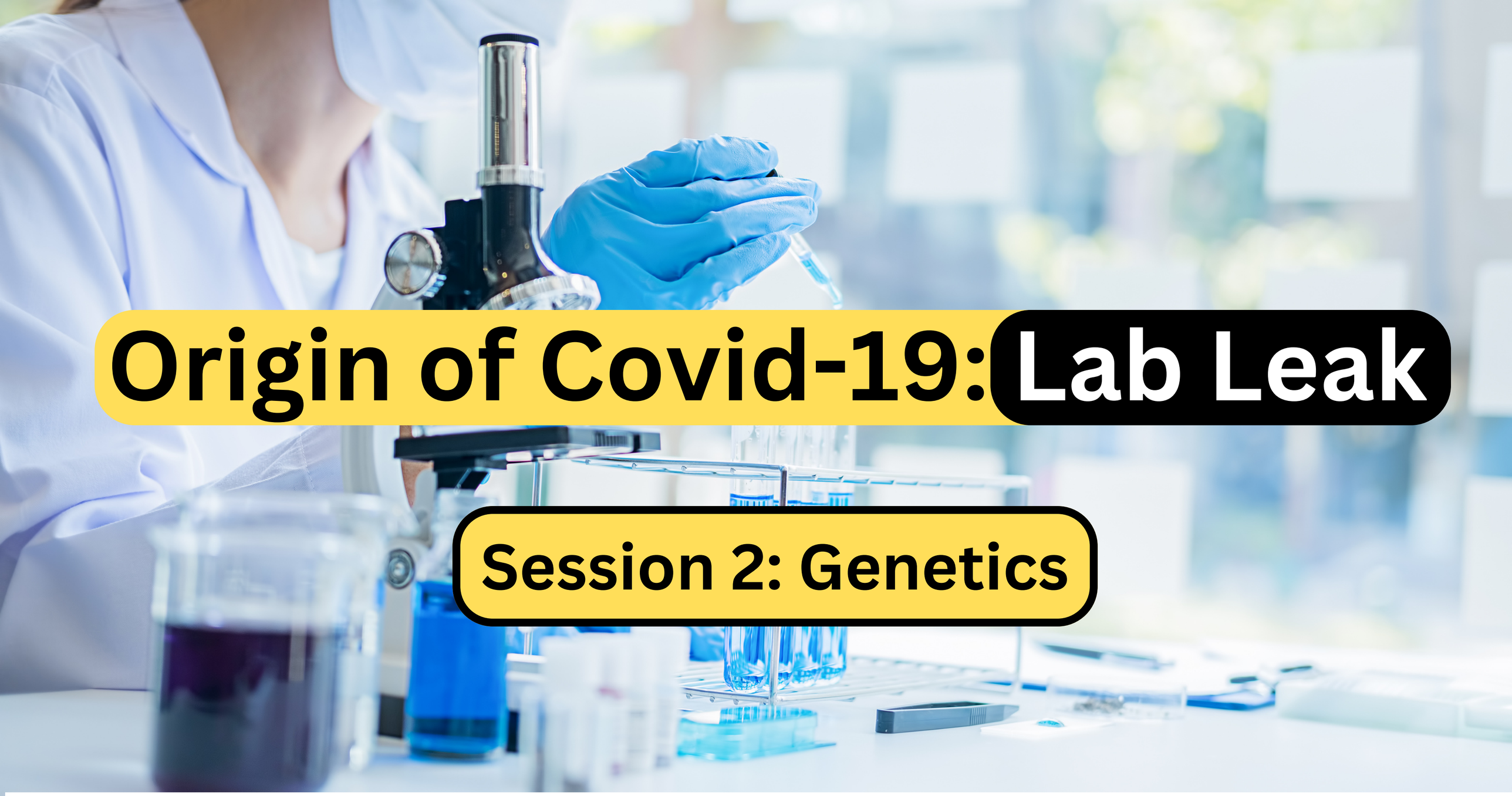


Origin of Covid-19: Lab Leak

Session 2: Genetics

A case by Rootclaim, presented by Yuri Deigin

An Introduction: Yuri Deigin

Drug developer and biotech entrepreneur currently leading a startup developing partial reprogramming gene therapies for Alzheimer’s and other diseases

I won't be delving into the probabilistic inference aspect of the analysis, except to showcase the numbers. That will be discussed by Saar in Session 3.

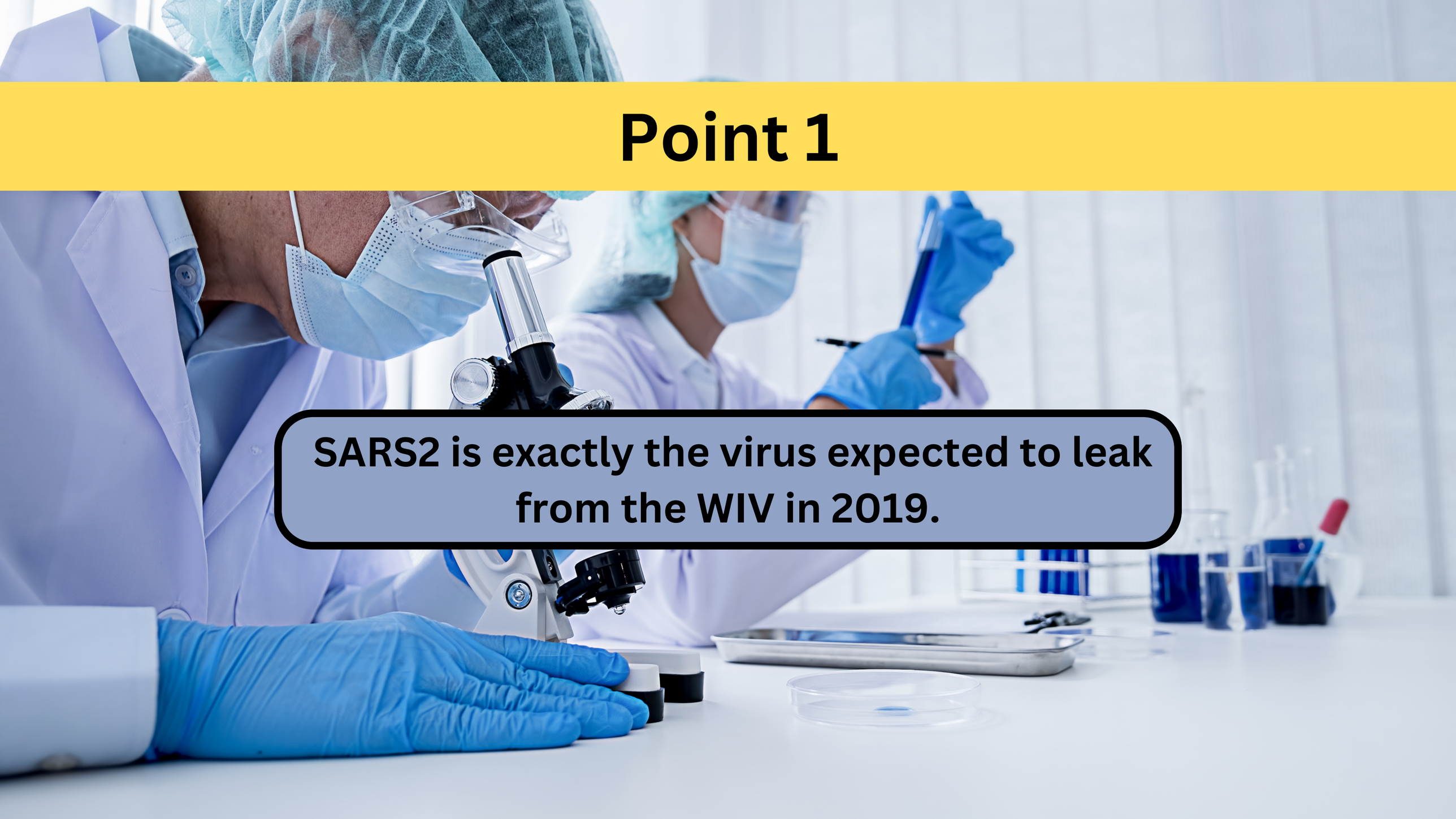
Genetics of SARS-CoV-2: Main Points

SARS2 is exactly the virus expected to leak from the WIV in 2019.

It has several genetic features which are extremely rare in nature, but reasonable to expect from a lab.

Low genetic variability early in the pandemic is indicative of a quick, localized jump of a virus that is already pre-selected for human tropism and possibly further adapted for it in human cells and/or humanized mice, as expected in a lab leak, but not in zoonosis.

Point 1



SARS2 is exactly the virus expected to leak from the WIV in 2019.

What the WIV works on

They collect and research SARS-like and MERS-like coronaviruses

Conduct gain-of-function research on them

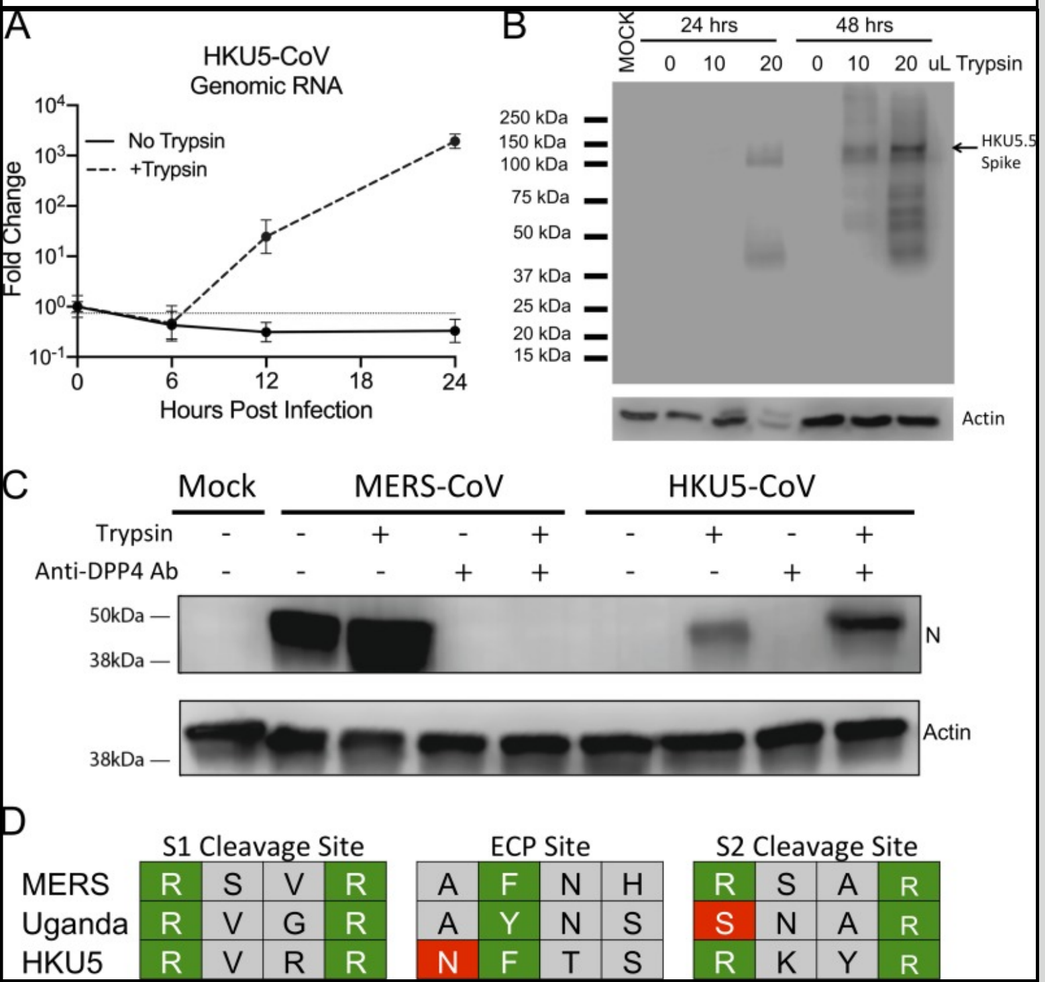
Special interest in Furin Cleavage Sites (FCS)

Furin Cleavage Sites were a focus of coronavirology in 2019

Published online 2020 Feb 14. Prepublished online 2019 Dec 4. doi: 10.1128/JVI.01774-19 PMID: 31801868

Trypsin Treatment Unlocks Barrier for Zoonotic Bat Coronavirus Infection

Vineet D. Menachery,[a,b] Kenneth H. Dinnon, III,[b,c] Boyd L. Yount, Jr.,[b] Eileen T. McAnarney,[a,b] Lisa E. Gralinski,[b] Andrew Hale,[c] Rachel L. Graham,[b] Trevor Scobey,[b] Simon J. Anthony,[d,e] Lingshu Wang,[f] Barney Graham,[f] Scott H. Randell,[g] W. Ian Lipkin,[d,e] and Ralph S. Baric[b,c]



2019 Beijing paper that engineered a novel RRKR furin cleavage site in a chicken coronavirus

Published online 2019 Oct 22. doi: 10.3390/v11100972 PMID: 31652591

The S2 Subunit of QX-type Infectious Bronchitis Coronavirus Spike Protein Is an Essential Determinant of Neurotropism

Jinlong Cheng, Ye Zhao, Gang Xu, Keran Zhang, Wenfeng Jia, Yali Sun, Jing Zhao, Jia Xue, Yanxin Hu, and Guozhong Zhang*

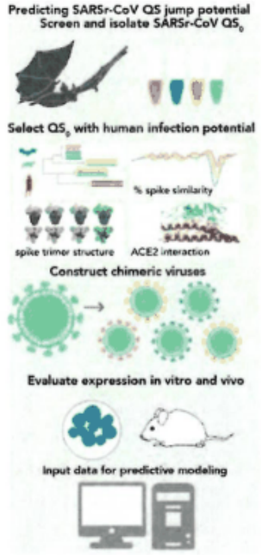
Ralph Baric speaking in China in early 2019 about engineering novel chimeric CoVs:

"Studies to alter pathogen properties of viruses can use several approaches, including selection pressure to drive evolution toward a phenotype as well as deliberate design. Potential opportunities might include building chimeric viruses with altered structures for the receptor for viral entry, or those that incorporate changes to other virulence determinants or that modulate host-pathogen interactions."

The DEFUSE Proposal

SARSr-CoV QS detection, sequencing, and recovery. We will screen samples for SARSr-CoV nucleic acid using our pan-CoV consensus one-step hemi-nested RT-PCR assay targeting a 440-nt fragment in the RNA-dependent RNA polymerase gene (RdRp) of all known α- and β-CoVs[1,53], and specific assays for known SARSr-CoVs[2,21,33,34]. PCR products will be gel purified, sequenced and qPCR performed on SARSr-CoV-positive samples to determine viral load. Full-length genomes or S genes of all SARSr-CoVs will be high-throughput sequenced followed by genome walking[2,3,34]. We will analyze the S gene for its ability to bind human ACE2 by Biocore or virus entry assay. *Synthesis of Chimeric Novel SARSr-CoV QS:* We will commercially synthesize SARSr-CoV S glycoprotein genes, designed for insertion into SHC014 or WIV16 molecular clone backbones (88% and 97% S-protein identity to epidemic SARS-Urbani). These are BSL-3, not select agents or subject to P3C0 (they use bat SARSr-CoV backbones which are exempt) and are pathogenic to hACE2 transgenic mice. Different backbone strains increase recovery of viable viruses identification of barriers for RNA recombination-mediated gene transfer between strains[34]. Recombinant viruses will be recovered in Vero cells, or in mouse cells over-expressing human, bat or civet ACE2 receptors to support cultivation of viruses with a weaker RBD-human ACE2 interface. *Recovery of Full length SARSr-CoV:* We will compile sequence/RNAseq data from a panel of closely related strains (<5% nucleotide variation) and compare full length genomes, scanning for unique SNPs representing sequencing errors[54-56]. Consensus candidates genomes will be synthesized commercially (e.g. BioBasic), using established techniques and genome-length RNA and electroporation to recover recombinant viruses[28,57].

Predicting strain-specific SARSr-CoV spillover risk. We will combine detailed experimental characterization of QS_0 at our test cave sites with state-of-the-art genotype-phenotype Bayesian network models. This will enable us to predict the jump probability of future QS that emerge with unique genetic recombinations. Our models will be parameterized with experimental data from a series of assays on the S genes of bat SARSr-CoVs (**Fig. 6, right**), with experimental and modeling work flowing together in iterative steps. Our prior data will act as baseline to parameterize spillover risk modeling[11,12,29,58]. This will be supplemented by characterization of isolated viruses under DEFUSE (at WIV), **approximately 15-20 bat SARSr-CoV spike proteins/year** (at UNC, WIV), and >180 bat SARSr-CoV strains sequenced in our prior work and not yet examined for spillover potential. All experiments will be performed in triplicate and data fed to models in real time:



Experimental assays of SARSr-CoV QS jump potential (Fig. 6, right). *Pre-screening via structural protein modeling, mutation identification, and pseudovirus assays:* Viral entry is the major species restriction preventing spillover of SARSr-CoVs[29,58]. To select QS for further characterization we will first use structural modeling of SARSr-CoV S protein binding to ACE2 receptors[59,60]. Mutations in the RBD[29,58,61,62], and host protease proteolytic processing of the S glycoprotein[63-65], regulate SARSr-CoV cell entry and cross-species infectivity. Mismatches in the S-RBD-ACE2 molecules or S proteolytic processing will prevent cell entry of SARS-CoV[29,58] and QS with these mismatches will be deprioritized. Single amino acid variations could dramatically alter these phenotypes and we will evaluate the impact of low abundant, high consequence micro-variation in the RBD using RNAseq to identify low abundant QS variants encoding mutations relevant to ACE2 binding. We will conduct *in vitro* pseudovirus binding assays, using established techniques[2], and live virus binding assays (at WIV to prevent delays and unnecessary dissemination of viral cultures) for isolated strains. Initial model predictions based on these data inputs will be used to guide strain selection for further characterization. *In vitro testing of chimeric viruses:* All chimeric viruses will be sequence verified and evaluated for: i) ACE2 receptor usage across species *in vitro*, ii) growth in primary HAE, iii) sensitivity to broadly cross neutralizing human monoclonal antibodies that recognize unique epitopes in the RBD[66,67]. Should some isolates prove highly resistant to our mAB panel, we will evaluate cross neutralization against a limited number of human SARS-CoV serum samples from the Toronto outbreak. Chimeric viruses that encode novel S genes with spillover potential will be used to identify SARSr-CoV strains for recovery as full genome length viable viruses. *In vivo pathogenesis:* Groups of 10 animals will be infected intranasally with 1.0×10^4 PFU of each vSARSr-CoV, clinical signs (weight loss, respiratory function, mortality, etc.) followed for 6 days

p.i., and sacrificed at day 2 or 6 p.i. for virologic analysis, histopathology and immunohistochemistry of the lung and for 22-parameter complete blood count (CBC) and bronchiolar alveolar lavage (BAL). *Validation with full-length genome QS:* We will validate results from chimeric viruses by re-characterizing full-length genome versions, testing whether backbone genome sequence alters full length SARSr-CoV spillover potential. QS for full-genome characterization will be selected to reflect strain differences in antigenicity, receptor usage, growth in human cells and pathogenesis. We will test growth in primary HAE cultures and *in vivo* in hACE2 transgenic mice. We anticipate recovering ~3-5 full length genome viruses/yr. *Testing Synthetic Modifications:* We will synthesize QS with novel combinations of mutations to determine the effects of specific genetic traits and the jump potential of future and unknown recombinants. *RBD deletions:* Small deletions at specific sites in the SARSr-CoV RBD alter risk of human infection. We will analyze the functional consequences of these RBD deletions on SARSr-CoV hACE2 receptor usage, growth in HAE cultures and *in vivo* pathogenesis. First, we will delete these regions, sequentially and in combination, in SHC014 and SARS-CoV Urbani, anticipating that the introduction of deletions will prevent virus growth in Vero cells and HAE[58]. In parallel, we will evaluate whether RBD deletion repair restores the ability of low risk strains to use human ACE2 and grow in human cells. *S2 Proteolytic Cleavage and Glycosylation Sites:* After receptor binding, a variety of cell surface or endosomal proteases[68-71] cleave the SARS-CoV S glycoprotein causing massive changes in S structure [72] and activating fusion-mediated entry[64,73]. We will analyze all SARSr-CoV S gene sequences for appropriately conserved proteolytic cleavage sites in S2 and for the presence of potential furin cleavage sites[74,75]. SARSr-CoV S with mismatches in proteolytic cleavage sites can be activated by exogenous trypsin or cathepsin L. Where clear mismatches occur, we will introduce appropriate human-specific cleavage sites and evaluate growth potential in Vero cells and HAE cultures. In SARS-CoV, we will ablate several of these sites based on pseudotyped particle studies and evaluate the impact of select SARSr-CoV S changes on virus replication and pathogenesis. We will also review deep sequence data for low abundant high risk SARSr-CoV that encode functional proteolytic cleavage sites, and if so, introduce these changes into the appropriate high abundant, low risk parental strain. *N-linked glycosylation:* Some glycosylation events regulate SARS-CoV particle binding DC-SIGN/L-SIGN, alternative receptors for SARS-CoV entry into macrophages or monocytes[76,77]. Mutations that introduced two new N-linked glycosylation sites may have been involved in the emergence of human SARS-CoV from civet and raccoon dogs[77]. While the sites are absent from civet and raccoon dog strains and clade 2 SARSr-CoV, they are present in WIV1, WIV16 and SHC014, supporting a potential role for these sites in host jumping. To evaluate this, we will sequentially introduce clade 2 disrupting residues of SARS-CoV and SHC014 and evaluate virus growth in Vero cells, nonpermissive cells ectopically expressing DC-SIGN, and in human monocytes and macrophages anticipating reduced virus growth efficiency. We will introduce the clade I mutations that result in N-linked glycosylation in rs4237 RBD deletion repaired strains, evaluating virus growth efficiency in HAE, Vero cells, or nonpermissive cells ± ectopic DC-SIGN expression[77]. *In vivo*, we will evaluate pathogenesis in transgenic hACE2 mice. *Low abundance micro-variations:* We will structurally model and identify highly variable residue changes in the SARSr-CoV S RBD, use commercial gene blocks to introduce these changes singly and in combination into the S glycoprotein gene of the low risk, parental strain and test ACE2 receptor usage, growth in HAE and *in vivo* pathogenesis.

