR_", RNA predicted model non-protein-coding transcript (27,725). The sequence analysis similarity was performed using the SARS-CoV-2 S gene 12 nucleotide insert as a query. For each SARS-CoV-2 S gene 12-nucletide insert PRRA coding, and for each human mRNA RefSeq transcript sequence a 12-nucleotide window was run through the entire human mRNA sequence. The 100% match were reported.

Task 3. Arginine codon usage bias or synonymous base substitution in the arginine pair of the SARS-CoV-2 furin site

Because the insertion of the PRRA created a novel RRAR furin cleavage site (with another R after the A in the sequence) that introduces two arginine codons CGG–CGG, using the "RRAR" query sequence that was run as four position window through the entire protein sequence the spike glycoprotein RRAR motif was identified. The protein RRAR motif position was multiplied by three to obtain the corresponding S gene codons. The cases in which the pair of the arginine codons were different from CGG-CGG were recorded.

Task 4. To identify PRRA-like insertions in the SARS-CoV-2 spike glycoprotein sequences

NCBI Virus protein sequences were downloaded with the following filters: host human, ambiguous characters (X) 0, and sequence length 1260–1300. Using a Perl script, the downloaded protein sequences were grouped into blocks of 4000 sequences to be used thorough the EMBL Clustal Omega tool [26], which can align up to 4000 sequences or a maximum file size of 4 MB. The reference SARS-CoV-2 spke glycoprotein sequences were included in each block. Then, using another Perl script, each large multiple sequence alignment was computationally analysed. The spike glycoprotein sequences generating four or more gaps strictly conserved in all other sequences in the block were identified. Then, the

coding region of the identified sequences (having the insert) and the reference S gene sequence were aligned using the HIV Sequence Database Codon Alignment v2.1.0 tool [27, 28]. The nucleotide fragment encoding the protein insert was identified.

Task 5. Tissue-specificity of human genes related to the SARS-CoV-2 PRRA and PRRA-like insertions

The human genes under study were those whose mRNAs had made a 100% match with the S gene inserts. Tissue specificity was retrieved from the Human Protein Atlas RNA-seq normal tissues (HPA RNA-seq normal tissues. NCBI BioProject, Accession: PRJEB4337 ID: 231263). In this BioProject, based on 95 human individuals, the normal human gene expression was determined in 27 different normal human organ or tissues. Data were download from NCBI Human Genome Resources, assembly GRCh38.p14. Units of transcript expression are normalized reads per kilobase of transcript, per million mapped reads (RPKM). The results are shown in bar graphs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-023-01169-8.

Additional file 1: Table 1S. NCBI human mRNA RefSeq transcripts matching the CTCCTCGGCGGG SARS-CoV-2 S gene insert.

Additional file 2: Table 2S. NCBI human mRNA RefSeq transcripts matching the TCCTCGGCGGGC SARS-CoV-2 S gene insert.

Additional file 3: Table 3S. NCBI human mRNA RefSeq transcripts matching the CCCGCCGAGGAG SARS-CoV-2 S gene insert.

Additional file 4: Table 4S. NCBI human mRNA RefSeq transcripts matching the GCCCGCCGAGGA SARS-CoV-2 S gene insert.

Additional file 5: Table 5S. NCBI Virus database SARS-CoV-2 isolates with synonymous base substitution at the arginine codons CGG-CGG.

Additional file 6: Table 6S. GISAID database SARS-CoV-2 isolates with synonymous base substitution at the arginine codons CGG-CGG.

Additional file 7: Table 7S. GISAID database SARS-CoV-2 lineages with synonymous base substitution at the arginine codons CGG-CGG.

Additional file 8: Table 8S. NCBI Virus database SARS-CoV-2 isolates with spike gycoprotein N-terminal domain SRWM insert.

Additional file 9: Table 9S. NCBI human mRNA RefSeq transcripts matching SARS-CoV-2 S gene inserts encoding S protein PRRA-like.

Additional file 10: Figure 1S. Tissue specificity of human genes matching SARS-CoV-2 S gene coding PRRA-like insertion at the NTD, RBD and furin site.

Acknowledgements
The author is very grateful to Dr. Francisco Pascual Velasco for the valuable comments. The author thanks Joe Hayes at British School Tarragona for proofreading this paper. The author acknowledges NCBI Virus and GISAID databases contributors.

Author's contributions
A.R. single author.

Funding
This work has not been awarded grants by any research-supporting institution.