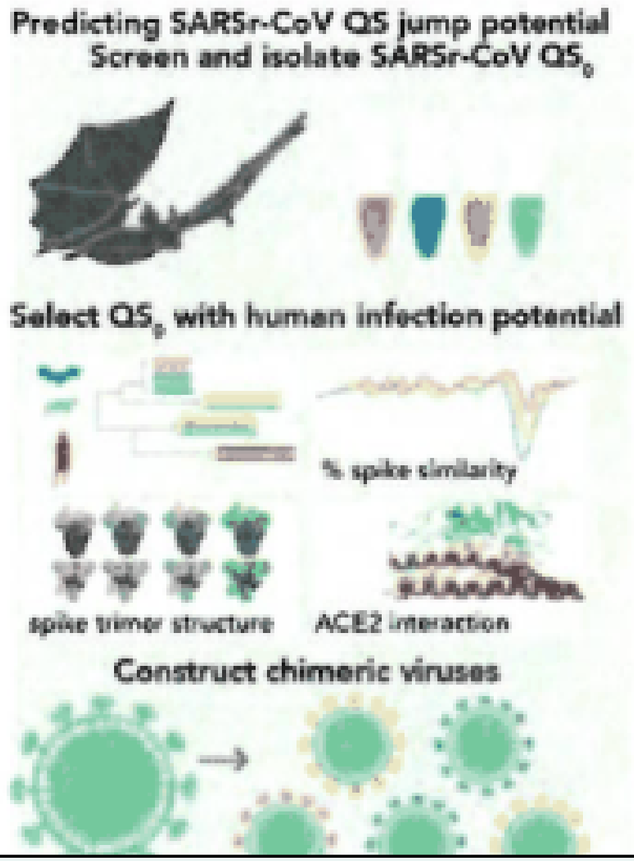
The DEFUSE Proposal

SARSr-CoV QS detection, sequencing, and recovery. We will screen samples for SARSr-CoV nucleic acid using our pan-CoV consensus one-step hemi-nested RT-PCR assay targeting a 440-nt fragment in the RNA-dependent RNA polymerase gene (RdRp) of all known α- and β-CoVs[1,53], and specific assays for known SARSr-CoVs[2,21,33,34]. PCR products will be gel purified, sequenced and qPCR performed on SARSr-CoV-positive samples to determine viral load. Full-length genomes or S genes of all SARSr-CoVs will be high-throughput sequenced followed by genome walking[2,3,34]. We will analyze the S gene for its ability to bind human ACE2 by Biocore or virus entry assay. *Synthesis of Chimeric Novel SARSr-CoV QS:* We will commercially synthesize SARSr-CoV S glycoprotein genes, designed for insertion into SHC014 or WIV16 molecular clone backbones (88% and 97% S-protein identity to epidemic SARS-Urbani). These are BSL-3, not select agents or subject to P3C0 (they use bat SARSr-CoV backbones which are exempt) and are pathogenic to hACE2 transgenic mice. Different backbone strains increase recovery of viable viruses identification of barriers for RNA recombination-mediated gene transfer between strains[34]. Recombinant viruses will be recovered in Vero cells, or in mouse cells over-expressing human, bat or civet ACE2 receptors to support cultivation of viruses with a weaker RBD-human ACE2 interface. *Recovery of Full length SARSr-CoV:* We will compile sequence/RNAseq data from a panel of closely related strains (<5% nucleotide variation) and compare full length genomes, scanning for unique SNPs representing sequencing errors[54-56]. Consensus candidates genomes will be synthesized commercially (e.g. BioBasic), using established techniques and genome-length RNA and electroporation to recover recombinant viruses[28,57].

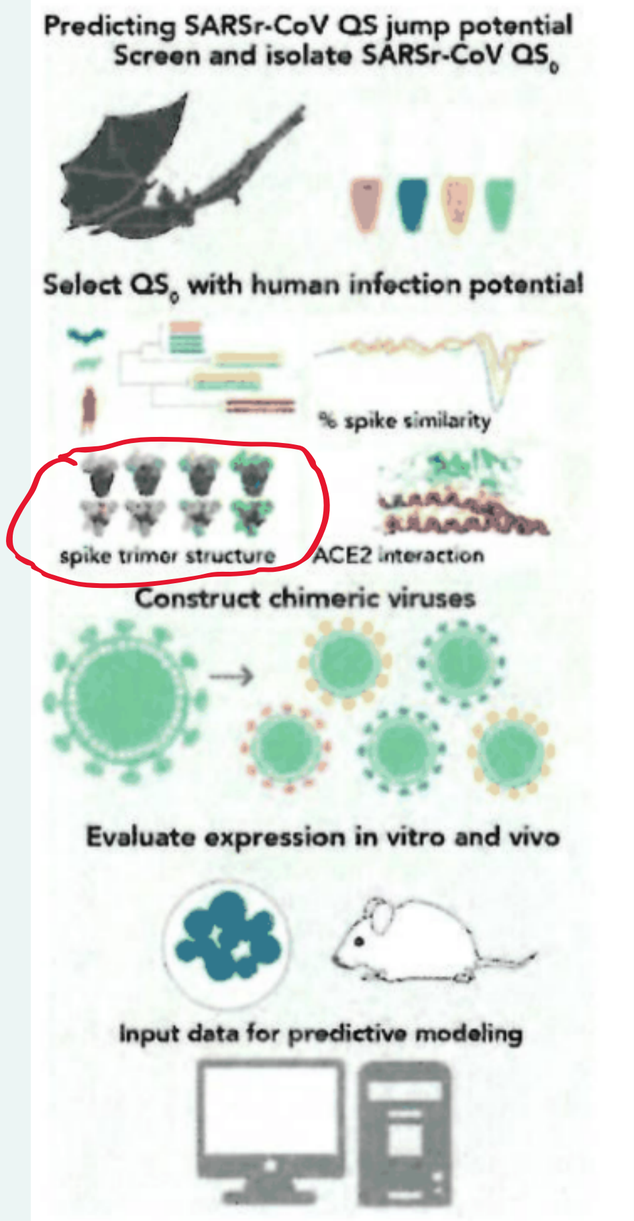
The DEFUSE Proposal

HR001118S0017 EcoHealth Alliance (Daszak) Project DEFUSE

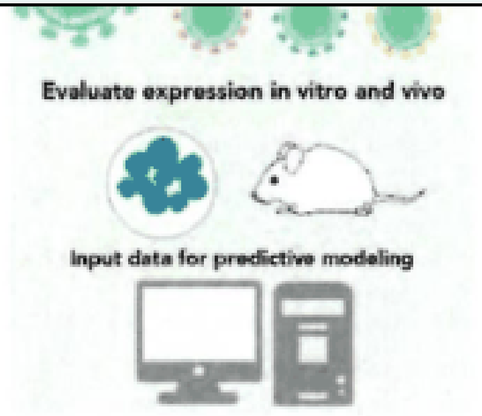
Predicting strain-specific SARSr-CoV spillover risk. We will combine detailed experimental characterization of QS_0 at our test cave sites with state-of-the-art genotype-phenotype Bayesian network models. This will enable us to predict the jump probability of future QS that emerge with unique genetic recombinations. Our models will be parameterized with experimental data from a series of assays on the S genes of bat SARSr-CoVs (**Fig. 6, right**), with experimental and modeling work flowing together in iterative steps. Our prior data will act as baseline to parameterize spillover risk modeling[11,12,29,58]. This will be supplemented by characterization of isolated viruses under DEFUSE (at WIV), **approximately 15-20 bat SARSr-CoV spike proteins/year** (at UNC, WIV), and >180 bat SARSr-CoV strains sequenced in our prior work and not yet examined for spillover potential. All experiments will be performed in triplicate and data fed to models in real time:



The DEFUSE Proposal



The DEFUSE Proposal



Experimental assays of SARSr-CoV QS jump potential (Fig. 6, right). *Pre-screening via structural protein modeling, mutation identification, and pseudovirus assays:* Viral entry is the major species restriction preventing spillover of SARSr-CoVs[29,58]. To select QS for further characterization we will first use structural modeling of SARSr-CoV S protein binding to ACE2 receptors[59,60]. Mutations in the RBD[29,58,61,62], and host protease proteolytic processing of the S glycoprotein[63-65], regulate SARSr-CoV cell entry and cross-species infectivity. Mismatches in the S-RBD-ACE2 molecules or S proteolytic processing will prevent cell entry of SARS-CoV[29,58] and QS with these mismatches will be deprioritized. Single amino acid variations could dramatically alter these phenotypes and we will evaluate the impact of low abundant, high consequence micro-variation in the RBD using RNAseq to identify low abundant QS variants encoding mutations relevant to ACE2 binding. We will conduct *in vitro* pseudovirus binding assays, using established techniques[2], and live virus binding assays (at WIV to prevent delays and unnecessary dissemination of viral cultures) for isolated strains. Initial model predictions based on these data inputs will be used to guide strain selection for further characterization. *In vitro testing of chimeric viruses:* All chimeric viruses will be sequence verified and evaluated for: i) ACE2 receptor usage across species *in vitro*, ii) growth in primary HAE, iii) sensitivity to broadly cross neutralizing human monoclonal antibodies that recognize unique epitopes in the RBD[66,67]. Should some isolates prove highly resistant to our mAB panel, we will evaluate cross neutralization against a limited number of human SARS-CoV serum samples from the Toronto outbreak. Chimeric viruses that encode novel S genes with spillover potential will be used to identify SARSr-CoV strains for recovery as full genome length viable viruses. *In vivo pathogenesis:* Groups of 10 animals will be infected intranasally with 1.0×10^4 PFU of each vSARSr-CoV, clinical signs (weight loss, respiratory function, mortality, etc.) followed for 6 days

The DEFUSE Proposal

Testing Synthetic Modifications: We will synthesize QS with novel combinations of mutations to determine the effects of specific genetic traits and the jump potential of future and unknown recombinants. *RBD deletions:* Small deletions at specific sites in the SARSr-CoV RBD alter risk of human infection. We will analyze the functional consequences of these RBD deletions on SARSr-CoV hACE2 receptor usage, growth in HAE cultures and *in vivo* pathogenesis. First, we will delete these regions, sequentially and in combination, in SHC014 and SARS-CoV Urbani, anticipating that the introduction of deletions will prevent virus growth in Vero cells and HAE[58]. In parallel, we will evaluate whether RBD deletion repair restores the ability of low risk strains to use human ACE2 and grow in human cells. *S2 Proteolytic Cleavage and Glycosylation Sites:* After receptor binding, a variety of cell surface or endosomal proteases[68-71] cleave the SARS-CoV S glycoprotein causing massive changes in S structure [72] and activating fusion-mediated entry[64,73]. We will analyze all SARSr-CoV S gene sequences for appropriately conserved proteolytic cleavage sites in S2 and for the presence of potential furin cleavage sites[74,75]. SARSr-CoV S with mismatches in proteolytic cleavage sites can be activated by exogenous trypsin or cathepsin L. Where clear mismatches occur, we will introduce appropriate human-specific cleavage sites and evaluate growth potential in Vero cells and HAE cultures. In SARS-CoV, we will ablate several of these sites based on pseudotyped particle studies and evaluate the impact of select SARSr-CoV S changes on virus replication and pathogenesis. We will also review deep sequence data for low abundant high risk SARSr-CoV that encode functional proteolytic cleavage sites, and if so, introduce these changes into the appropriate high abundant, low risk parental strain. *N-linked glycosylation:* Some glycosylation events regulate SARS-CoV particle binding DC-SIGN/L-SIGN, alternative receptors for SARS-CoV entry into macrophages or

The DEFUSE Proposal

bronchiolar alveolar lavage (BAL). *Validation with full-length genome QS:* We will validate results from chimeric viruses by re-characterizing full-length genome versions, testing whether backbone genome sequence alters full length SARSr-CoV spillover potential. QS for full-genome characterization will be selected to reflect strain differences in antigenicity, receptor usage, growth in human cells and pathogenesis. We will test growth in primary HAE cultures and *in vivo* in hACE2 transgenic mice. We anticipate recovering ~3-5 full length genome viruses/yr.

DEFUSE vs. SARS2 Comparison

DEFUSE Proposal	SARS2
Screen for / create human ACE2 match.	A spike that is unusually well adapted to human ACE2 from day 1. No need to adapt like in SARS1.
Manipulate N-glycans.	Missing N-glycan that increases infectivity in human lung cells (but bad for enteric).
Introduce human specific cleavage sites if missing.	FCS, first one ever in sarbecoviruses.

Comparing all DEFUSE activities

These are all the activities mentioned in the DEFUSE chapter relevant to GoF. To verify there is no cherry-picking.

DEFUSE activity	SARS2 evidence
SARSr-CoV QS detection, sequencing and recovery	
Synthesis of Chimeric Novel SARSr-CoV QS:	Unknown if SARS2 is chimeric. Probably not, given BANAL
Validation with full-length genome QS	n/a
Predicting strain-specific SARSr-CoV spillover risk	Yes, most likely
Experimental assays of SARSr-CoV QS jump potential	
Pre-screening via structural protein modeling, mutation identification, and pseudovirus assays:	Human ACE2 match
In vitro testing of chimeric viruses:	Chimeric viruses already discussed above
In vivo pathogenesis in hACE2 mice	Unknown. Worth noting that such work increases the probability of a leak
Validation with full-length genome QS:	See above, yes
Testing synthetic modifications	Unknown
RBD deletions	No
S2 proteolytic cleavage and glycosylation sites	Yes
N-linked glycosylation	Yes, although unclear if removal also applies
Low abundance micro-variations	n/a

Highest Affinity to Human ACE2

Table 2 Binding free energies of SARS-Cov-2 spike to ACE2 for different species and infection susceptibility reported by other studies.

From: In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin

Species	ΔG_{eqn1} (kcal/mol)	ΔG_{MMPBSA} (kcal/mol)	SARS-Cov-2 infectivity
Homo sapiens (human)	−52.8	−57.6 ± 0.25	Permissive, high infectivity, severe disease in 5–10%,
Manis javanica (pangolin)	−52.0	−56.3 ± 0.4	Permissive[23,24]
Canis luparis (dog)	−50.8	−49.5	Permissive, low/mod infectivity, no overt disease[25,26]
Macaca fascicularis (monkey)	−50.4	−50.8	Permissive, high infectivity, lung disease[11]
Mesocricetus auratus (hamster)	−49.7	−50.0	Permissive, high infectivity, lung disease[27,28]
Mustela putorius furo (ferret)	−48.6	−49.2	Permissive, moderate infectivity, no overt disease[28,29,30]
Felis catus (cat)	−47.6	−48.9	Permissive, high infectivity, lung disease[26,29,31]
Panthera tigris (tiger)	−47.3	−42.5	Permissive, overt disease, RNA positive[26]
Rhinolophus sinicus (bat)	−46.9	−50.1 ± 1.0	Not permissive[11]
Paguma larvata (civet)	−45.1	−46.1	No reported infection
Equus ferus caballus (horse)	−44.1	−49.2	No naturally occurring infections[26]
Bos taurus (cow)	−43.6	−42.5	No naturally occurring infections[26]
Ophiophagus hannah (king cobra)	−39.5	−40.7 ± 1.2	No reported infection
Mus musculus (mouse)	−38.8	−39.4	Resistant to infection[28]

https://www.nature.com/articles/s41598-021-92388-5/tables/2

Early adaptation will be discussed later

Missing N-Glycan

SARS2 has another unique feature mentioned in DEFUSE not yet seen in any natural SARS-like viruses – an ablated N-linked glycan at position N370. Importantly, the T327A mutation greatly increases SARS2 infectivity in human lung cells but, just like an FCS, this kind of a mutation seems to have selective pressure AGAINST it in ancestral bat viruses.

DEFUSE's interest in N-linked glycans stems from a very curious observation about SARS1 whose bat progenitor seems to have temporarily lost two of its N-linked glycans in civet SARS1 progenitors before re-acquiring them, and this led virologists to hypothesize that those glycans could be relevant for host switching. This is described in DEFUSE in a somewhat convoluted way:

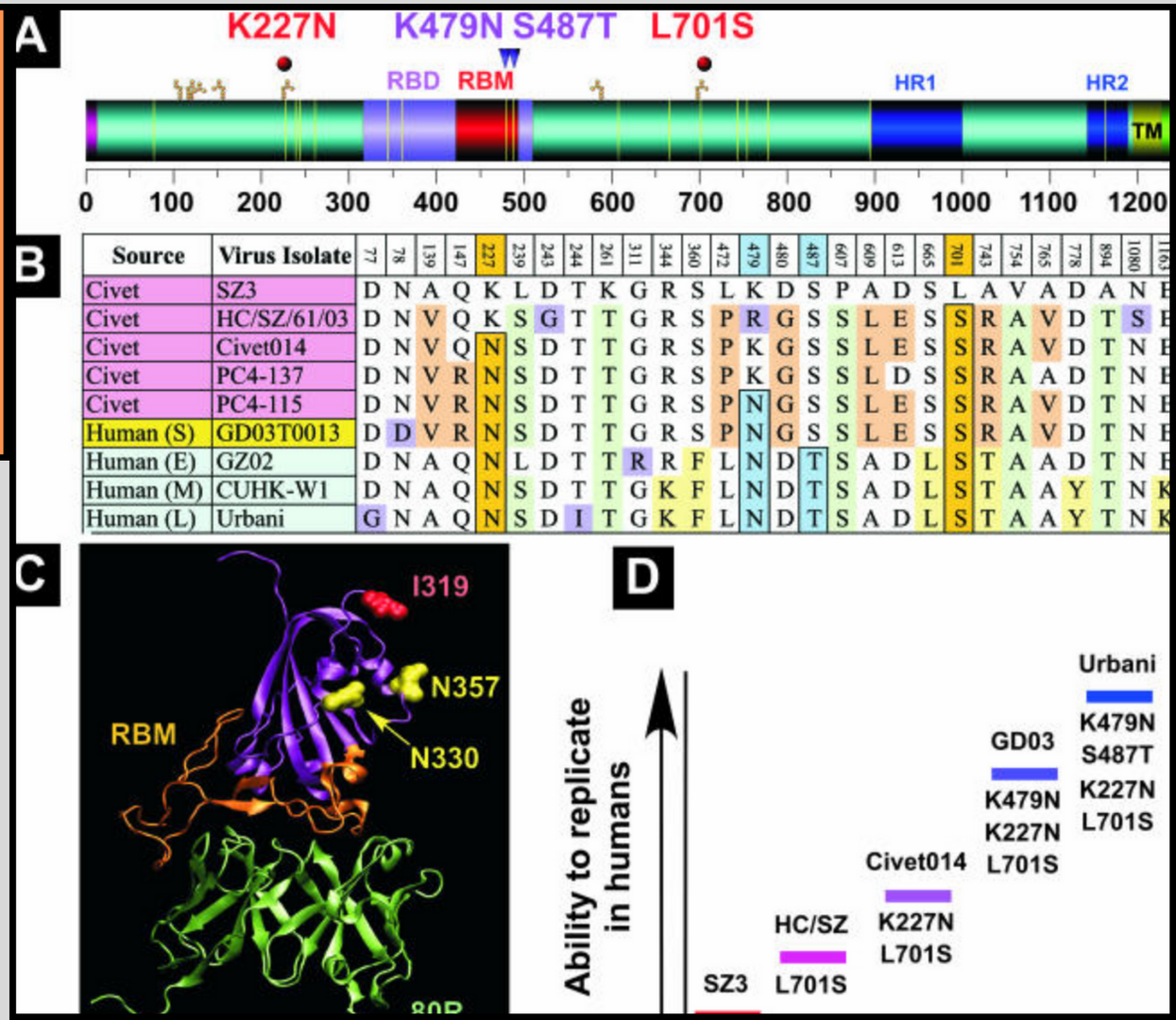
Missing N-Glycan

"N-linked glycosylation: Some glycosylation events regulate SARS-CoV particle binding DC-SIGN/L-SIGN, alternative receptors for SARS-CoV entry into macrophages or monocytes. **Mutations that introduced two new N-linked glycosylation sites may have been involved in the emergence of human SARS-CoV from civet and raccoon dogs. While the sites are absent from civet and raccoon dog strains and clade 2 SARSr-CoV, they are present in WIV1, WIV16 and SHC014, supporting a potential role for these sites in host jumping.** To evaluate this, we will sequentially introduce clade 2 disrupting residues of SARS-CoV and SHC014 and evaluate virus growth in Vero cells, nonpermissive cells ectopically expressing DC-SIGN, and in human monocytes and macrophages anticipating reduced virus growth efficiency. "

Missing N-Glycan

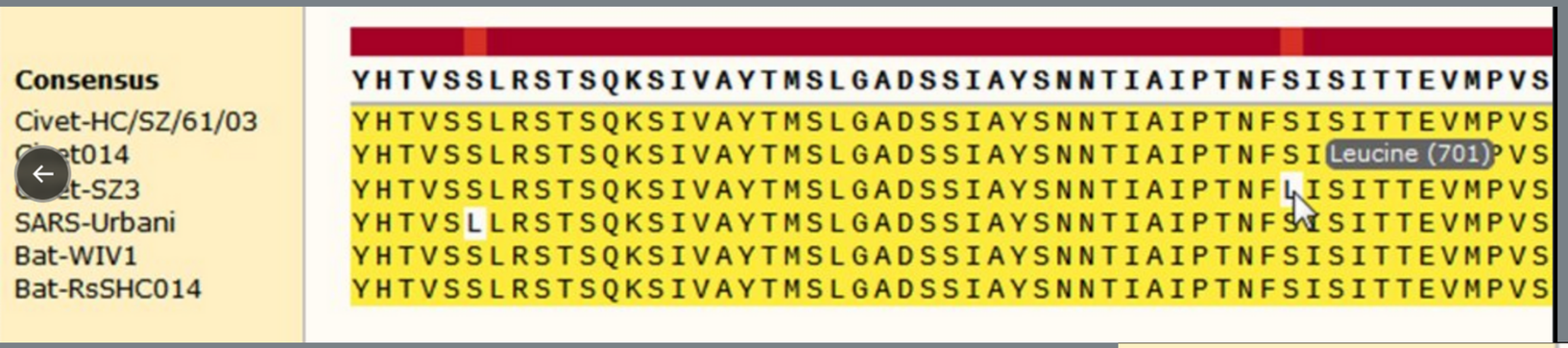
A paper cited in DEFUSE 2007 researched the 5 civet progenitor strains of SARS1 and showed that initially those strains did not have glycans around positions N227 and N699 but then eventually acquired them in civet progenitors and kept in human SARS1.

SOURCE: The paper cited in DEFUSE is a 2007 work by Han et al. titled “Specific Asparagine-Linked Glycosylation Sites Are Critical for DC-SIGN- and L-SIGN-Mediated Severe Acute Respiratory Syndrome Coronavirus Entry”.



Missing N-Glycan

The DEFUSE authors noted that the bat progenitor strains like WIV1/Rs3367 or SHC014 also have glycans at those positions. This is what likely made the DEFUSE authors interested in the host jumping potential of these glycans and potentially genetically modifying them to further study their role:



Missing N-Glycan

Circling back to the DEFUSE proposal, the N370 glycan in SARS2 is the same glycan as N357 in SARS1 which was found to be important for DC-SIGN binding in 2006:

```
Consensus        VVRFPNITNLCPFGEVFNATKFPSVYAWERKRISNCVADYSVLYNSAsparagine (357)G
SARS-GD03T0013   VVRFPNITNLCPFGEVFNATKFPSVYAWERKRISNCVADYSVLYNSTSFSTFKCYG
Civet014         VVRFPNITNLCPFGEVFNATKFPSVYAWERKRISNCVADYSVLYNSTSFSTFKCYG
Civet-HC/SZ/61/03 VVRFPNITNLCPFGEVFNATKFPSVYAWERKRISNCVADYSVLYNSTSFSTFKCYG
Civet-SZ3        VVRFPNITNLCPFGEVFNATKFPSVYAWERKRISNCVADYSVLYNSTSFSTFKCYG
SARS-CUHK-W1     VVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTFKCYG
SARS-Urbani      VVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTFKCYG
SARS-GZ02        VVRFPNITNLCPFGEVFNATKFPSVYAWERKRISNCVADYSVLYNSTFFSTFKCYG
Bat-WIV1         VVRFPNITNLCPFGEVFNATTFPSVYAWERKRISNCVADYSVLYNSTSFSTFKCYG
Bat-RsSHC014     VVRFPNITNLCPFGEVFNATTFPSVYAWERKRISNCVADYSVLYNSTSFSTFKCYG
SARS2            IVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYG
```

Missing N-Glycan

Now, the loss of the N370 glycan by SARS2 has been shown to greatly increase its infectivity in human cells:

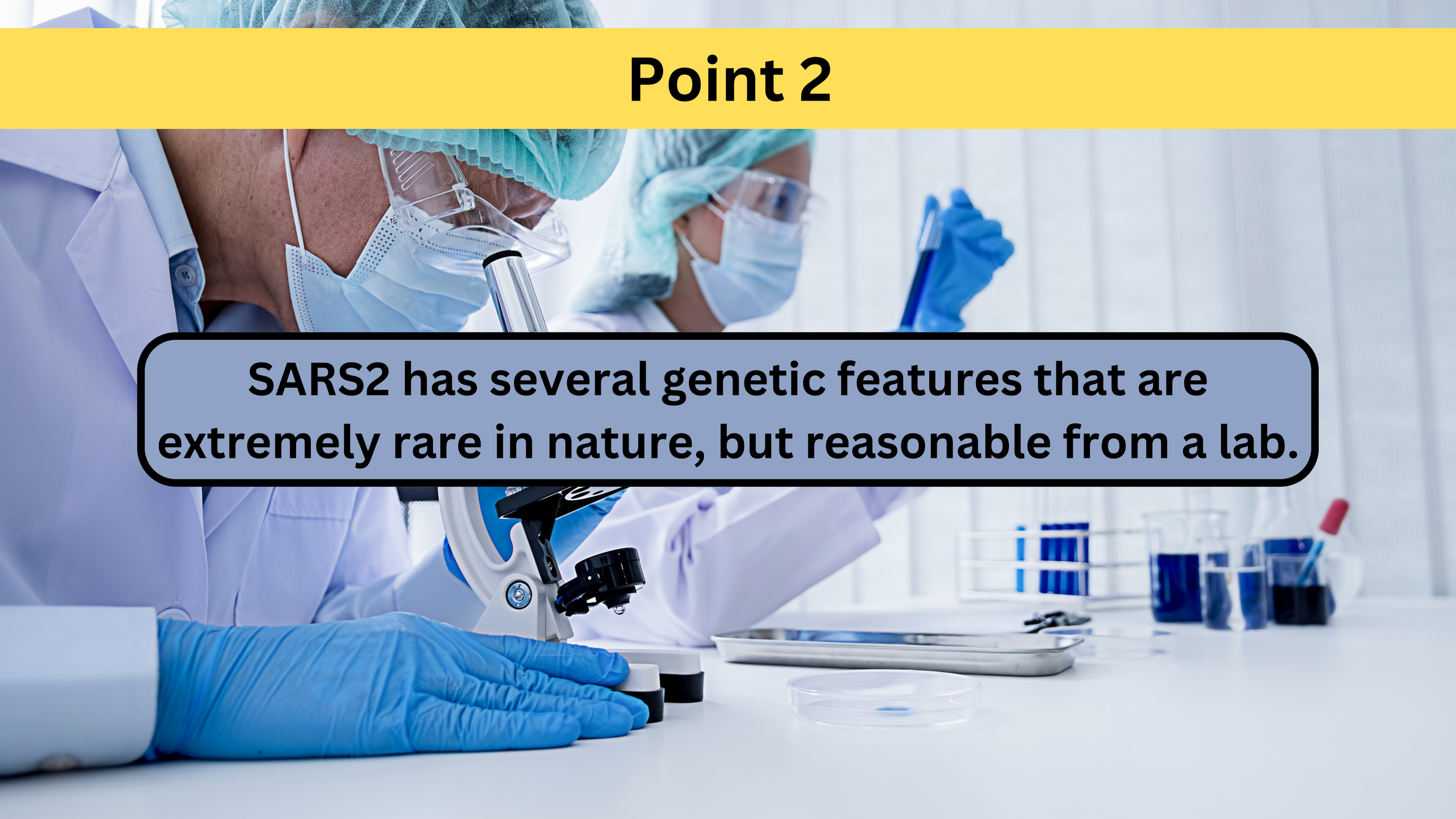
"Using a reverse genetics system to generate a SARS-CoV-2 mutant containing the putative ancestral SNP, we show that the A372T S mutant virus replicates over 60-fold less efficiently than WT SARS-CoV-2 in Calu-3 human lung epithelial cells (Figure 4d). Further, growth of the A372T S mutant was reduced greatly for multiple days, which may be indicative of an effect on viral shedding kinetics in humans. We also generated the D614G S mutant here—reported widely to increase SARS-CoV-2 infectivity (Korber et al., 2020)—which only increased viral titers by a maximum of 2.9-fold in Calu-3 cells compared with the WT, a finding that is consistent with previous results (Plante et al., 2021)."

Missing N-Glycan

However, this mutation is unlikely to have arisen in bats as it is detrimental to oral-fecal transmission (which SARS-like CoVs rely on in bats; this is also likely why we don't see an FCS in bat SARS-like CoVs).

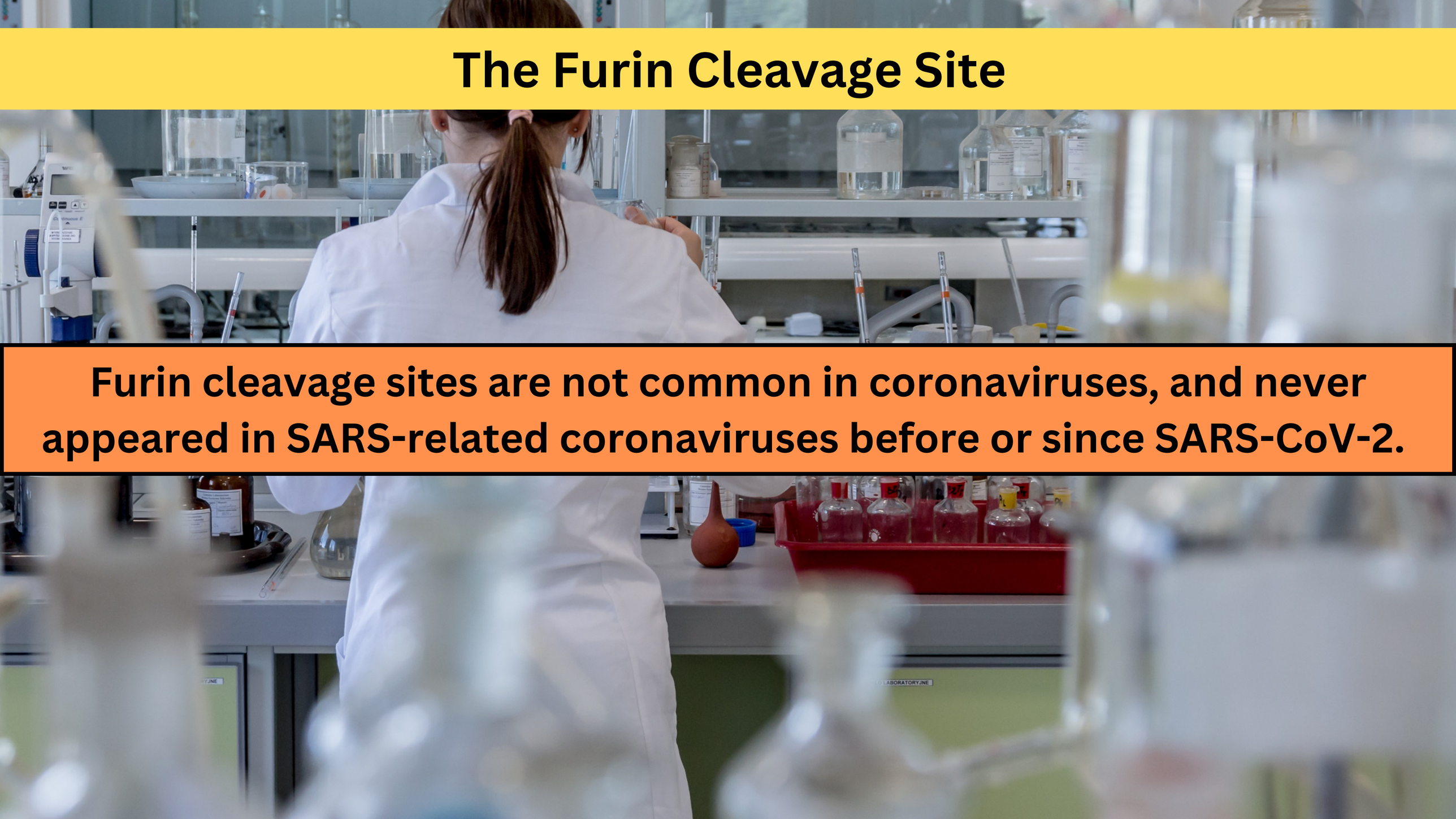
"Why do all bat SC2r-CoVs retain T372, not A372, in their spike proteins, even though the A372 mutant showed substantially higher infectivity than T372? Since the fecal-oral route plays a vital role in bat CoV transmission among bats, we hypothesized that fecal-oral transmission might favor S proteins in all "down" conformation during natural selection, and T372A change might cause some RBDs to assume "up" conformation, which might be detrimental for the survival of S proteins during their passage through the bat stomach. The pH of an insectivorous bat stomach is around 5.633. To test this hypothesis, WT and T372A mutant S pseudovirions were treated with TPCK trypsin at pH 5.5 at 37 °C, a condition roughly mimicking bat stomach digestion. With increase of trypsin concentration, both WT and T372A pseudovirions lost significant amount of infectivity (Fig. 4b, c). However, the speed and extent of infectivity loss varied significantly between WT and T372A mutants (Fig. 4b, c). While a brief 10 min treatment of trypsin at 2.5 μg/mL resulted in over 96.6% and 99.9% loss of infectivity for BANAL-20-52 T372A and BANAL-20-236 T368A mutants, respectively, WT BANAL-20-52 and BANAL-20-236 S pseudovirions retained more than 37% and 21% of infectivity (Fig. 4b, c). Moreover, even after 40 min digestion with trypsin at 2.5 μg/mL, WT BANAL-20-52 and BANAL-20-236 pseudoviruses still retain over 23% and 14% of infectivity, respectively, whereas T372A and T368A mutants almost completely lost infectivity (Fig. 4d, e)."

Point 2



SARS2 has several genetic features that are extremely rare in nature, but reasonable from a lab.

The Furin Cleavage Site



Furin cleavage sites are not common in coronaviruses, and never appeared in SARS-related coronaviruses before or since SARS-CoV-2.

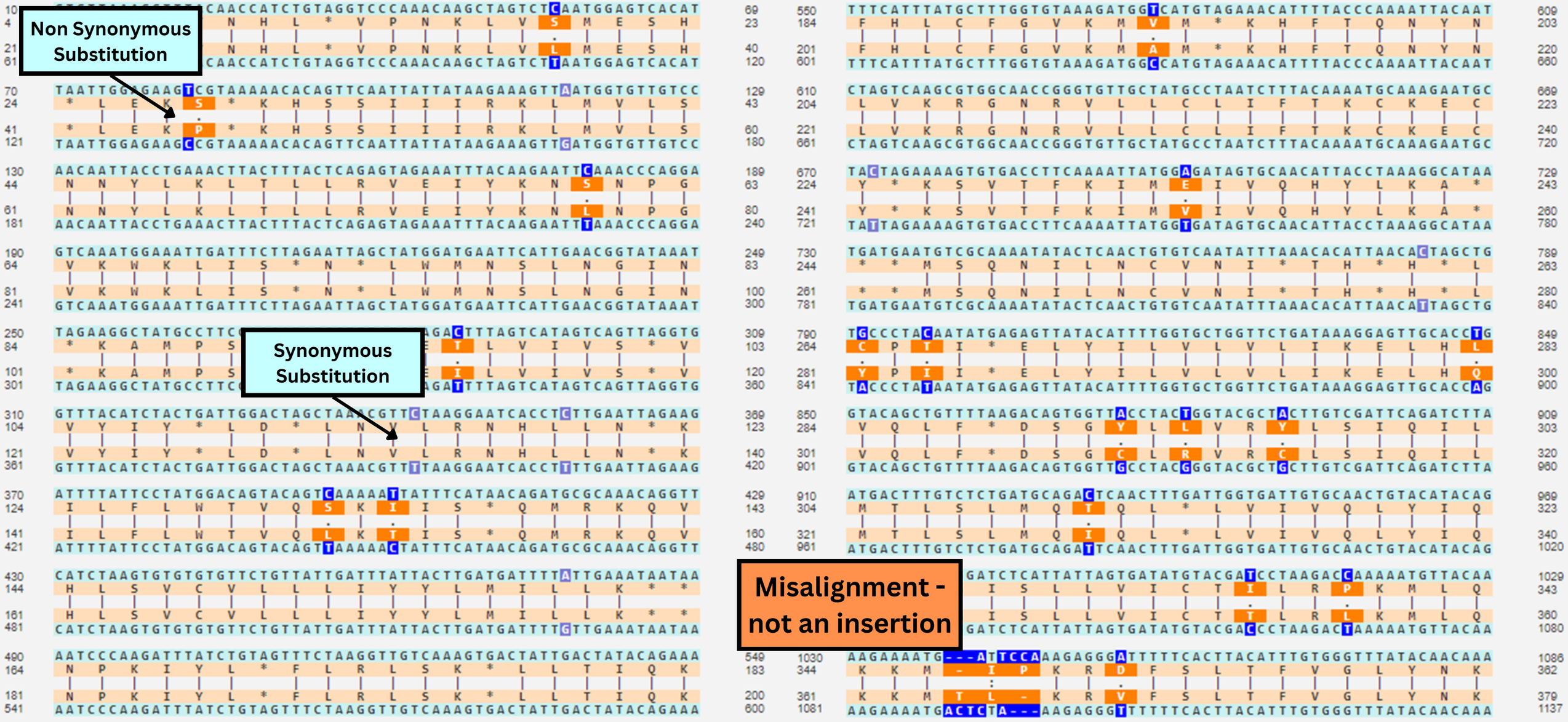
The FCS Alone is Given Low Weight

- Given that we have a bat coronavirus pandemic, it is not unreasonable to expect the new virus to have a novel feature that increases its ability to infect humans.
- An FCS seems unlikely for sarbecoviruses specifically, but hard to estimate by how much.
- So by itself, the FCS is given little probabilistic weight.

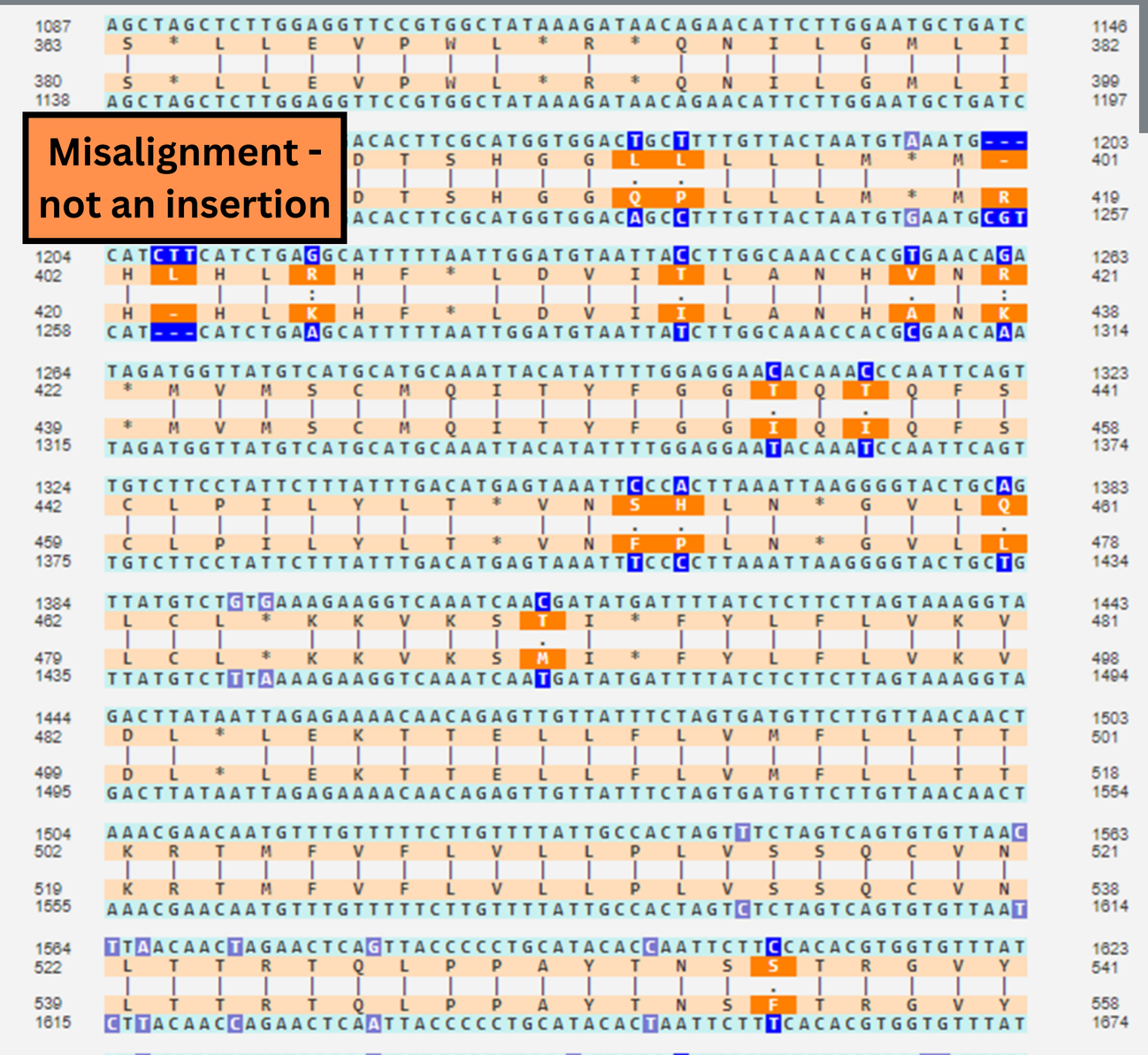
The strong evidence lies in the specific way it appears: as a clean 12nt insertion that uses rare CGG-CGG codons.

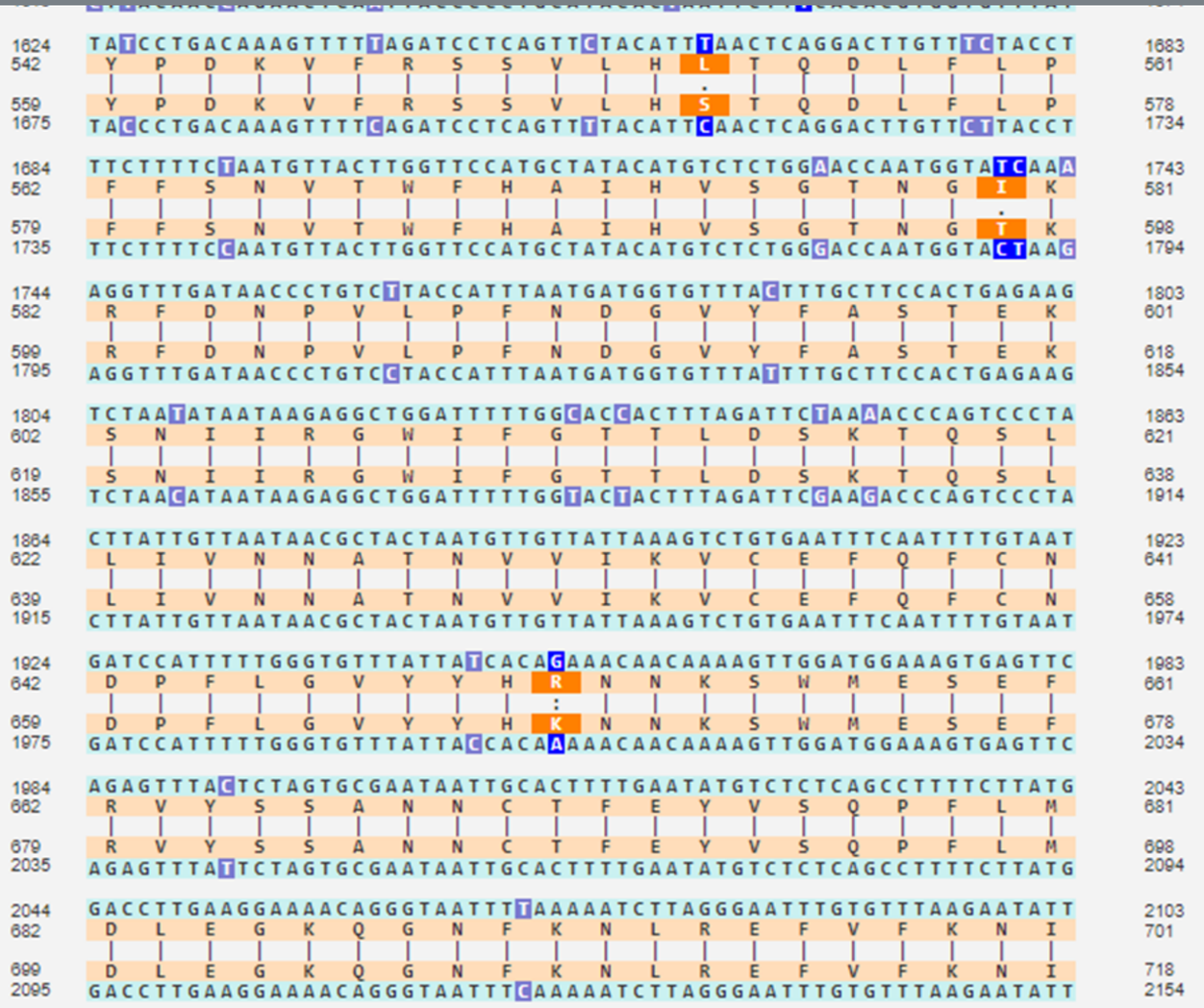
Factor 1 - The FCS is cleanly inserted

The following slides compare the last third of SARS2 to BANAL-52.