Availability of data and materials
The datasets generated and/or analysed during the current study are all available in the Tables and Figures of both the article itself and the Supplementary Information. All data are the results of sequence analyses. The sequences have been downloaded from NCBI Virus and GISAID databases. In all datasets, the sequence identifier (id) is indicated. The datasets generated in this study are not the type to be uploaded to a life science digital content repository, such as: proteomics data and/or Protein sequences; DNA and/or RNA sequences; genetic polymorphisms; linked genotype and/or phenotype data; Macromolecular structure; gene expression data; or crystallographic data.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 June 2022 Accepted: 26 October 2023
Published online: 21 November 2023

References

1. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26:450–2.
2. Xia S, Lan Q, Su S, Wang X, Xu W, Liu Z, et al. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. Signal Transduct Target Ther. 2020;5:92.
3. Liu S, Wang W, Brown LE, Qiu C, Lajkiewicz N, Zhao T, et al. The PRRA insert at the S1/S2 site modulates cellular tropism of SARS-CoV-2 and ACE2 usage by the closely related bat RaTG13. J Virol. 2021;95:e01751–20.
4. Xia X. Domains and functions of spike protein in Sars-Cov-2 in the context of vaccine design. Viruses. 2021;13:109.
5. Hoffmann M, Kleine-Weber H, Pöhlmann S. Multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Mol Cell. 2020;78:779–84.
6. Temmam S, Vongphayloth K, Baquero E, Munier S, Bonomi M, Regnault B, et al. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature. 2022;604:330–6.
7. Hassanin A, Tu V, Görfö T, Ngon L, Pham P, Hang C, et al. Phylogeographic evolution of horseshoe bat sarbecoviruses in Vietnam and implications for the origins of SARS-CoV and SARS-CoV-2. Preprint Res Square. 2023; https://doi.org/10.21203/rs.3.rs-3227228/v1.
8. Seyran M, Pizzol D, Adadi P, El-Aziz TMA, Hassan SS, Soares A, et al. Questions concerning the proximal origin of SARS-CoV-2. J Med Virol. 2021;93:1204–6.
9. Romeu AR, Ollé E. The SARS-CoV-2 arginine dimers. Preprint Res Square. 2021; https://doi.org/10.21203/rs.3.rs-770380/v1.
10. Kandeel M, Ibrahim A, Fayez M, Al-Nazawi M. From SARS and MERS CoVs to SARS-CoV-2: moving toward more biased codon usage in viral structural and nonstructural genes. J Med Virol. 2020;92:660–6.
11. Romeu AR, Ollé E. SARS-CoV-2 and the secret of the Furin site. Preprints. 2021. 2021020264 https://doi.org/10.20944/preprints202102.0264.v1.
12. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucl Acids Res. 2016;44:D733–45.
13. National Center for Biotechnology Information (NCBI). Human Genome Resources at NCBI. RefSeq Transcripts (GRCh38). Access August 4, 2023.

https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml.
14. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5:1403–7.
15. Tian S, Huang Q, Fang Y, Wu J. FurinDB: A database of 20-residue furin cleavage site motifs, substrates and their associated drugs. Int J Mol Sci. 2011;12:1060–5.
16. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol. 2021;19:155–70.
17. Barbagallo F, Calogero AE, Cannarella Condorelli RA, Mongioì LM, Aversa A, et al. The testis in patients with COVID-19: virus reservoir or immunization resource? Transl Androl Urol. 2020;9:1897–900.
18. Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? Nat Rev Microbiol. 2011;9:617–26.
19. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581:215–20.
20. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367:1260–3.
21. Takeda M. Proteolytic activation of SARS-CoV-2 spike protein. Microbiol Immunol. 2022;66:15–23.
22. Global Initiative on Sharing Avian Influenza Data (GISAID) database. Accessed August 4, 2023. https://www.gisaid.org/
23. National Center for Biotechnology Information (NCBI) Virus database, SARS-CoV-2 Data Hub. Accessed August 4, 2023. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2.
24. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. Euro Surveill. 2017;22:30494.
25. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7:203–14.
26. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucl Acids Res. 2019;47:W636–41.
27. HIV Sequence Database Codon Alignment v2.1.0. Accessed August 4, 2023. https://www.hiv.lanl.gov/content/sequence/CodonAlign/codonalign.html.
28. Korber B. HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences. In: Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers; 2000. Chapter 4, p. 55–72.

Publisher's Note