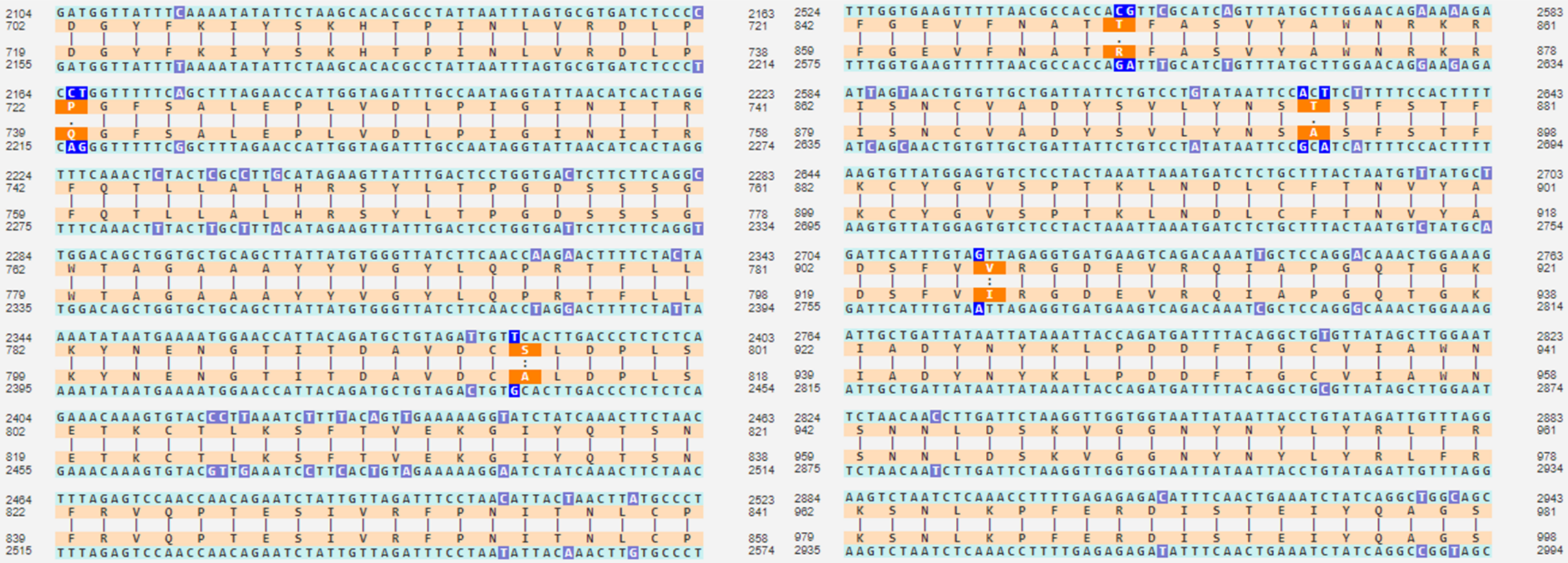


Factor 1 - The FCS is cleanly inserted

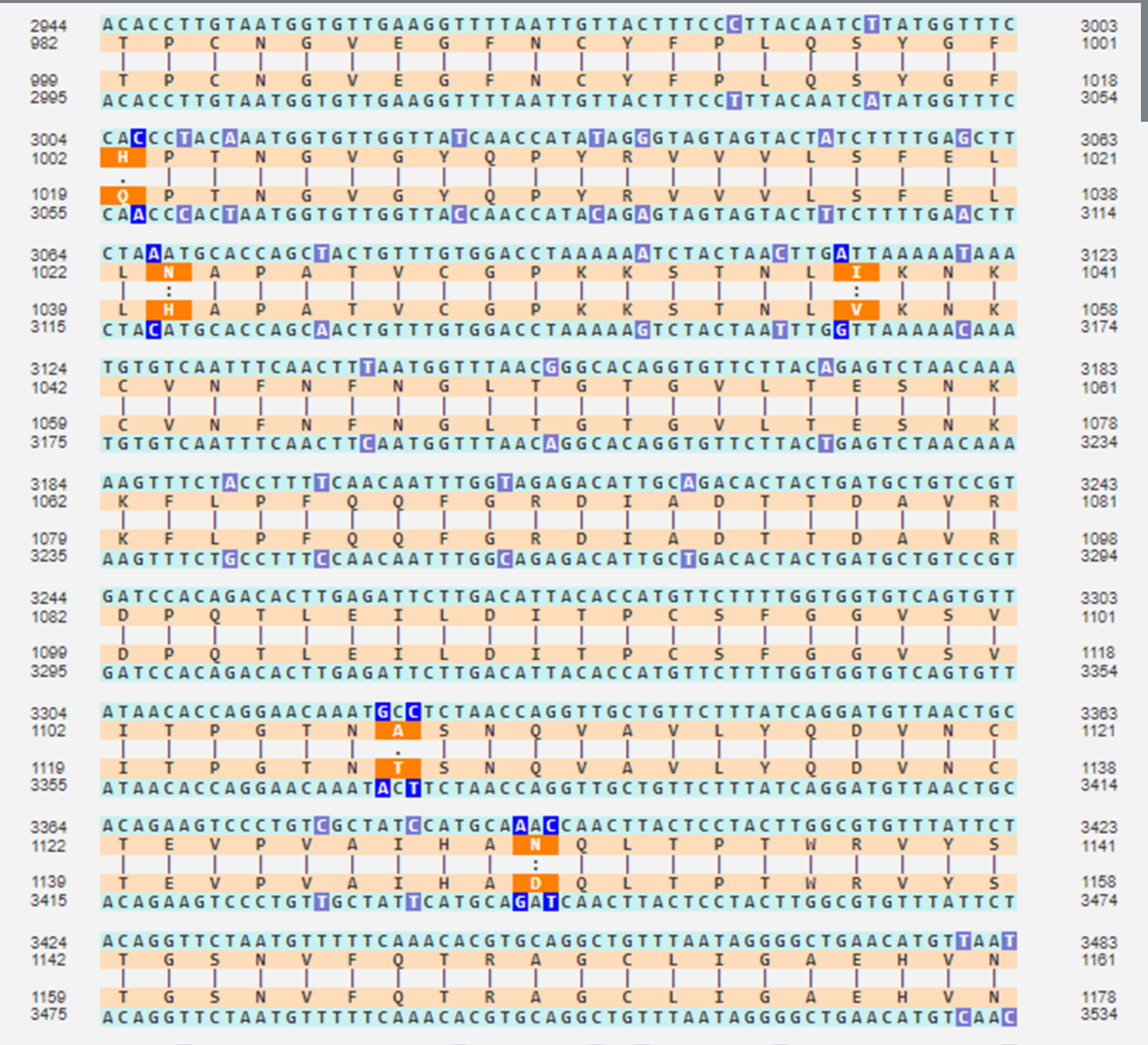


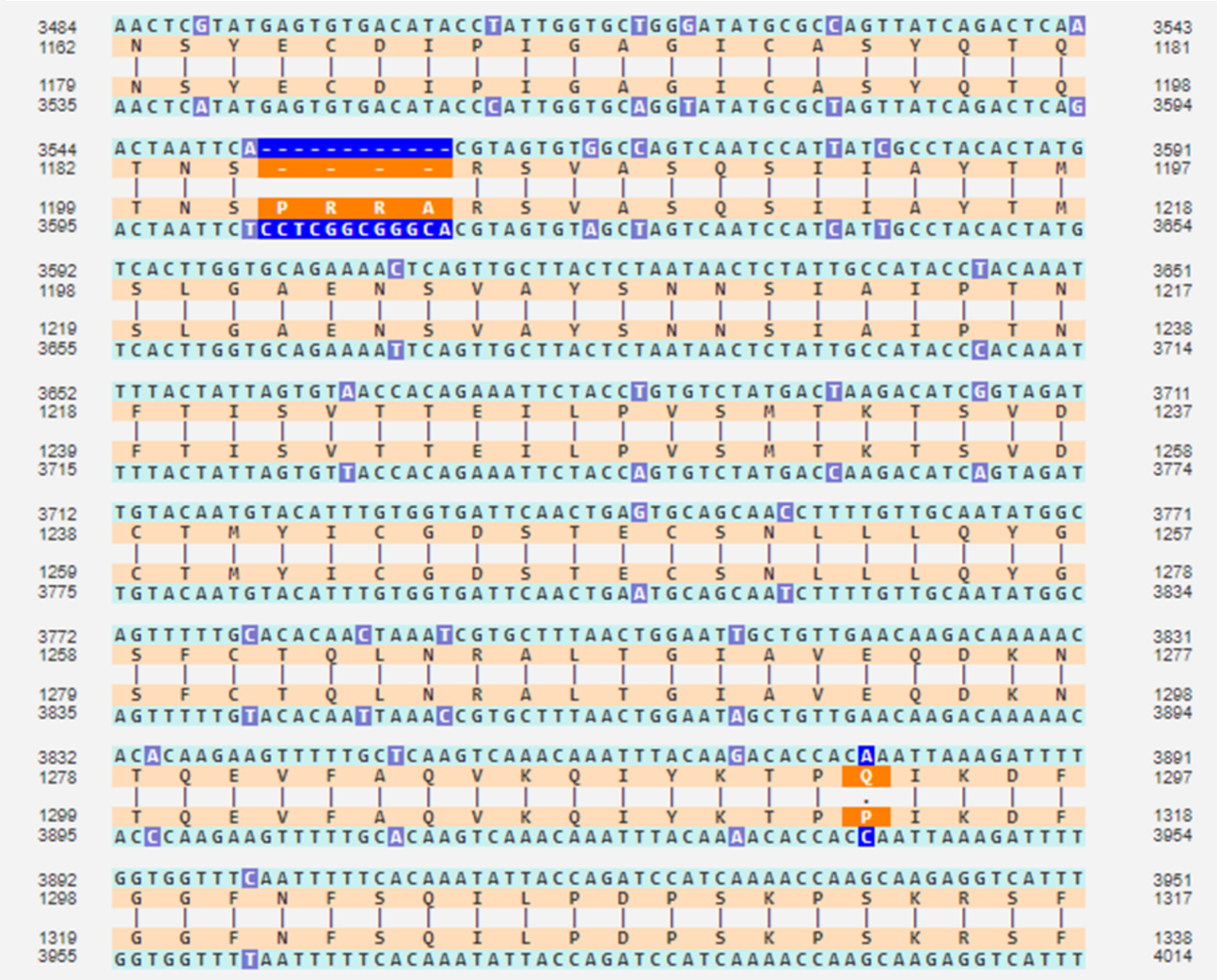


Factor 1 - The FCS is cleanly inserted



Factor 1 - The FCS is cleanly inserted



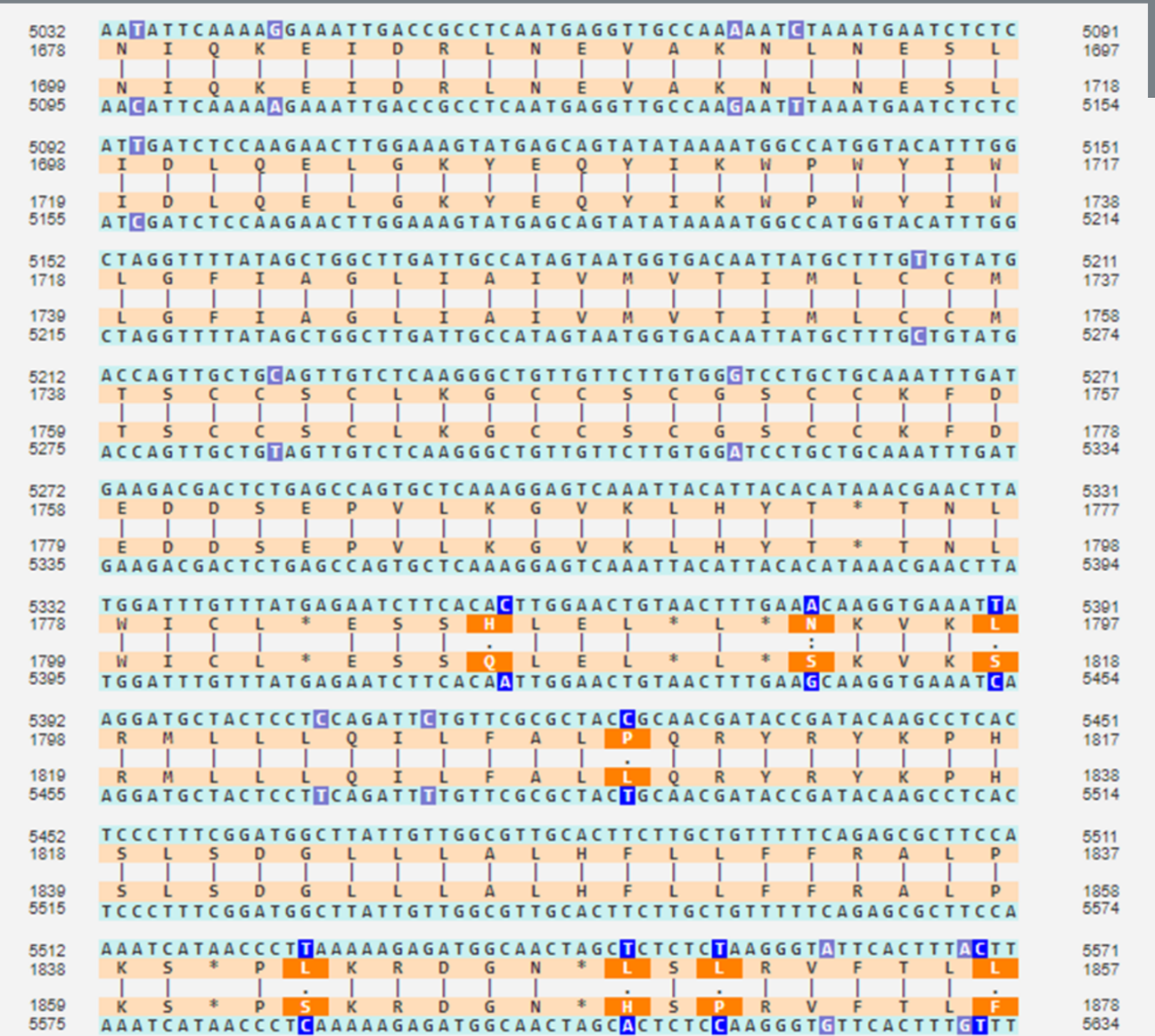


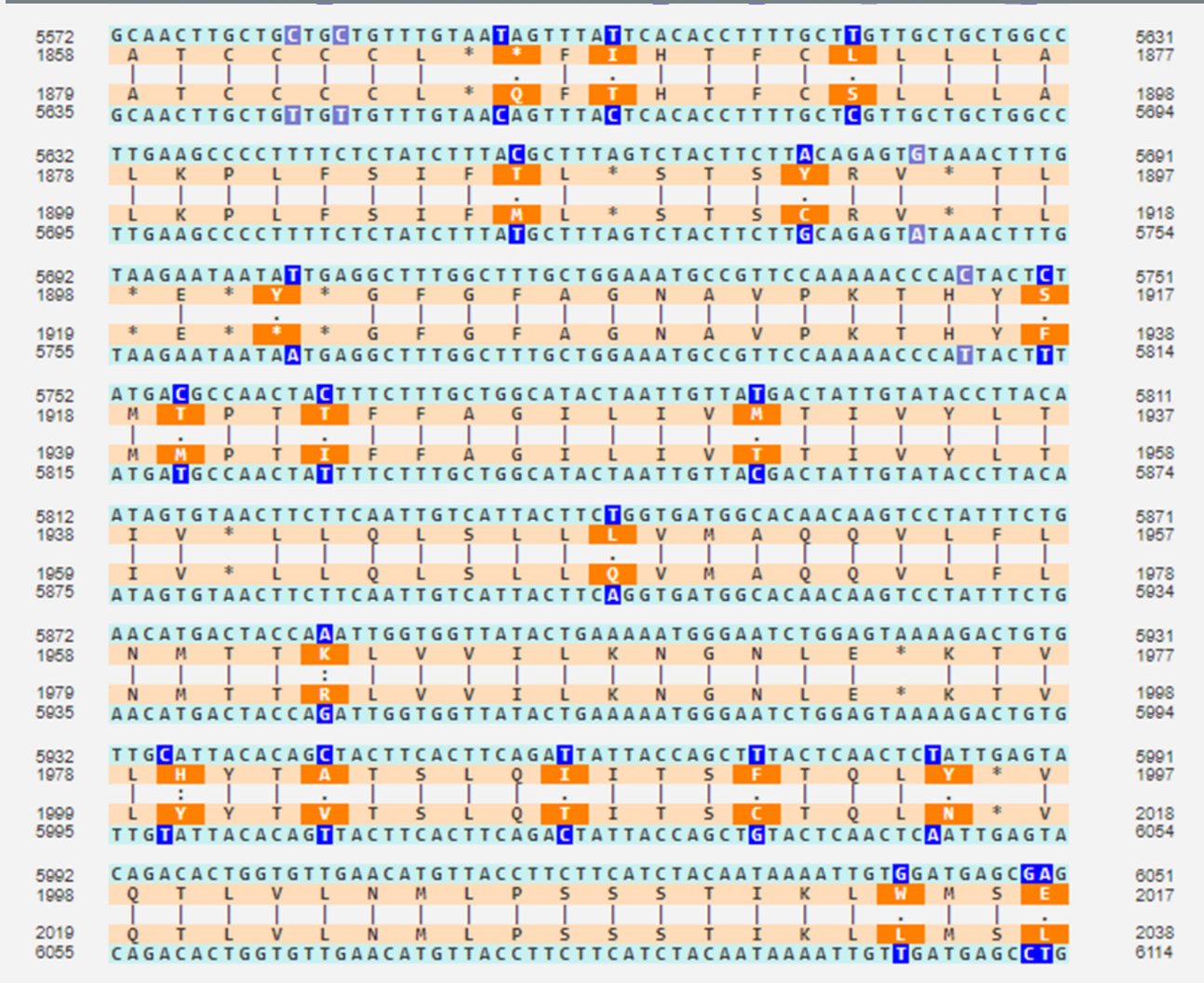
Factor 1 - The FCS is cleanly inserted

```
3952 ATTGAGGACTTGCTCTTCAACAAAGTGACACTTGCTGATGCTGGCTTCATCAAACAATAT
1318  I  E  D  L  L  F  N  K  V  T  L  A  D  A  G  F  I  K  Q  Y
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1339  I  E  D  L  L  F  N  K  V  T  L  A  D  A  G  F  I  K  Q  Y
4015 ATTGAAGATCTACTTTTCAACAAAGTGACACTTGCAGATGCTGGCTTCATCAAACAATAT

4012 GGTGATTGCCTTGGTGATATTGCTGCTAGAGATCTTATTTGTGCTCAAAAGTTTAATGGC
1338  G  D  C  L  G  D  I  A  A  R  D  L  I  C  A  Q  K  F  N  G
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1359  G  D  C  L  G  D  I  A  A  R  D  L  I  C  A  Q  K  F  N  G
4075 GGTGATTGCCTTGGTGATATTGCTGCTAGAGACCTCATTTGTGCACAAAAGTTTAACGGC

4072 CTTACTGTTCTGCCACCTTTGCTCACAGATGAAATGATTGCTCAATACACTTCTGCACTA
1358  L  T  V  L  P  P  L  L  T  D  E  M  I  A  Q  Y  T  S  A  L
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1379  L  T  V  L  P  P  L  L  T  D  E  M  I  A  Q  Y  T  S  A  L
4135 CTTACTGTTTGCCACCTTTGCTCACAGATGAAATGATTGCTCAATACACTTCTGCACTG

4132 TTAGCGGGTACAATCACTTCTGGTTGGACCTTTGGTGCAGGTGCTGCATTACAAATACCA
1378  L  A  G  T  I  T  S  G  W  T  F  G  A  G  A  A  L  Q  I  P
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1399  L  A  G  T  I  T  S  G  W  T  F  G  A  G  A  A  L  Q  I  P
4195 TTAGCGGGTACAATCACTTCTGGTTGGACCTTTGGTGCAGGTGCTGCATTACAAATACCA

4192 TTTGCTATGCAAATGGCTTATAGGTTTAATGGTATTGGAGTTACACAGAATGTTCTCTAT
1398  F  A  M  Q  M  A  Y  R  F  N  G  I  G  V  T  Q  N  V  L  Y
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1419  F  A  M  Q  M  A  Y  R  F  N  G  I  G  V  T  Q  N  V  L  Y
4255 TTTGCTATGCAAATGGCTTATAGGTTTAATGGTATTGGAGTTACACAGAATGTTCTCTAT

4252 GAGAACCAAAAATTGATTGCCAACCAATTTAATAGTGCTATTGGCAAAATCCAAGATTCA
1418  E  N  Q  K  L  I  A  N  Q  F  N  S  A  I  G  K  I  Q  D  S
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1439  E  N  Q  K  L  I  A  N  Q  F  N  S  A  I  G  K  I  Q  D  S
4315 GAGAACCAAAAATTGATTGCCAACCAATTTAATAGTGCTATTGGCAAAATTCAAGACTCA

4312 CTTTCTTCTACAGCAAGTGCACTTGGAAAACTCCAAGATGTTGTCAACCAAAATGCACAA
1438  L  S  S  T  A  S  A  L  G  K  L  Q  D  V  V  N  Q  N  A  Q
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1459  L  S  S  T  A  S  A  L  G  K  L  Q  D  V  V  N  Q  N  A  Q
4375 CTTTCTTCCACAGCAAGTGCACTTGGAAAACTTCAAGATGTGGTCAACCAAAATGCACAA

4372 GCTTTAAACACGCTTGTCAAACAACTTAGCTCCAACTTTGGTGCAATTTCAAGTGTGTTA
1458  A  L  N  T  L  V  K  Q  L  S  S  N  F  G  A  I  S  S  V  L
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1479  A  L  N  T  L  V  K  Q  L  S  S  N  F  G  A  I  S  S  V  L
4435 GCTTTAAACACGCTTGTTAAACAACTTAGCTCCAATTTTGGTGCAATTTCAAGTGTTTTA

4432 AATGACATCCTTTCACGTCTTGACAAAGTTGAGGCTGAAGTGCAGATTGATAGGTTGATC
1478  N  D  I  L  S  R  L  D  K  V  E  A  E  V  Q  I  D  R  L  I
     |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1499  N  D  I  L  S  R  L  D  K  V  E  A  E  V  Q  I  D  R  L  I
4495 AATGATATCCTTTCACGTCTTGACAAAGTTGAGGCTGAAGTGCAAATTGATAGGTTGATC
```

```
4011 4492 ACAGGCAGACTTCAAAGCTTGCAGACATATGTGACTCAACAACTAATTAGAGCTGCAGAA 4551
1337 1498  T  G  R  L  Q  S  L  Q  T  Y  V  T  Q  Q  L  I  R  A  A  E  1517
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1358 1519  T  G  R  L  Q  S  L  Q  T  Y  V  T  Q  Q  L  I  R  A  A  E  1538
4074 4555 ACAGGCAGACTTCAAAGTTTGCAGACATATGTGACTCAACAATTAATTAGAGCTGCAGAA 4614

4071 4552 ATCAGAGCTTCTGCTAATCTTGCTGCTACTAAAATGTCAGAGTGTGTACTCGGACAATCA 4611
1357 1518  I  R  A  S  A  N  L  A  A  T  K  M  S  E  C  V  L  G  Q  S  1537
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1378 1539  I  R  A  S  A  N  L  A  A  T  K  M  S  E  C  V  L  G  Q  S  1558
4134 4615 ATCAGAGCTTCTGCTAATCTTGCTGCTACTAAAATGTCAGAGTGTGTACTTGGACAATCA 4674

4131 4612 AAAAGAGTTGATTTTTGTGGAAAAGGCTATCACCTCATGTCCTTCCCTCAGTCAGCACCT 4671
1377 1538  K  R  V  D  F  C  G  K  G  Y  H  L  M  S  F  P  Q  S  A  P  1557
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1398 1559  K  R  V  D  F  C  G  K  G  Y  H  L  M  S  F  P  Q  S  A  P  1578
4194 4675 AAAAGAGTTGATTTTTGTGGAAAGGGCTATCATCTTATGTCCTTCCCTCAGTCAGCACCT 4734

4191 4672 CATGGTGTAGTTTTCTTGCACGTGACATATGTCCCTGCACAAGAAAAGAACTTCACAACT 4731
1397 1558  H  G  V  V  F  L  H  V  T  Y  V  P  A  Q  E  K  N  F  T  T  1577
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1418 1579  H  G  V  V  F  L  H  V  T  Y  V  P  A  Q  E  K  N  F  T  T  1598
4254 4735 CATGGTGTAGTCTTCTTGCATGTGACTTATGTCCCTGCACAAGAAAAGAACTTCACAACT 4794

4251 4732 GCCCCTGCCATTTGTCATGATGGAAAAGCACACTTTCCTCGCGAGGGTGTTTTTGTTTCA 4791
1417 1578  A  P  A  I  C  H  D  G  K  A  H  F  P  R  E  G  V  F  V  S  1597
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1438 1599  A  P  A  I  C  H  D  G  K  A  H  F  P  R  E  G  V  F  V  S  1618
4314 4795 GCTCCTGCCATTTGTCATGATGGAAAAGCACACTTTCCTCGTGAAGGTGTCTTTGTTTCA 4854

4311 4792 AATGGCACACATTGGTTTGTAACACAAAGGAATTTTTATGAACCACAAATTATTACTACA 4851
1437 1598  N  G  T  H  W  F  V  T  Q  R  N  F  Y  E  P  Q  I  I  T  T  1617
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1458 1619  N  G  T  H  W  F  V  T  Q  R  N  F  Y  E  P  Q  I  I  T  T  1638
4374 4855 AATGGCACACACTGGTTTGTAACACAAAGGAATTTTTATGAACCACAAATCATTACTACA 4914

4371 4852 GATAACACATTTGTATCTGGTAACTGTGATGTTGTAATAGGAATTGTCAACAACACAGTT 4911
1457 1618  D  N  T  F  V  S  G  N  C  D  V  V  I  G  I  V  N  N  T  V  1637
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1478 1639  D  N  T  F  V  S  G  N  C  D  V  V  I  G  I  V  N  N  T  V  1658
4434 4915 GACAACACATTTGTGTCTGGTAACTGTGATGTTGTAATAGGAATTGTCAACAACACAGTT 4974

4431 4912 TATGATCCTTTGCAACCAGAACTTGATTCATTCAAGGAGGAGTTGGATAAATACTTTAAA 4971
1477 1638  Y  D  P  L  Q  P  E  L  D  S  F  K  E  E  L  D  K  Y  F  K  1657
          |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
1498 1659  Y  D  P  L  Q  P  E  L  D  S  F  K  E  E  L  D  K  Y  F  K  1678
4494 4975 TATGATCCTTTGCAACCTGAATTAGACTCATTCAAGGAGGAGTTAGATAAATATTTTAAG 5034

4491 4972 AATCATACATCACCAGATGTAGATTTAAGTGACATCTCTGGCATTAATGCTTCAGTTGTC 5031
1497 1658  N  H  T  S  P  D  V  D  L  S  D  I  S  G  I  N  A  S  V  V  1677
          |  |  |  |  |  |  |  |  |  .  |  |  |  |  |  |  |  |  |  |
1518 1679  N  H  T  S  P  D  V  D  L  G  D  I  S  G  I  N  A  S  V  V  1698
4554 5035 AATCATACATCACCAGATGTTGATTTAGGTGACATCTCTGGCATTAATGCTTCAGTTGTA 5094
```

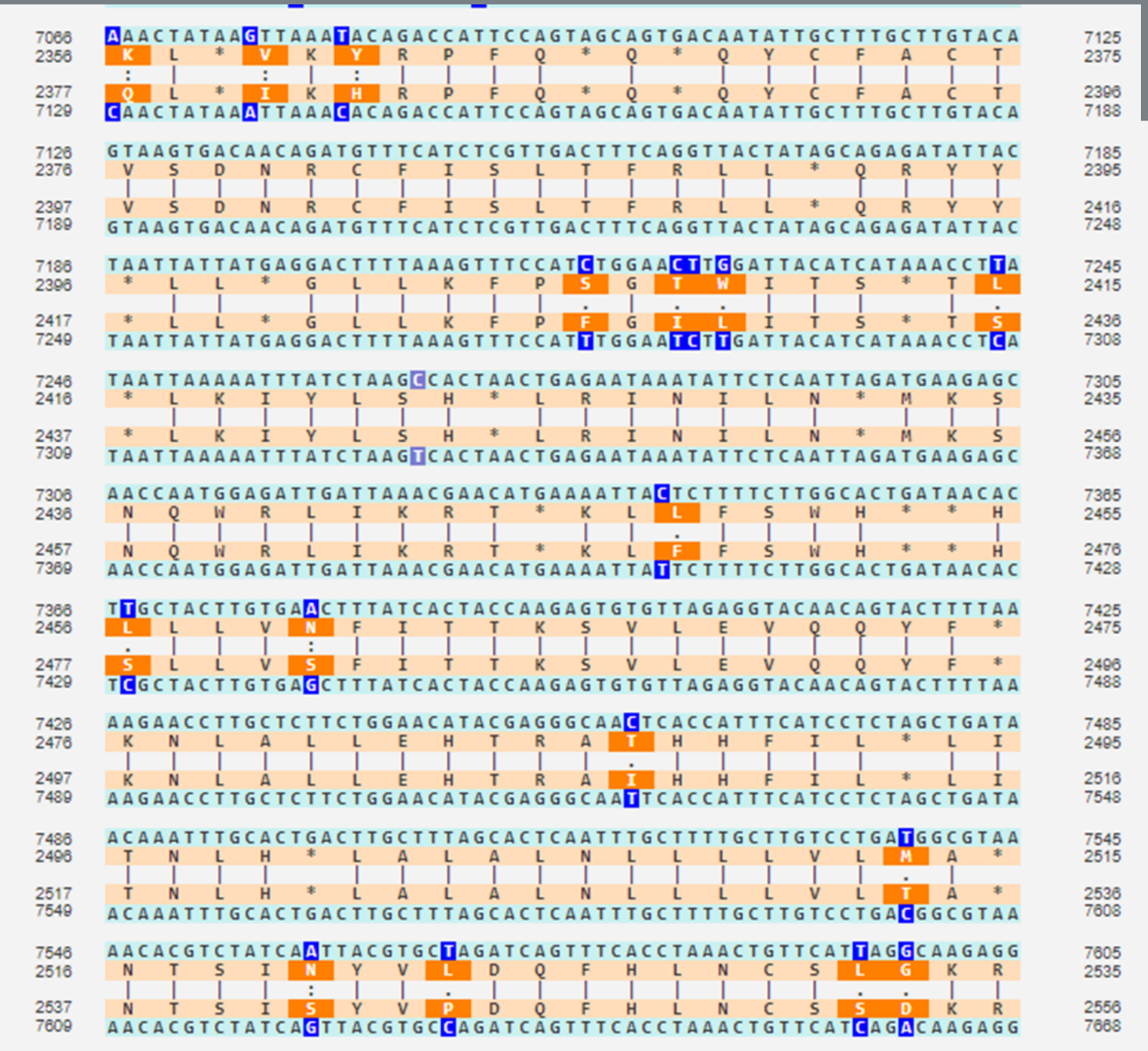
Factor 1 - The FCS is cleanly inserted

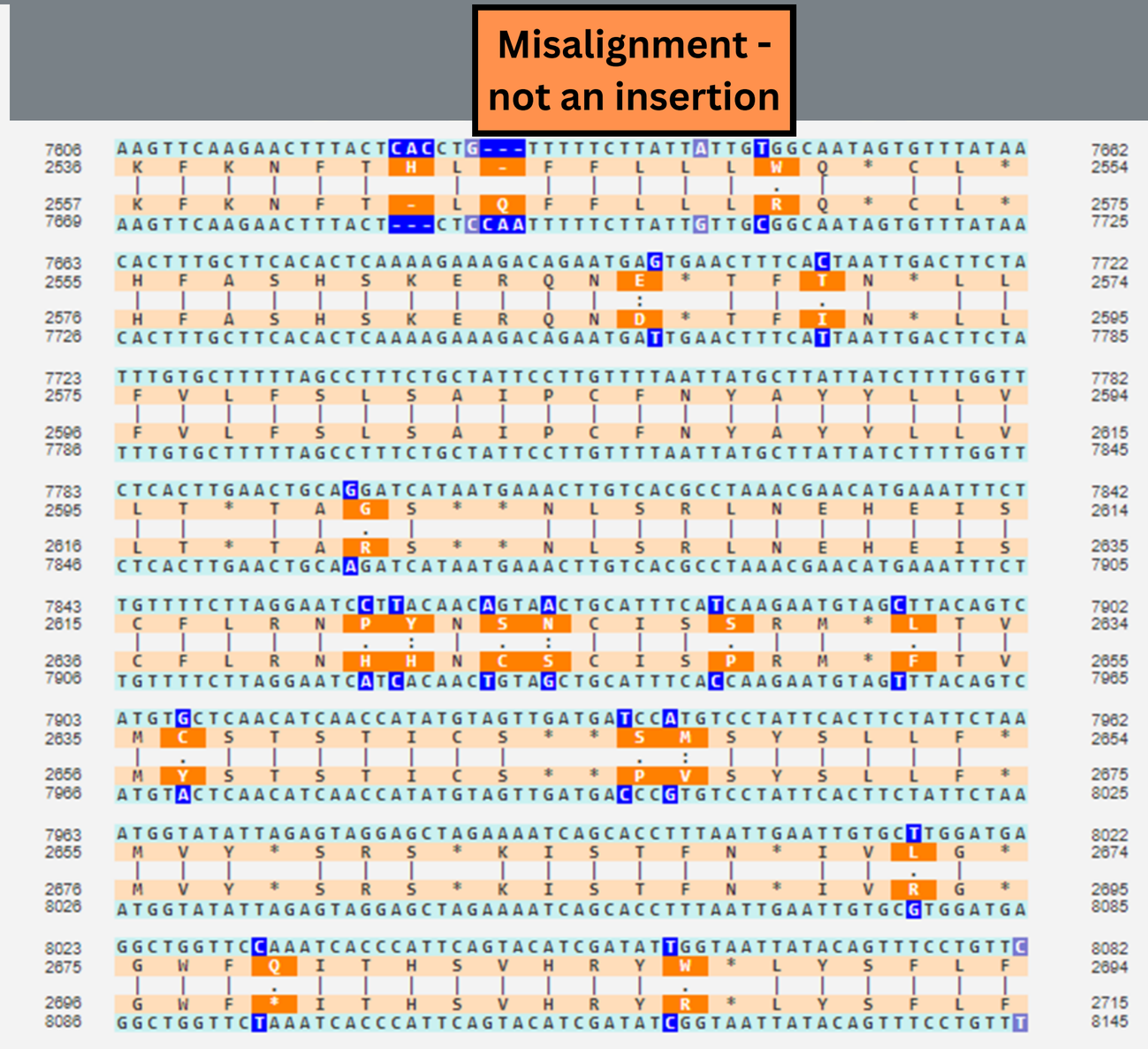




Factor 1 - The FCS is cleanly inserted

Factor 1 - The FCS is cleanly inserted



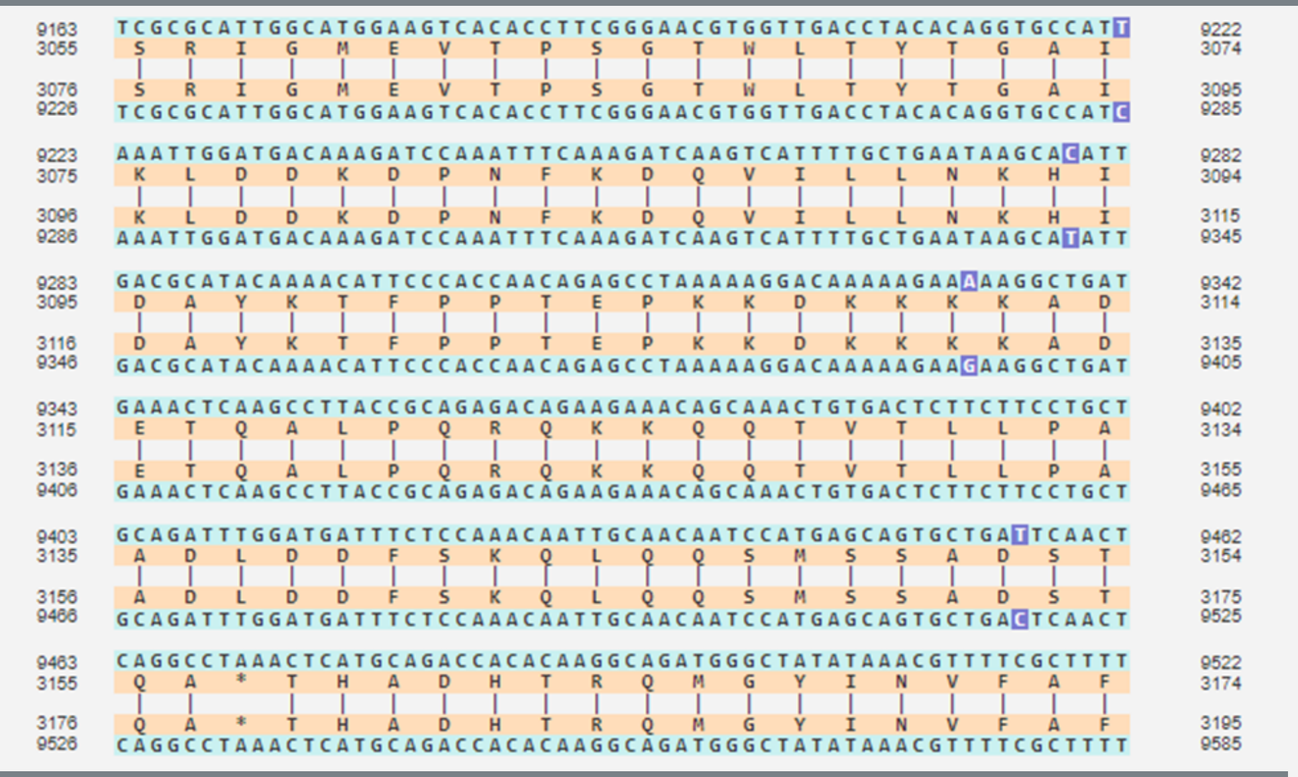


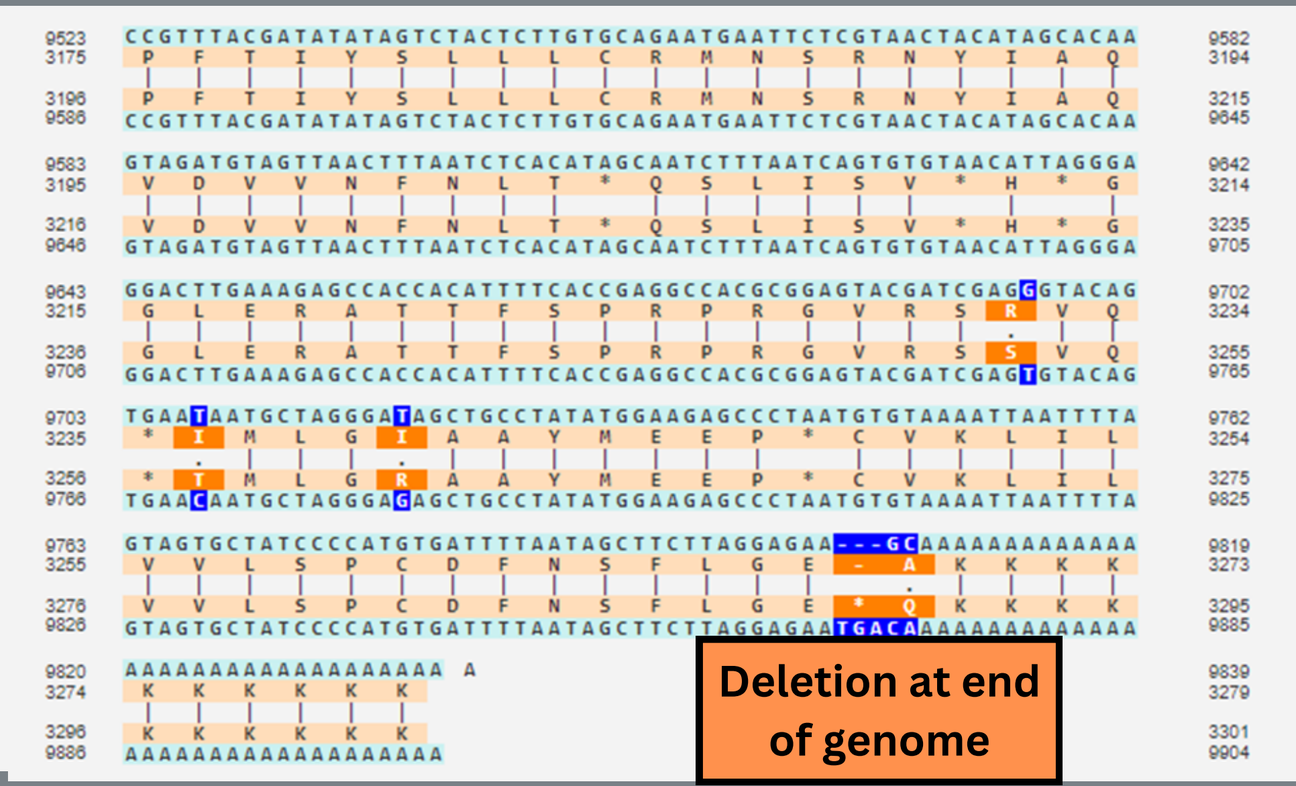
Factor 1 - The FCS is cleanly inserted

```
8083  ACCTTTTACAATTAATTGCCAGGAACCTAAATTGGGTAGTCTTGTAGTGCGTTGTTCGTT  8142
2695   T  F  Y  N  *  L  P  G  T  *  I  G  *  S  C  S  A  L  F  V   2714
       |  |  |  |     |  |  |  |     |  |     |  |  |  |  |  |  |
2716   T  F  Y  N  *  L  P  G  T  *  I  G  *  S  C  S  A  L  F  V   2735
8146  ACCTTTTACAATTAATTGCCAGGAACCTAAATTGGGTAGTCTTGTAGTGCGTTGTTCGTT  8205

8143  CTATGAAGACTTTTTAGAGTATCATGACGTTCGTGTTGTTTTAGATTTCATCTAAACGAA  8202
2715   L  *  R  L  F  R  V  S  *  R  S  C  C  F  R  F  H  L  N  E   2734
       |     |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |  |
2736   L  *  R  L  F  R  V  S  *  R  S  C  C  F  R  F  H  L  N  E   2755
8206  CTATGAAGACTTTTTAGAGTATCATGACGTTCGTGTTGTTTTAGATTTCATCTAAACGAA  8265

8203  CAAACTAAAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTT  8262
2735   Q  T  K  M  S  D  N  G  P  Q  N  Q  R  N  A  P  R  I  T  F   2754
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2756   Q  T  K  M  S  D  N  G  P  Q  N  Q  R  N  A  P  R  I  T  F   2775
8266  CAAACTAAAATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTT  8325

8263  GGTGGACCCTCAGATTCAACTGGCAGTAACCAGAATGGAGAACGCAGTGGGGCGCGACCA  8322
2755   G  G  P  S  D  S  T  G  S  N  Q  N  G  E  R  S  G  A  R  P   2774
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  .
2776   G  G  P  S  D  S  T  G  S  N  Q  N  G  E  R  S  G  A  R  S   2795
8326  GGTGGACCCTCAGATTCAACTGGCAGTAACCAGAATGGAGAACGCAGTGGGGCGCGATCA  8385

8323  AAACAACGTCGGCCTCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCTCTCACT  8382
2775   K  Q  R  R  P  Q  G  L  P  N  N  T  A  S  W  F  T  A  L  T   2794
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2796   K  Q  R  R  P  Q  G  L  P  N  N  T  A  S  W  F  T  A  L  T   2815
8386  AAACAACGTCGGCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCTCTCACT  8445

8383  CAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAAT  8442
2795   Q  H  G  K  E  D  L  K  F  P  R  G  Q  G  V  P  I  N  T  N   2814
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2816   Q  H  G  K  E  D  L  K  F  P  R  G  Q  G  V  P  I  N  T  N   2835
8446  CAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAAT  8505

8443  AGCAGTCCAGACGACCAAATTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGT  8502
2815   S  S  P  D  D  Q  I  G  Y  Y  R  R  A  T  R  R  I  R  G  G   2834
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2836   S  S  P  D  D  Q  I  G  Y  Y  R  R  A  T  R  R  I  R  G  G   2855
8506  AGCAGTCCAGATGACCAAATTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGT  8565

8503  GACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTACCTAGGAACTGGGCCA  8562
2835   D  G  K  M  K  D  L  S  P  R  W  Y  F  Y  Y  L  G  T  G  P   2854
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2856   D  G  K  M  K  D  L  S  P  R  W  Y  F  Y  Y  L  G  T  G  P   2875
8566  GACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTACCTAGGAACTGGGCCA  8625

8563  GAAGCTGGACTTCCCTATGGTGCTAACAAAGATGGCATCATATGGGTTGCAACTGAGGGA  8622
2855   E  A  G  L  P  Y  G  A  N  K  D  G  I  I  W  V  A  T  E  G   2874
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2876   E  A  G  L  P  Y  G  A  N  K  D  G  I  I  W  V  A  T  E  G   2895
8626  GAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGA  8685
```

```
8623  GCCTTGAATACACCAAAAGACCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATT  8682
2875   A  L  N  T  P  K  D  H  I  G  T  R  N  P  A  N  N  A  A  I   2894
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2896   A  L  N  T  P  K  D  H  I  G  T  R  N  P  A  N  N  A  A  I   2915
8686  GCCTTGAATACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATC  8745

8683  GTGCTACAACTTCCTCAAGGAACAACATTGCCAAAAGGCTTCTACGCAGAGGGGAGCAGG  8742
2895   V  L  Q  L  P  Q  G  T  T  L  P  K  G  F  Y  A  E  G  S  R   2914
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2916   V  L  Q  L  P  Q  G  T  T  L  P  K  G  F  Y  A  E  G  S  R   2935
8746  GTGCTACAACTTCCTCAAGGAACAACATTGCCAAAAGGCTTCTACGCAGAAGGGAGCAGA  8805

8743  GGCGGCAGTCAAGCTTCTTCTCGCTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTCA  8802
2915   G  G  S  Q  A  S  S  R  S  S  S  R  S  R  N  S  S  R  N  S   2934
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2936   G  G  S  Q  A  S  S  R  S  S  S  R  S  R  N  S  S  R  N  S   2955
8806  GGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTCA  8865

8803  ACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGGATGGCTGGCAATGGCGGTGATGCT  8862
2935   T  P  G  S  S  R  G  T  S  P  A  R  M  A  G  N  G  G  D  A   2954
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2956   T  P  G  S  S  R  G  T  S  P  A  R  M  A  G  N  G  G  D  A   2975
8866  ACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCT  8925

8863  GCTCTTGCTTTGCTGCTGCTTGACAGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAA  8922
2955   A  L  A  L  L  L  L  D  R  L  N  Q  L  E  S  K  M  S  G  K   2974
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2976   A  L  A  L  L  L  L  D  R  L  N  Q  L  E  S  K  M  S  G  K   2995
8926  GCTCTTGCTTTGCTGCTGCTTGACAGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAA  8985

8923  GGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCAGAGGCTTCTAAGAAA  8982
2975   G  Q  Q  Q  Q  G  Q  T  V  T  K  K  S  A  A  E  A  S  K  K   2994
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
2996   G  Q  Q  Q  Q  G  Q  T  V  T  K  K  S  A  A  E  A  S  K  K   3015
8986  GGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGGCTTCTAAGAAG  9045

8983  CCTCGGCAAAAACGTACTGCCACTAAACAATACAATGTAATACAAGCTTTTGGCAGACGT  9042
2995   P  R  Q  K  R  T  A  T  K  Q  Y  N  V  I  Q  A  F  G  R  R   3014
       |  |  |  |  |  |  |  |  |  .  |  |  |  .  |  |  |  |  |  |
3016   P  R  Q  K  R  T  A  T  K  A  Y  N  V  T  Q  A  F  G  R  R   3035
9046  CCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGT  9105

9043  GGTCCAGAACAAACCCAAGGAAACTTTGGGGACCAGGAACTAATCAGACAAGGAACTGAT  9102
3015   G  P  E  Q  T  Q  G  N  F  G  D  Q  E  L  I  R  Q  G  T  D   3034
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
3036   G  P  E  Q  T  Q  G  N  F  G  D  Q  E  L  I  R  Q  G  T  D   3055
9106  GGTCCAGAACAAACCCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGAT  9165

9103  TACAAACATTGGCCGCAAATTGCACAATTTGCTCCCAGCGCTTCTGCATTCTTTGGAATG  9162
3035   Y  K  H  W  P  Q  I  A  Q  F  A  P  S  A  S  A  F  F  G  M   3054
       |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
3056   Y  K  H  W  P  Q  I  A  Q  F  A  P  S  A  S  A  F  F  G  M   3075
9166  TACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGCGCTTCAGCGTTCTTCGGAATG  9225
```

Factor 1 - The FCS is cleanly inserted



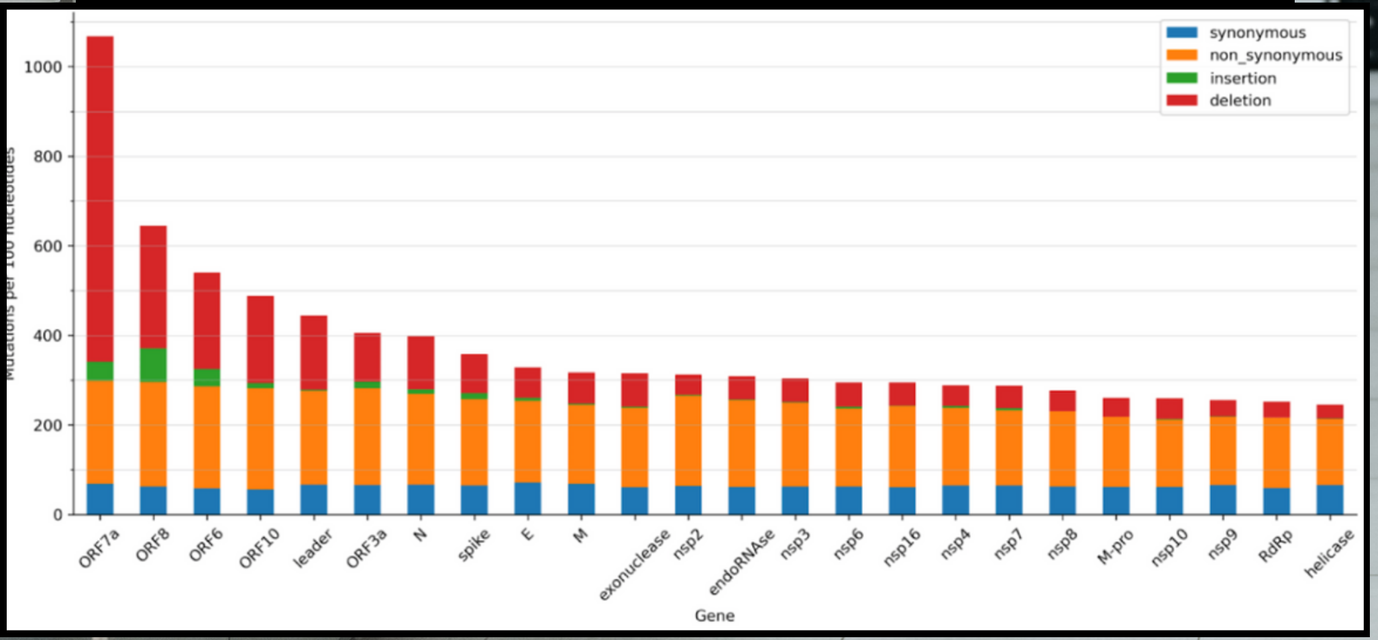


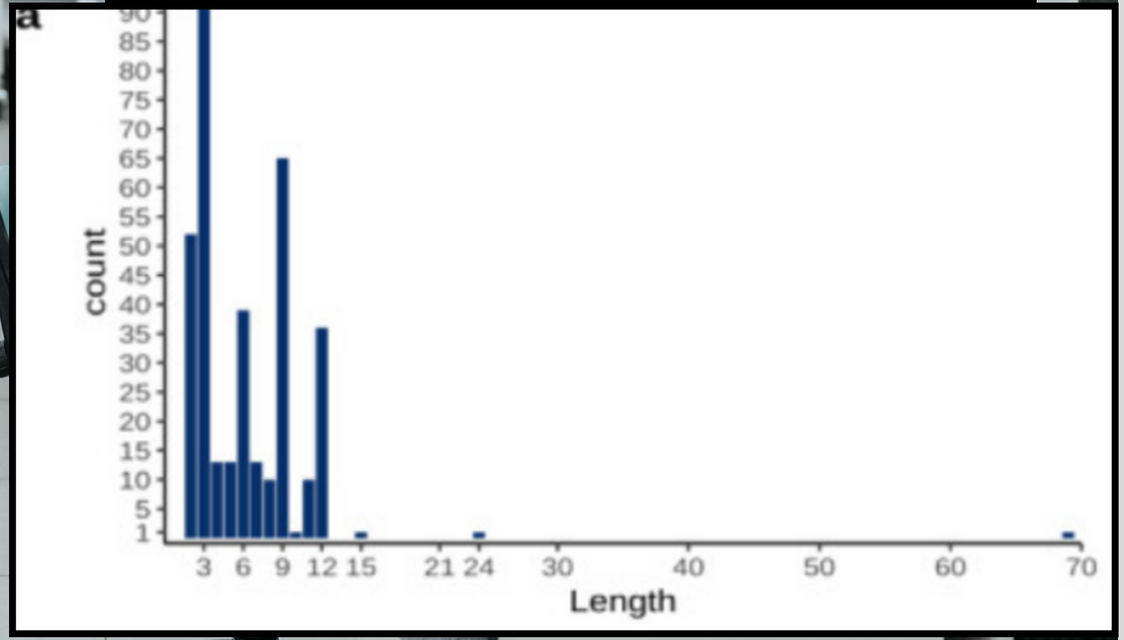
Deletion at end of genome

Rarity of long insertions in SARS-CoV-2

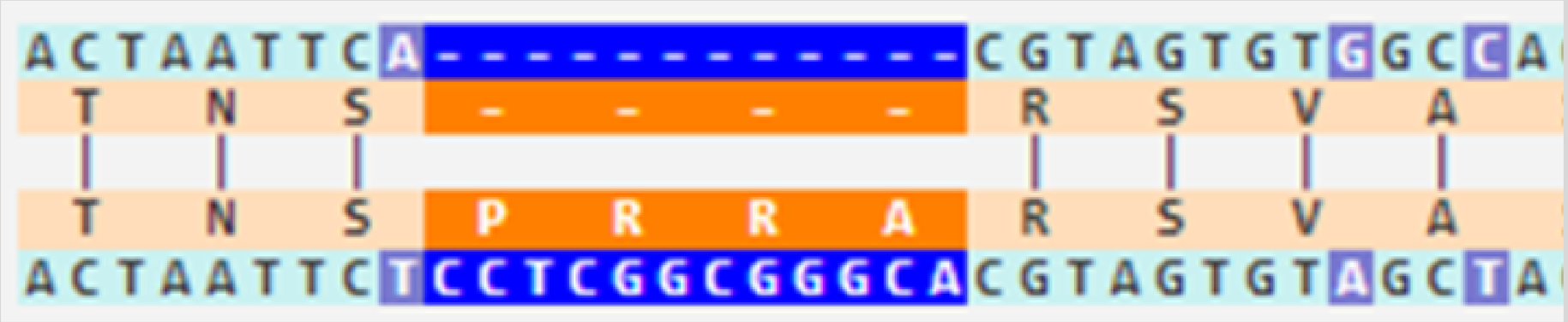
An insertion is the <u>rarest type of mutation</u>.

Long insertions are <u>even rarer</u>.





Factor 2 - Arginine CGG-CGG Coding Unlikely in Nature



The two Arginine (R) amino acids use the CGG codon.

- **CGG is the rarest codon in SARS-like viruses (and most viruses).**
- **<u>Appears</u> in 2.6% of Rs in the SARS2 genome (outside the FCS).**
- **Here it appears in both Rs, in the most critical feature of SARS2.**
- ***<u>"In fact, we have checked all 255 sarbecovirus strains present in GenBank that have protein annotations, and with the exception of SARS-CoV-2, none have two consecutive arginines coded by CGGCGG anywhere in their genomes (on average, each sarbecovirus strain has 12 arginine doublets in its annotated proteins)."</u>***
- **<u>Doesn't appear in any FCS of other viruses.</u>**

Factor 2 - Arginine CGG Frequency in SARS2

Table 3. Arginine codon usage in NC_045512.2 SARS-CoV-2, isolate Wuhan-Hu-1, genome

Gene	AGG	AGA	CGG	CGA	CGT	CGC	Total
nsp1	0	0	0	1	7	2	10
nsp2	2	5	0	2	7	3	19
nsp3	6	24	3	2	8	2	45
nsp4	2	11	0	0	5	2	20
3C-like proteinase	4	3	0	1	2	1	11
nsp6	1	6	0	0	1	1	9
nsp7	1	1	0	0	0	0	2
nsp8	2	3	0	0	2	0	7
nsp9	2	2	0	1	1	0	6
nsp10	0	0	0	0	1	1	2
nsp12	5	19	2	1	9	7	43
nsp13	2	14	1	2	9	2	30
nsp14A2	1	14	0	0	5	2	22
nsp15-A1	1	4	1	0	2	1	9
nsp16_OMT	2	6	0	0	0	1	9
S	10	20	2	0	9	1	42
ORF3a	1	3	0	0	1	1	6
ORF4	0	1	0	1	1	0	3
ORF5	3	3	0	1	5	2	14
ORF6	1	0	0	0	0	0	1
ORF7a	0	4	0	0	1	0	5
ORF8	0	2	0	0	2	0	4
ORF9	1	10	2	5	6	5	29
ORF10	0	1	0	0	1	0	2
Total	47	156	11	17	85	34	350

Following Peter's feedback, we obtained another source for CGG frequency, which agrees with our previous source:

(11-2)/(350-2) = 2.6%

Recreating from scratch in code the exact alignment of a virus is an error prone process. We could not invest the time to identify the exact problem but suspect the issue lies in correctly identifying open reading frames which can be tricky in CoVs.

Factor 2 - Arginine CGG-CGG Coding Reasonable for a Lab

Why would WIV choose these codons?

- **The exact reason WIV may choose these specific codons is yet unknown.**
- **However, we know they are not limited by whatever natural selection pressures made it rare in nature.**
- **Even as a random choice (1/6 vs 2.6%, squared) it is much more likely (41x)**
- **CGG is top of mind for human genetic engineers: Moderna has recoded 39 of the 42 SARS2 spike arginines by CGG while Pfizer has recoded 19 of 20 CGx spike arginines by CGG**
- **One possible reason: Using a rare codon allows easy screening of samples where the FCS has mutated away.**
  - **Specifically, this sequence introduces a new FauI restriction site.**

Table 1

Optimization of compound codon families in the two mRNA vaccines.

AA	Codon	RP [1]	Bkground [2]	S_{Ref} [3]	$S_{BNT\text{-}162b2}$ [3]	$S_{mRNA\text{-}1273}$ [3]
R	AGA	257	0.2640	20	21	0
R	AGG	230	0.2262	10	1	2
R	CGA	169	0.2640	0	0	0
R	CGC	306	0.2178	1	1	0
R	CGG	229	0.2262	2	19	39
R	CGU	171	0.2920	9	0	1
L	CUA	69	0.2640	9	0	1
L	CUC	215	0.2178	12	3	2
L	CUG	440	0.2262	3	105	103
L	CUU	203	0.2920	36	0	1
L	UUA	50	0.2640	28	0	1
L	UUG	172	0.2262	20	0	0
S	AGC	144	0.2178	5	64	96
S	AGU	95	0.2920	17	0	0
S	UCA	80	0.2640	26	0	2
S	UCC	194	0.2178	12	22	1
S	UCG	37	0.2262	2	0	0
S	UCU	180	0.2920	37	13	0

Open in a separate window

[1] Coding sequences of ribosomal proteins (34 and 53 in the small and large subunits, respectively. Only longest isoform for each gene is included); [2] Nucleotide frequencies from all introns in human chromosomes 18–22 (NC_000018-NC_000022) as a proxy of mutation bias at the third codon site. An A-ending codon has nucleotide frequency of nucleotide A; [3] Spike protein gene in reference SARS-CoV-2 genome (NC_045512) and BNT-162b2.

Factor 2 - Arginine CGG-CGG Coding Reasonable for a Lab

- **And yes, FauI has been used in virology before, even for RFLP screening (e.g. in 2019):**

Plant Pathol J. 2019 Aug; 35(4): 389–392.
Published online 2019 Aug 1. doi: 10.5423/PPJ.NT.12.2018.0306
PMCID: PMC6706017
PMID: 31481862

Genetic Diversity of Seven *Strawberry mottle virus* Isolates in Poland

Mirosława Cieślińska*

Author information · Article notes · Copyright and License information · PMC Disclaimer

A loss-of-function mutation in *Itgal* contributes to the high susceptibility of Collaborative Cross strain CC042 to *Salmonella* infections

Click here

Jing Zhang, Megan Teh, Jamie Kim, Megan M. Eva, Romain Cayrol, Rachel Meade, Anastasia Nijnik, Xavier Montagutelli, Danielle Malo, Jean Jaubert

doi: https://doi.org/10.1101/723478

This article is a preprint and has not been certified by peer review [what does this mean?].

0 0 0 0 0 8

Previous

Posted August 02, 2019.

Download PDF

Print/Save Options

Plant Pathol J

The RFLP analysis showed a restriction fragment length polymorphism of the amplified RNA2 fragment of the seven virus isolates. Six different profiles were obtained after digestion of RT-PCR products with the enzymes *Bfa*I, **FauI**, *Hae*III, *Hinc*I and *Taq*I. Only two isolates - Pink-1108 and Granat-1108 showed the same restriction patterns for each of the separately used enzymes (Fig. 1, Table 1). This result indicated that when selecting the suitable restriction enzymes for digesting of the RNA2 fragment, the RFLP technique can be useful and reliable method for the study on genetic variability of SMoV strains.

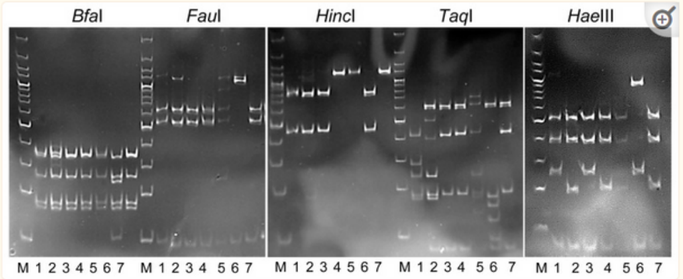


Fig. 1

Polyacrylamide gel showing the RFLP patterns of RNA2 fragment of the *Strawberry mottle virus* isolates amplified with 2SMoV1108F-2SMoV2386R primers digested with *Bfa*I, **FauI**, *Hae*III, *Hinc*I, *and Taq*I restriction enzymes of strawberry samples collected in 2015 in Bulgaria and Poland. Lanes: L – Thermo Scientific GeneRuler 100 bp Plus DNA Ladder, fragment sizes in base pairs (from top to bottom): 3000, 2000, 1500, 1200, 1000, 900, 800, 700, 600, 500, 400, 300, 200, 100. Samples: 1. Unkn-1108, 2. Granda-1108, 3. Markat-1108, 4. Pink-1108, 5. karkas-1108, 6. Pegat-1108, 7. Granat-1108.

Infect Immun

Itgal genotyping. Amplification of the region containing the CC042 *Itgal* deletion was conducted using a standard PCR (forward primer, 5′-TGCTTGGGTGTAGGCAGCCTCA-3′; reverse primer, 5′-CTTCAATCTGCAAGACCTGGTA-3′). DNA amplicons were digested using **FauI** with CutSmart buffer (catalog number R0651S; New England BioLabs) for 4 h at 55°C. The reaction was stopped by incubating the samples at 80°C for 15 min. The digested DNA was run on a 1.5% agarose gel in Tris-borate-EDTA (TBE) buffer.

Why the Leading Proline?

The amino acid sequence of the insert is PRRA, while theoretically RRA would suffice. There are several coronaviruses in nature with a P near the S1/S2 junction. Since this feature exists in nature, and is not necessary in engineering, it is claimed to be evidence for zoonosis.

In general, it is difficult to claim a very low probability for specific features based only on having 'no known reason to engineer them', as we don't know the exact intentions of the scientist.

Nevertheless, there are two reasonable explanations for inserting the Proline

MERS also has a Proline just before its FCS, which could provide inspiration for a lab to experiment with it.

RmYN02 S1/S2 cleavage site is PAAR, similar to SARS2's PRRAR. So if anyone in WIV came across a RmYN02-like virus with a PAA fragment (they had 180 unpublished viruses), they could choose to simulate how this could turn into a MERS-like FCS in nature. PRRAR FCSs were actually found in nature, among felines. Moreover, in 2017 Ben Hu of WIV thanked Libiao Zhang for collecting many samples across China. A 2019 paper by Zhang was based on samples from a location just 15 km away from where RmYN02 was found.

FCS Summary

While it's possible a coronavirus will develop an FCS naturally and start a pandemic. If that happens, we expect it to look wildly different:

- Not be the first virus in its family to have an FCS.
- Created by a small number of SNVs relative to original virus, rather than a clean insertion (e.g. PAAR in RmYN02 is just a single nucleotide mutation away from RAAR).
- If by insertion, a small one (e.g. 3 nt).
- If by a long insertion, the sequence would come from another part of the virus (but even that would already be very rare).
- Use common codons of the original virus.

Point 3

Low genetic variability early in the pandemic is indicative of a quick, localized jump of a virus that is already pre-selected for human tropism and possibly futher adapted for it in human cells and/or humanized mice, as expected in a lab leak but not in zoonosis.

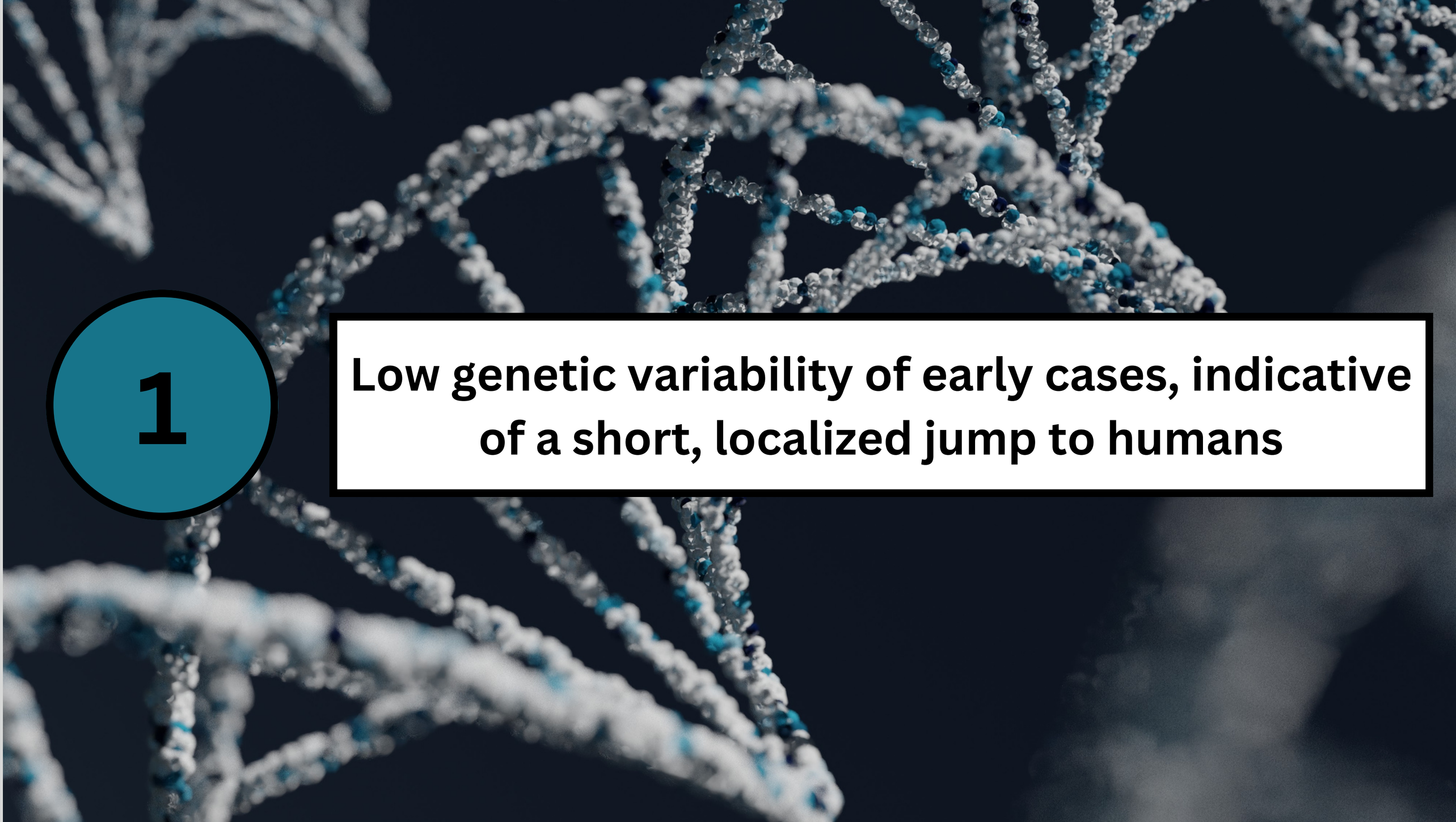
Genetics of Early Cases

In this section, we will show that:

Early cases had low genetic variability, indicative of a short, localized jump to humans.

The market was dominated by later strains, indicative it is not a spillover location.

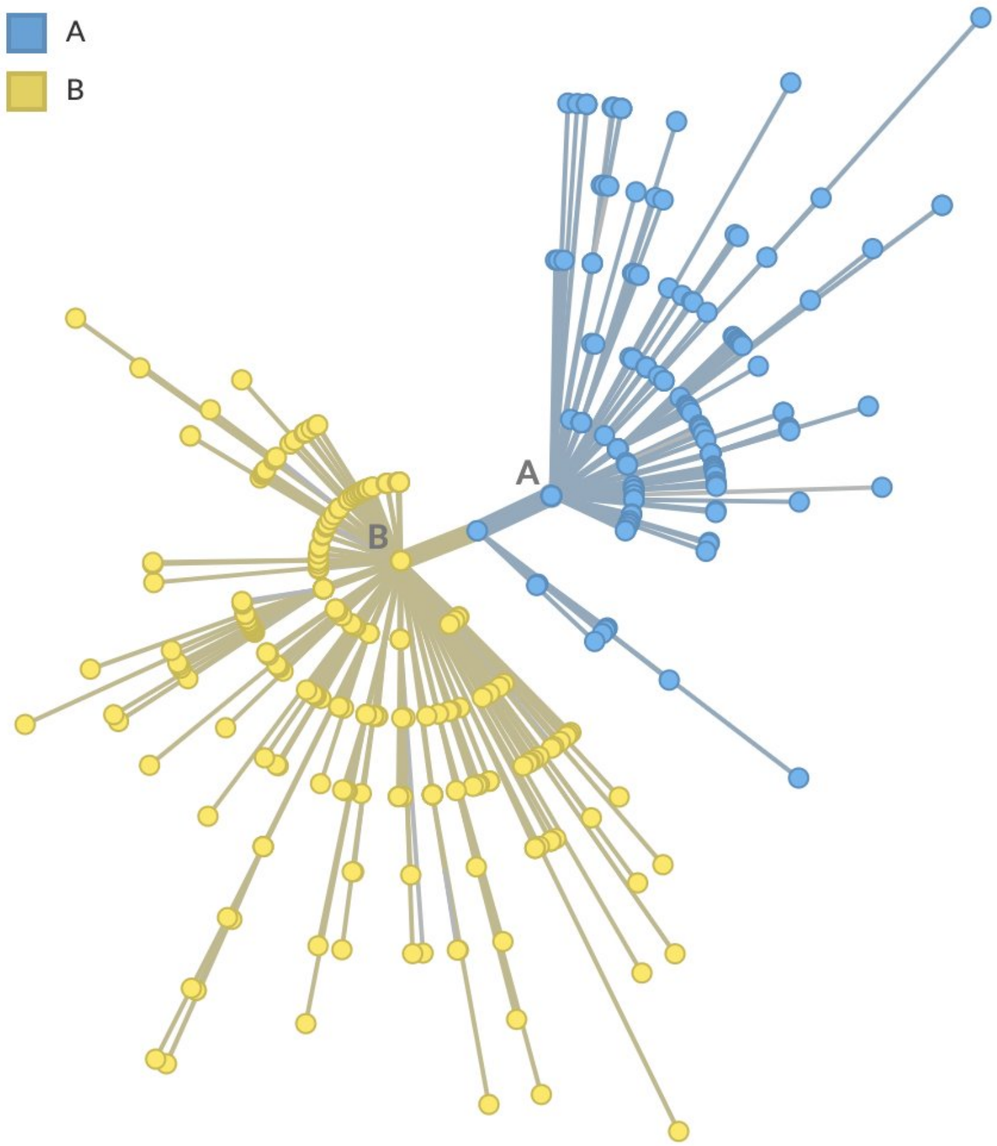
Low early mutation rate, indicative of prior adaptation to humans.



1

Low genetic variability of early cases, indicative of a short, localized jump to humans

Low Genetic Variability



nextstrain build targeted at SARS-CoV-2 genomes from Dec 2019 through Jan 2020, totaling 549 viruses

Pekar et al. claim of 2 separate jumps is erroneous

Table S5. Frequencies of observed topologies in epidemic simulations and corresponding Bayes factor in favor of multiple introductions versus a single introduction across varying doubling times, varying ascertainment rate, minimum polytomy size, and phylogenetic rooting method.

Analysis			Topology			Bayes factor	
DT	Asc	Min. polytomy size	C/C	A/B	Polytomy	Unconstrained	recCA
2.65	0.15	100	0.0	1.2	58.6	28.8	29.5
3.47[1]	0.15	100	0.0	0.5	47.5	60.0	61.6
4.45	0.15	100	0.1	0.3	43.1	86.2	87.7
3.50	0.05	100	0.0	0.5	45.7	57.7	59.2
3.52	0.25	100	0.2	1.0	47.3	26.7	27.2
3.47	0.15	20	0.1	1.6	60.7	21.5	22.0
3.47	0.15	50	0.1	0.8	53.6	37.2	38.0
3.47	0.15	200	0.0	0.3	40.7	85.4	87.7
3.47	0.15	500	0.0	0.2	31.7	99.7	102.3

DT, Median doubling time
Asc, Ascertainment rate
Min., Minimum

Erratum reduced Bayes Factors by ~6x

Table S5. Frequencies of observed topologies in epidemic simulations and corresponding Bayes factor in favor of multiple introductions versus a single introduction across varying doubling times, varying ascertainment rate, minimum polytomy size, and phylogenetic rooting method.

Analysis			Topology			Bayes factor	
DT	Asc	Min. polytomy size	C/C	A/B	Polytomy	Unconstrained	recCA
2.65	0.15	100	0.0	3.4	58.6	5.8	6.0
3.47[1]	0.15	100	0.0	3.1	47.5	4.2	4.3
4.45	0.15	100	0.1	3.5	43.1	3.0	3.0
3.50	0.05	100	0.0	3.4	45.7	3.5	3.6
3.52	0.25	100	0.2	3.5	47.3	3.6	3.7
3.47	0.15	20	0.1	5.1	60.7	4.1	4.2
3.47	0.15	50	0.1	3.9	53.6	4.2	4.3
3.47	0.15	200	0.0	2.1	40.7	4.5	4.6
3.47	0.15	500	0.0	1.1	31.7	5.2	5.4

DT, Median doubling time

https://www.science.org/doi/10.1126/science.adl0585

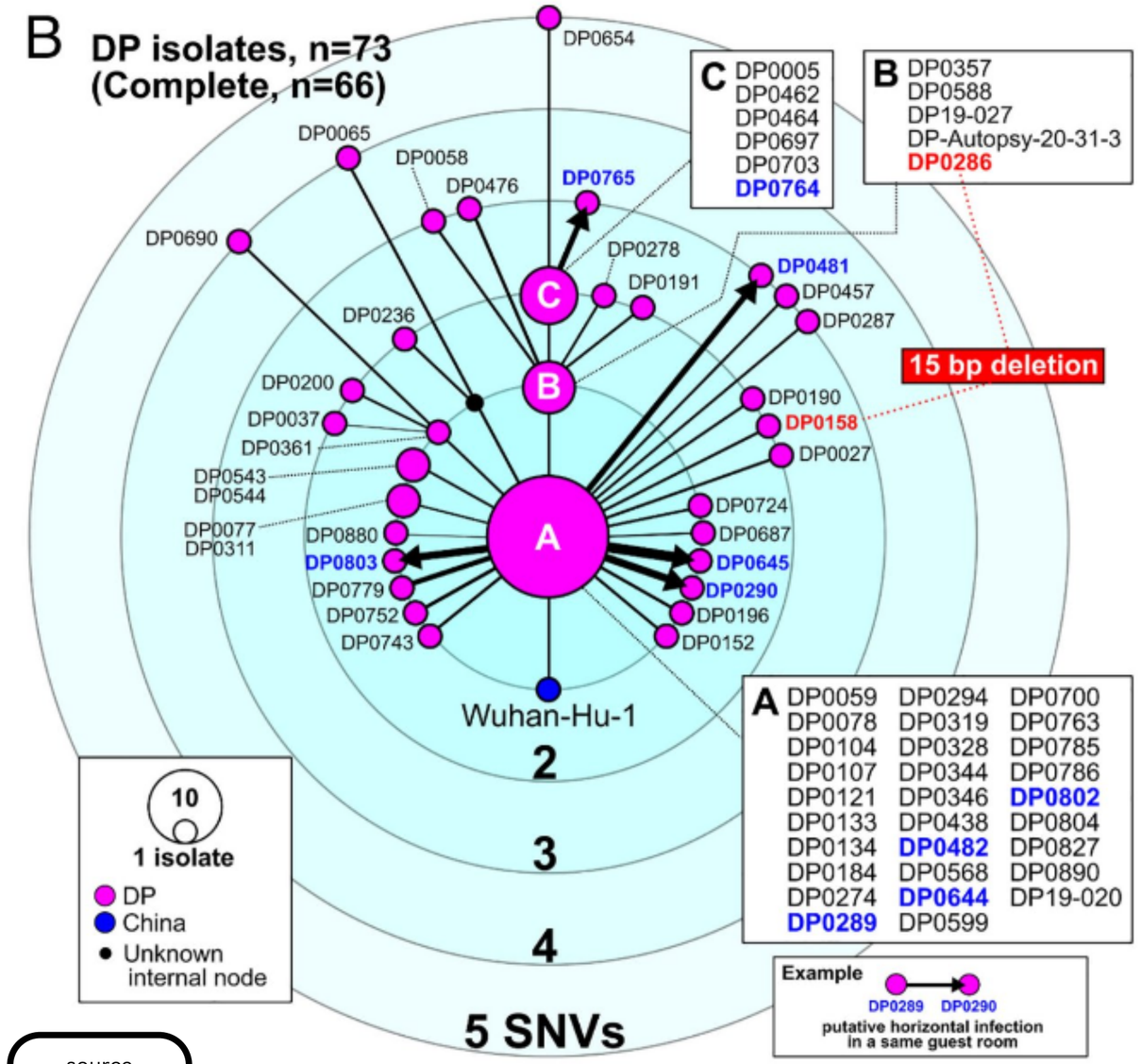
A/B lineages are likely not separate jumps

Only 2 mutations apart, not 10

This can happen in a single individual, as seen in the Diamond Princess cruise ship analysis.

In that case we will never see an intermediate genome

Alternatively, the intermediate could have died out, which is likely to happen when still few people are infected.



source

Other Pekar et al. issues

- **Pekar et al. is just one model based on a simulation, there are others:**

TopHap: rapid inference of key phylogenetic structures from common haplotypes in large genome collections with limited diversity

Marcos A Caraballo-Ortiz, Sayaka Miura, Maxwell Sanderford, Tenzin Dolker, Qiqing Tao, Steven Weaver, Sergei L K Pond, Sudhir Kumar Author Notes

- **Pekar threw out intermediate genomes:**

Unwarranted Exclusion of Intermediate Lineage A-B SARS-CoV-2 Genomes Is Inconsistent with the Two-Spillover Hypothesis of the Origin of COVID-19

by Steven E. Massey [1,*], Adrian Jones [2], Daoyu Zhang [3], Yuri Deigin [4] and Steven C. Quay [5]

Pekar et al. threw out valid A/B intermediate genomes

Steve Massey
@stevenemassey

We have discovered 6 more A/B intermediate SARS2 genomes (with C/C genotype) from Wuhan 🧵

This is in addition to the 7 new intermediates (also C/C) we previously identified from Sichuan 👇

Traduire le post

Steve Massey @stevenemassey · 13 sept. 2022

We have discovered 5 more A/B intermediate genomes from Sichuan (with a C/C genotype), in addition to 2 we found previously 🧵

These were not considered by Pekar et al, despite 2 of the 5 conforming to their inclusion criteria 👇

Materials and Methods

Sequence data. We queried the GISAID database SARS-CoV-2 viral genome alignment for sequences collected by 14 February 2020 (*57*). We selected this date to have a data set whose size is appropriate for Bayesian phylodynamic analyses (*i.e.*, under 1000 genomes). We restricted our data set to sequences that (i) were ≥29,000 nucleotides, (ii) had high coverage with ≤0.5% unique amino acid mutations, (iii) had fewer than 1% 'N's, (iv) were not identified as potentially problematic via NextStrain (*67*), and (v) had a year-month-day sampling date reported. We additionally queried for the

This further challenges the Pekar et al. hypothesis of two distinct SARS2 introductions

Earliest sampled A genome had an extra mutation

Earliest lineage B and lineage A were only 5 days apart and A already had an extra mutation (T4946C) away from bat consensus, meaning that it likely was circulating for some time before being sampled.

The earliest unambiguous case of COVID-19, with symptom onset on 10 December and hospitalization on 16 December, was a seafood vendor at the Huanan market. Unfortunately no published genome is available for this case (*8*). Nonetheless, we can reasonably assume this individual had a lineage B virus (supplementary text), as an environmental sample (EPI_ISL_408512) from the stall this vendor operated was lineage B. The earliest lineage A genome (IME-WH01) is from a familial cluster where the earliest symptom onset is 15 December and earliest hospitalization is 25 December (*34*). Accounting for these dates and using the recCA rooting, we inferred the infection date of the lineage B primary case to be 18 November (95% HPD: 23 October to 8 December) and the infection date of the primary case of lineage A to be 25 November (95% HPD: 29 October to 14 December). The lineage B primary case predated that of lineage A in 64.6% of the posterior sample, by a median of 7 days (Fig. 3D and table S6).

S13	2019/12/26	2019/12/30	SARS-CoV-2/Wuhan_IME-WH01/human/2019/CHN	4946, 8782, 28144	4946, 8782, 28144	ThermoFisher S5Plus	176	0.53

Download

Query range 81: 4801 to 4860

```
Query       4801   ACCACATTCCACCTAGATGGTGAAGTTATCACCTTTGACAATCTTAAGACACTTCTTCCT   4860   Wuhan_IME-WH01 (Lin A)
MN908947.3  4889   ............................................................T..   4948   Wuhan-Hu-1 (Lin B)
MN996532.2  4886   .TT.....T...................................T......CT..   4945   RaTG13
```

Pekar et al - Issues with Spillover Time Modeling

Due to the market bias, lineage A is expected to have been sampled less and later

This bias must be corrected before using these data to assume jump dates (”garbage in garbage out”)

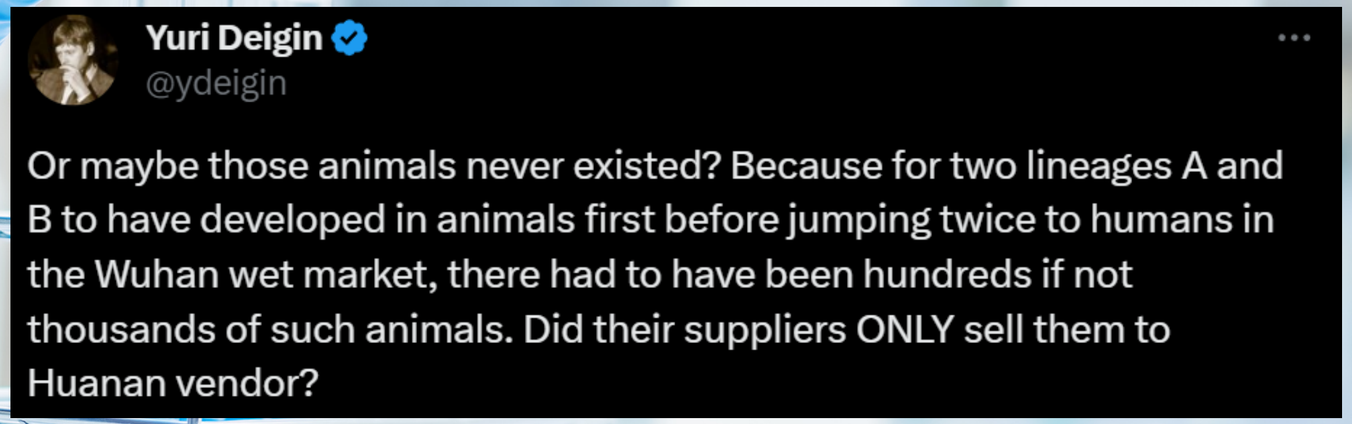
One of the two mutations from A to B is not synonymous, and B appears to be more infectious.

It is therefore possible that this specific mutation was under strong selective pressure and therefore emerged faster than others, making genetic clock models inaccurate.

Two Jumps in this Pattern are Likelier in a Lab Leak

Even if there were two separate jumps, since they occurred in the same location within a short time frame, they don't strengthen the zoonosis case:

Two spillovers can well happen in a lab; One of the three SARS1 lab leaks <u>had two jumps</u> from the same lab.

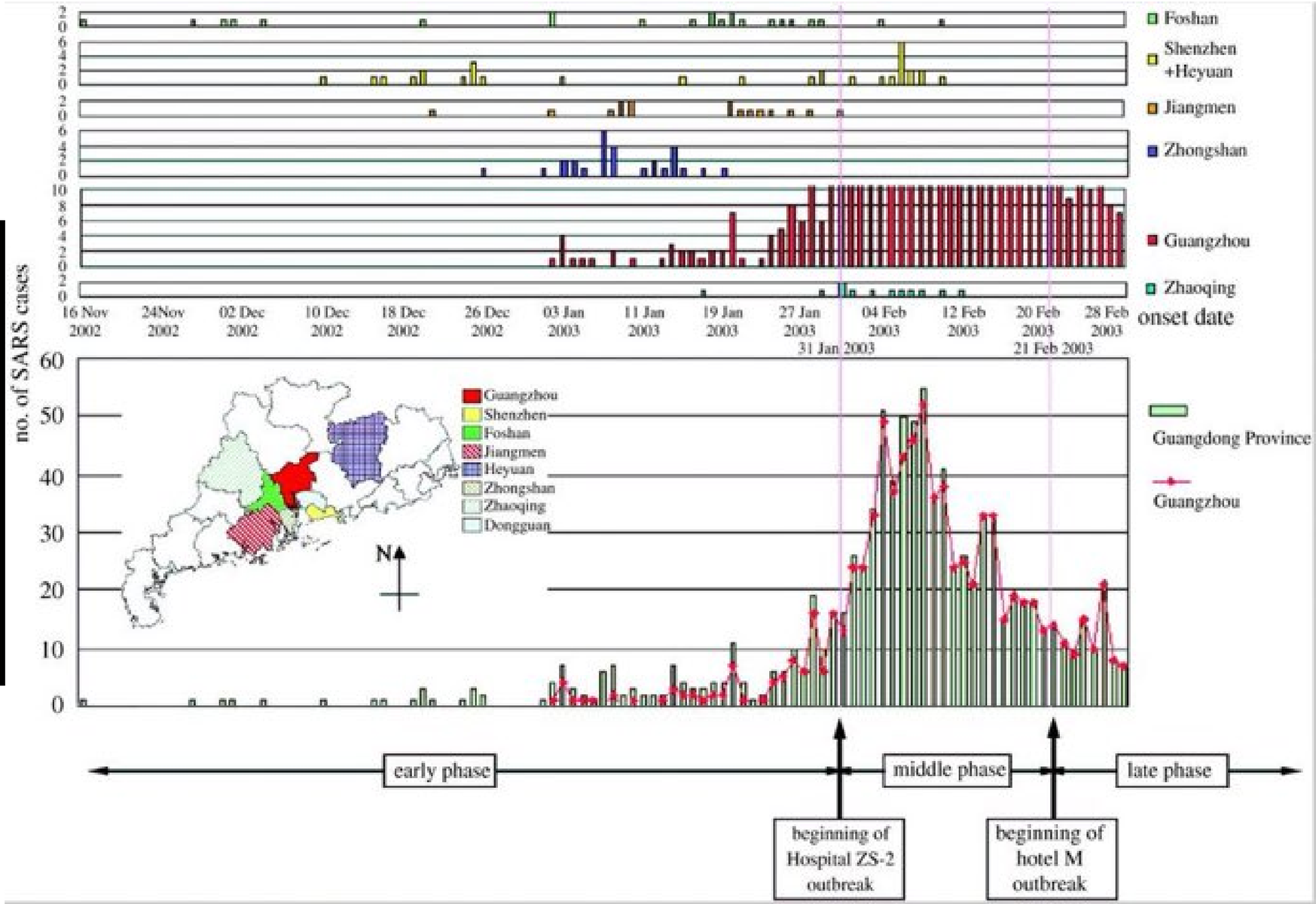


Importantly, two spillovers from wildlife imply many infected animals in contact with humans, which would make it much more unlikely that Wuhan will be the only outbreak.

Comparing to SARS1 emergence

A long period of sparse infections over multiple locations, until the superspreading event

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2435571/



Comparing to HIV

ABSTRACT

Go to:

The major cause of acquired immune deficiency syndrome (AIDS) is human immunodeficiency virus type 1 (HIV-1). We have been using evolutionary comparisons to trace (i) the origin(s) of HIV-1 and (ii) the origin(s) of AIDS. The closest relatives of HIV-1 are simian immunodeficiency viruses (SIVs) infecting wild-living chimpanzees (*Pan troglodytes troglodytes*) and gorillas (*Gorilla gorilla gorilla*) in west central Africa. Phylogenetic analyses have revealed the origins of HIV-1: chimpanzees were the original hosts of this clade of viruses; four lineages of HIV-1 have arisen by independent cross-species transmissions to humans and one or two of those transmissions may have been via gorillas.

Source

Comparing to MERS

From sequence data, we identify at least 50 zoonotic introductions of MERS-CoV into humans from the reservoir (Figure 1), from which we extrapolate that hundreds more such introductions must have taken place (Figure 3). Although we recover migration rates from our model (Figure 1

Source

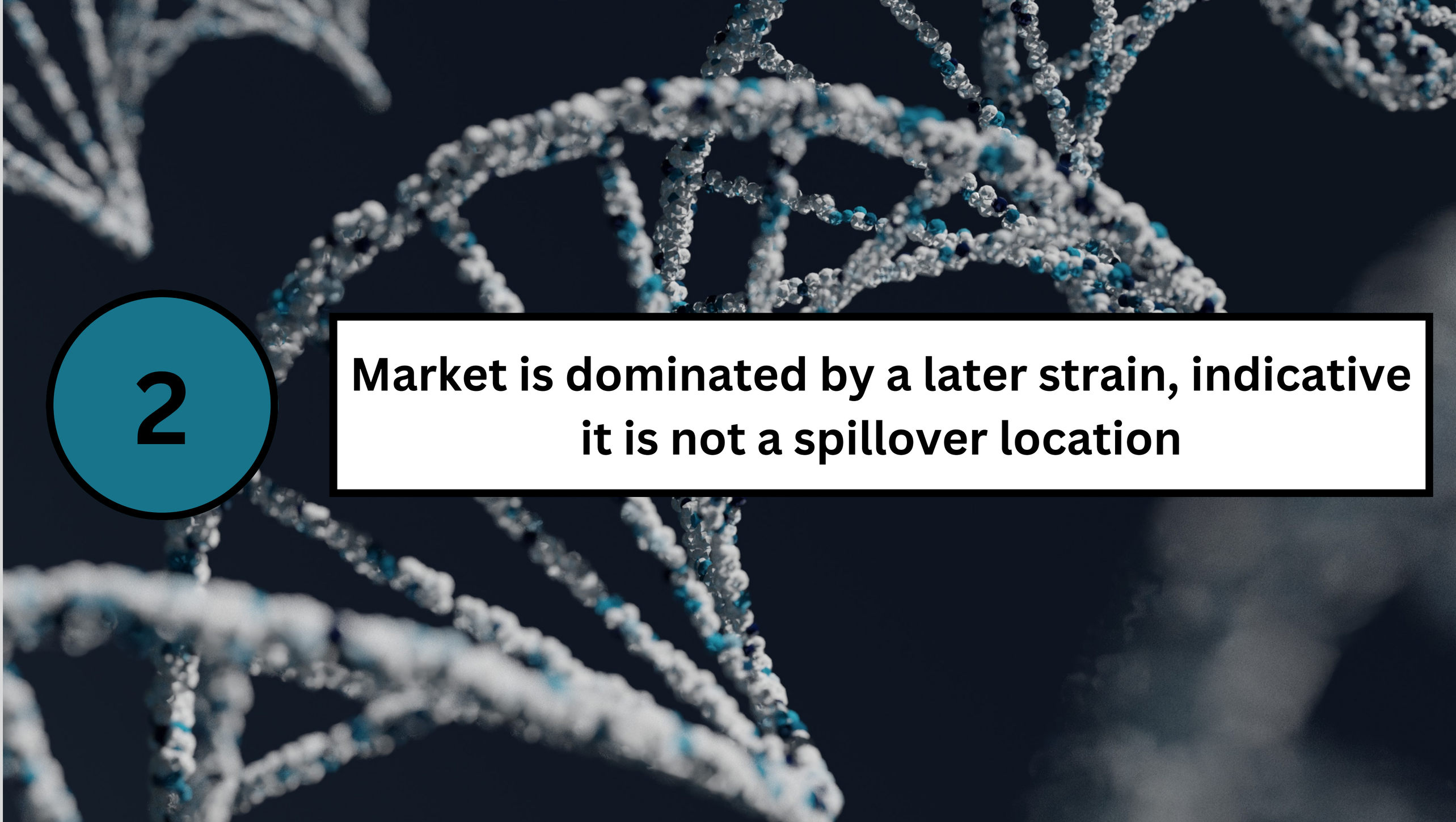
A/B lineages - Summary

Expectation

Zoonosis should show multiple jumps, more than 2 mutations apart, in multiple locations.

Reality

SARS2 shows one location in short time, likely from a single jump.



2

Market is dominated by a later strain, indicative it is not a spillover location

A is ancestral to B, and the market is dominated by B

Lineage A is two mutations closer to ancestral bat viruses than lineage B and almost certainly B evolved from A before eventually outcompeting it into oblivion.

But all 16 earliest Wuhan patients with link to the market had lineage B

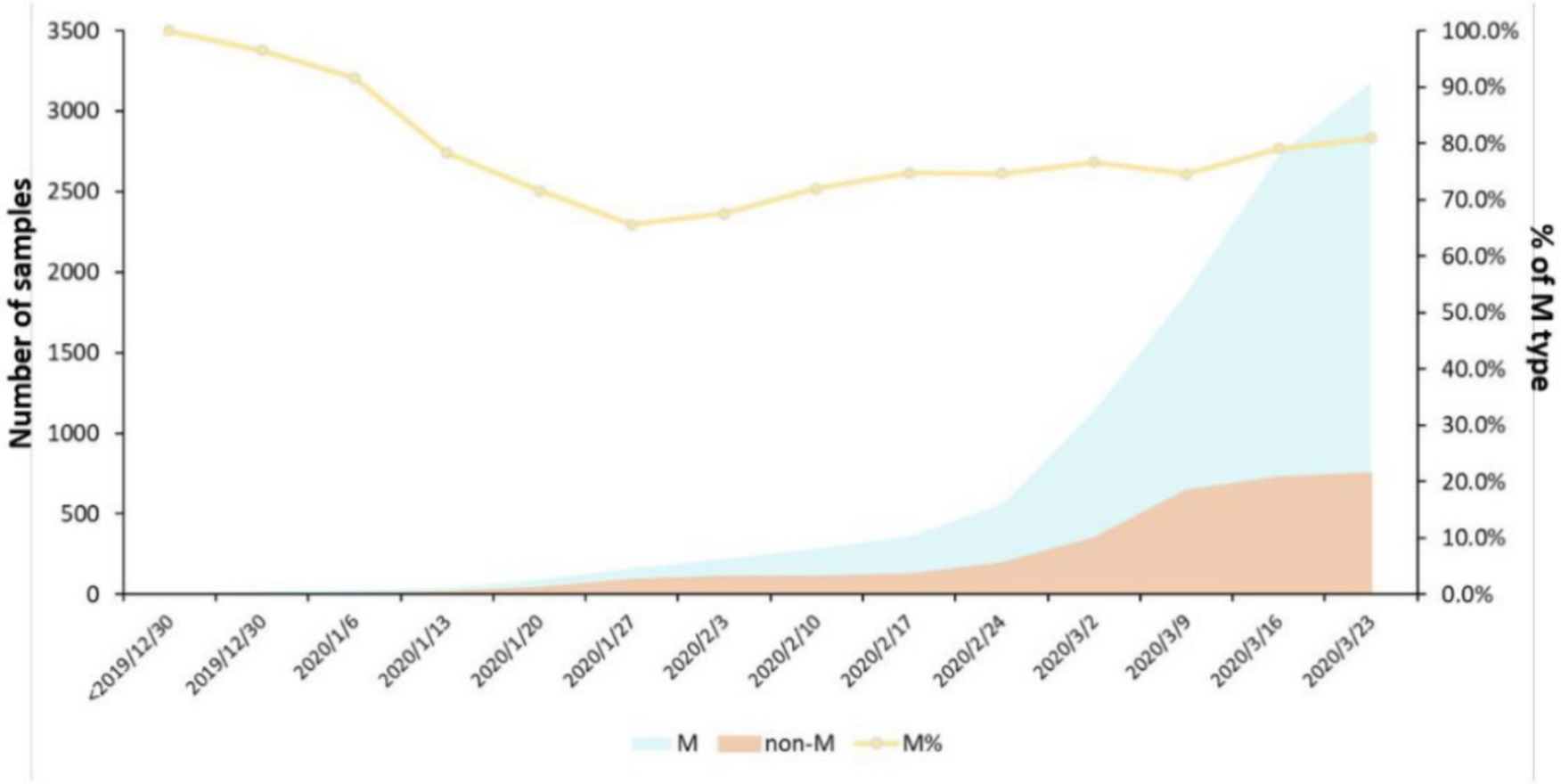
Also, all positive environmental samples in the market, except one, were lineage B, and the sole lineage A sample has provenance issues (to be addressed in detail later)

A is ancestral to B, and the market is dominated by B

Supplemental figures

Figure S1

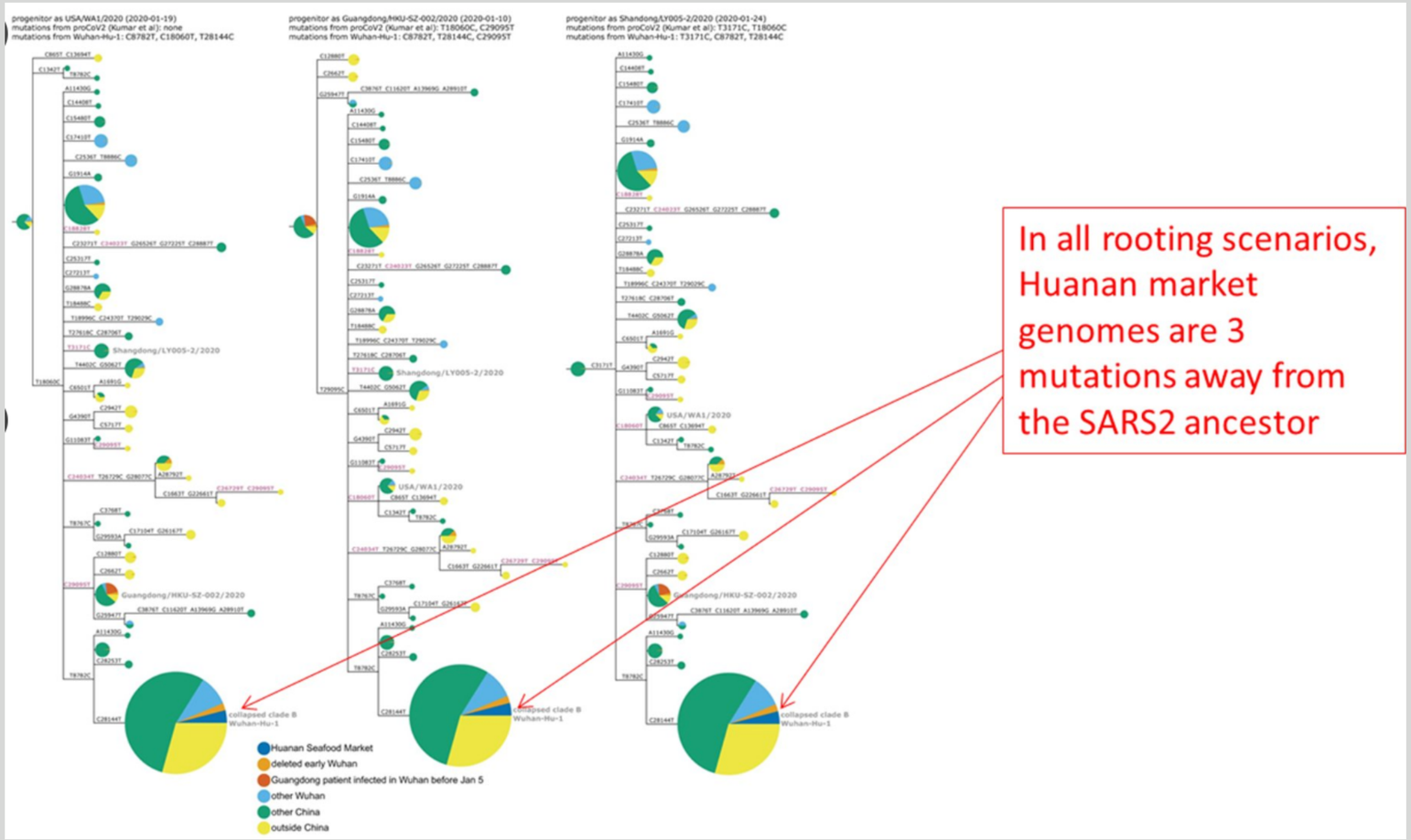
Outside the market, lineage A accounted for 33% of Wuhan early cases, and was quickly overtaken by B



M is same as lineage B

A is ancestral to B, and the market is dominated by B

However, lineage A itself is not at the root of the SARS2 ancestry tree because several phylogenetically earlier genomes are known, i.e. ones that have even fewer mutations than lineage A when compared to bat viruses like RaTG13 or BANAL-52.



A is ancestral to B, and the market is dominated by B

- One such mutation is C18060T and several investigators of SARS2 phylogeny (like Bloom or Kumar et al. 2021) think that it is likely that the earliest ancestor of all human SARS2 viruses had that mutation.

- Such an ancestor is sometimes called proCoV2, and it is basically lineage A with the C18060T mutation (so, in total, it is 3 mutations away from Huanan market's lineage B: C8782T, C18060T, and T28144C).

An Evolutionary Portrait of the Progenitor SARS-CoV-2 and Its Dominant Offshoots in COVID-19 Pandemic

Sudhir Kumar, Qiqing Tao, Steven Weaver, Maxwell Sanderford, Marcos A Caraballo-Ortiz, Sudip Sharma, Sergei L K Pond, Sayaka Miura

Molecular Biology and Evolution, Volume 38, Issue 8, August 2021, Pages 3046–3059, https://doi.org/10.1093/molbev/msab118

Published: 04 May 2021

A is ancestral to B, and the market is dominated by B

- While the situation with early Wuhan patient data is unclear, we do have some evidence that a number of early patients in Wuhan were infected by proCoV2:
  - The result from forensic metagenomics efforts by I. Csabai & N. Solymosi
  - Bloom has shown that there were a number of reads that are potentially consistent with a proCoV2 infection (as we see reads for all 3 of its key mutations, C8782T, C18060T, and T28144C)
  - While read count is low, they are the most popular in all samples, making a misread unlikely. The lower counts are likely due to mixing with lineage B.

site	accession	identity	Nucleotide at this site in reference genome (Wuhan-Hu-1)	Alternative nucleotide at this site	# of reads containing same nucleotide as in reference genome	# of reads containing alternative nucleotide at this site	Proportion of reads with alternative nucleotide	Nucleotide at this site in RaTG13	Nucleotide at this site in BANAL-20-52
8782	SRR13441704	high	C	T	1	3	75%	T	T
8782	SRR13441708	high	C	T	1	3	75%	T	T
18060	SRR13441705	high	C	T	4	6	60%	T	T
28144	SRR13441704	high	T	C	1	2	67%	C	C
28144	SRR13441705	high	T	C	1	6	86%	C	C
28144	SRR13441708	high	T	C	1	4	80%	C	C

A is ancestral to B, and the market is dominated by B

Bloom has found early sequences even ancestral to A

Yuri Deigin
@ydeigin

6/ Of course, @jbloom_lab has previously put together another compelling dataset based on his recovery of deleted sequencing data of early patient samples. In that dataset we see 3 patients infected by either lineage A or a proCoV2 lineage, 7 patients infected with lineage B or potentially an A/B intermediate lineage (we don't have sequencing data for positions 8782 or 18060 for these patients), and finally a patient with an additional mutation closer to bat ancestors, C29095T, which could mean they were infected even by a pro-proCoV2 progenitor.

This means either:

- **The mutation was a reversion (around 3% probability to hit an existing mutation, and not necessarily to the original nt)**
- **This was the strain that jumped from wildlife**
  - **This moves the jump date earlier to allow time for an extra mutation, invalidating all the A/B dating.**

A is ancestral to B, and the market is dominated by B

Table 1.
Samples for which the SARS-CoV-2 sequence could be called at ≥90% of sites between 21,570 and 29,550, and the substitutions in this region relative to the putative SARS-CoV-2 progenitor proCoV2 inferred by Kumar et al. (2021).

Sample	Fraction sites called (21,570–29,550)	Patient group	Substitutions relative to proCoV2
A4	0.9266	Early outpatient	None
C1	0.9396	Early outpatient	G22081A (A=924, C=4, G=9), C28144T (C=6, T=1185), T29483G (C=1, G=45, T=1)
C2	0.9397	Early outpatient	C29095T (C=1, G=1, T=751)
C9	0.9005	Early outpatient	C28144T (C=3, T=823), G28514T (G=1, T=36)
D9	0.9051	Early outpatient	C28144T (C=4, T=1653)
D12	0.9400	Early outpatient	C28144T (C=8, T=2400)
E1	0.9223	Early outpatient	C28144T (T=125)
E5	0.9227	Early outpatient	C24034T (A=5, C=3, T=74), T26729C (C=12), G28077C (C=142, G=4)
E11	0.9321	Early outpatient	C25460T (C=2, T=246), C28144T (C=1, T=412)
F11	0.9054	Early outpatient	T25304A (A=9, T=1), C28144T (C=6, G=1, T=1328)
G1	0.9396	Early outpatient	None
G11	0.9112	Early outpatient	None
H9	0.9381	Early outpatient	C28144T (C=2, T=1254)
R11	0.9422	Hospital patient (Feb)	C21707T (T=401), C28144T (A=1, C=18, T=4265)

Numbers in parentheses after each substitution give the deep sequencing reads with each nucleotide identity.

Sample C2 is missing C28144T, meaning it is lineage A.
There are a total of 4 mutations in the 5 lineage A samples, making a reversion possible but unlikely (3% x 4 times x reverting to the original nt)

Note: Only mutations above 21,570 are shown

A is ancestral to B, and the market is dominated by B

Some early genomes from both outside Wuhan and outside China show that there were dozens of early patients infected by strains that were ancestral to those seen in the Huanan market lineage B patients

In a nutshell

Earlier lineages were in circulation before a lineage B variant triggered the Huanan market superspreading event

It is clear from the genetics of the market cases that they were too far away from being the origin of the virus.

The most likely explanation is that earlier lineages were in circulation before a lineage B variant triggered the Huanan market superspreading event, thus further explaining why it concentrated early search efforts

A is ancestral to B, and the market is dominated by B

Huanan had only 1 lineage A sample but:

it was an environmental sample, found on a glove

had additional mutations (G26262T, C6145T, and possibly T24979C)

The lineage A genome was recovered only after passaging the A20 sample in culture, while direct sequencing of the A20 sample yielded only 22 SARS2 reads and no reads covering positions 8782 or 28144. Additionally, the original A20 sample had a very high PCR Ct value so it's very surprising to see a viable virus come out of that sample

Thus it is possible the lineage A genome in A20 was not present originally but was introduced during viral passaging of the sample in culture

Market is dominated by a later strain

- No other jump detected, whereas SARS1 had at least three:

Singapore - 2003	BSL-3	Foshan, Guangdong Province, China (Nov 2002) - Started with a farmer
Taiwan - 2003	BSL-4	Guangdong Province, China (Jan 2003) - Hotel Guest
China - 2004	BSL-2	Guangdong Province, China (Jan 2004) - Restaurant serving civets

- This is especially difficult to explain when claiming wildlife transported over 1000s of km did not reach any place other than Wuhan, or infect others on the way.

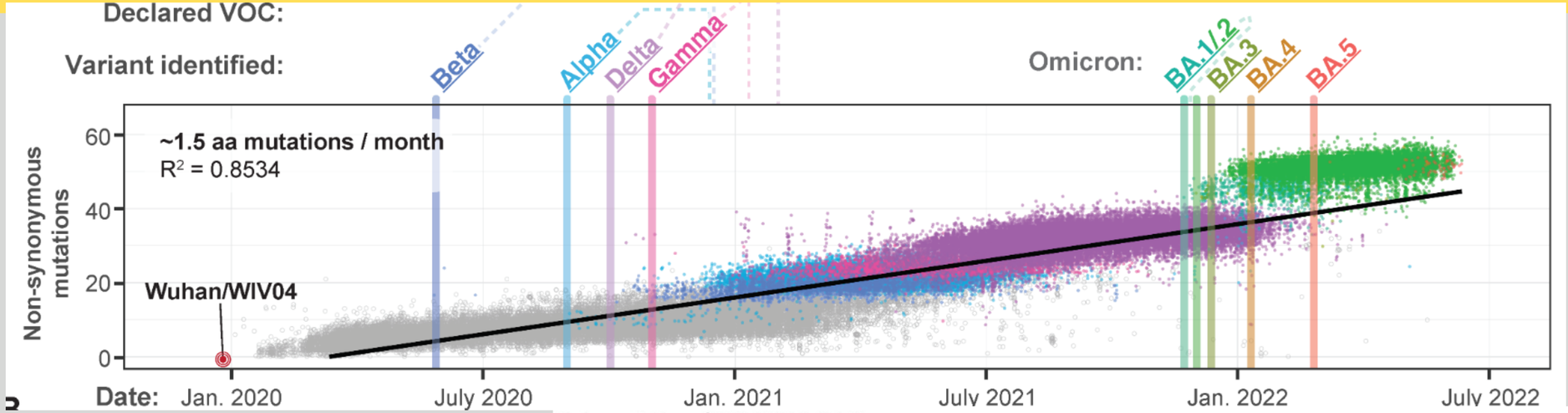
- Additionally, that is inconsistent with the two spillovers claim
  - two spillovers from two animals imply an even more widespread animal trade which surely should have left many traces and intermediate animal genomes — as in the case of SARS1

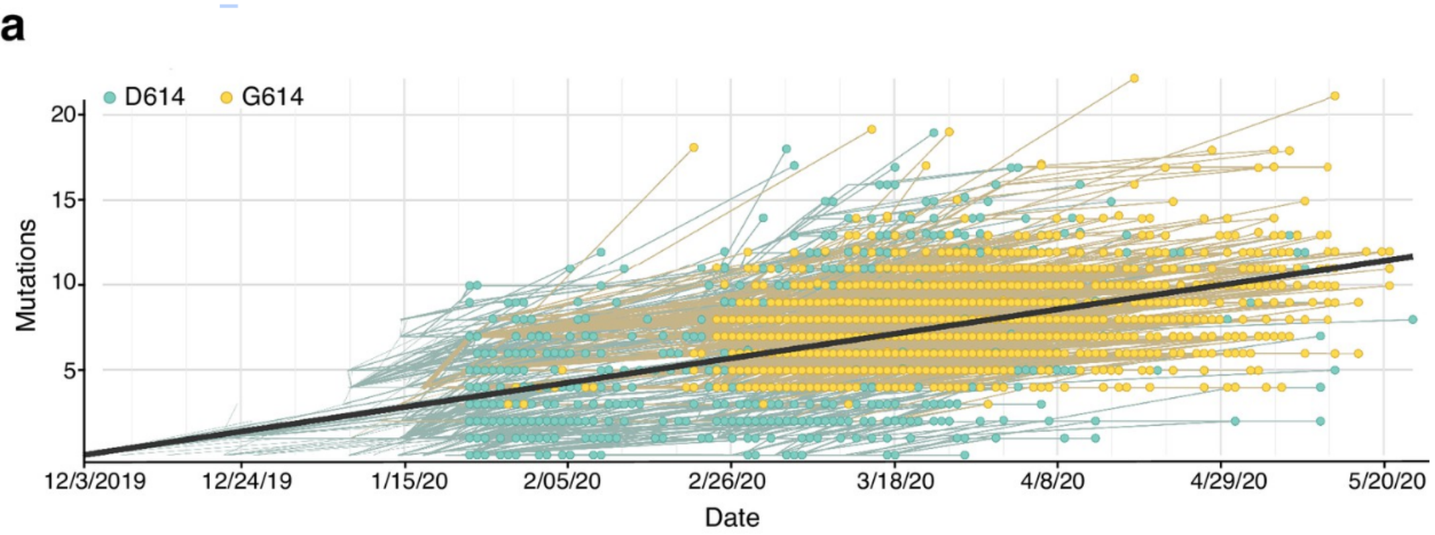
3

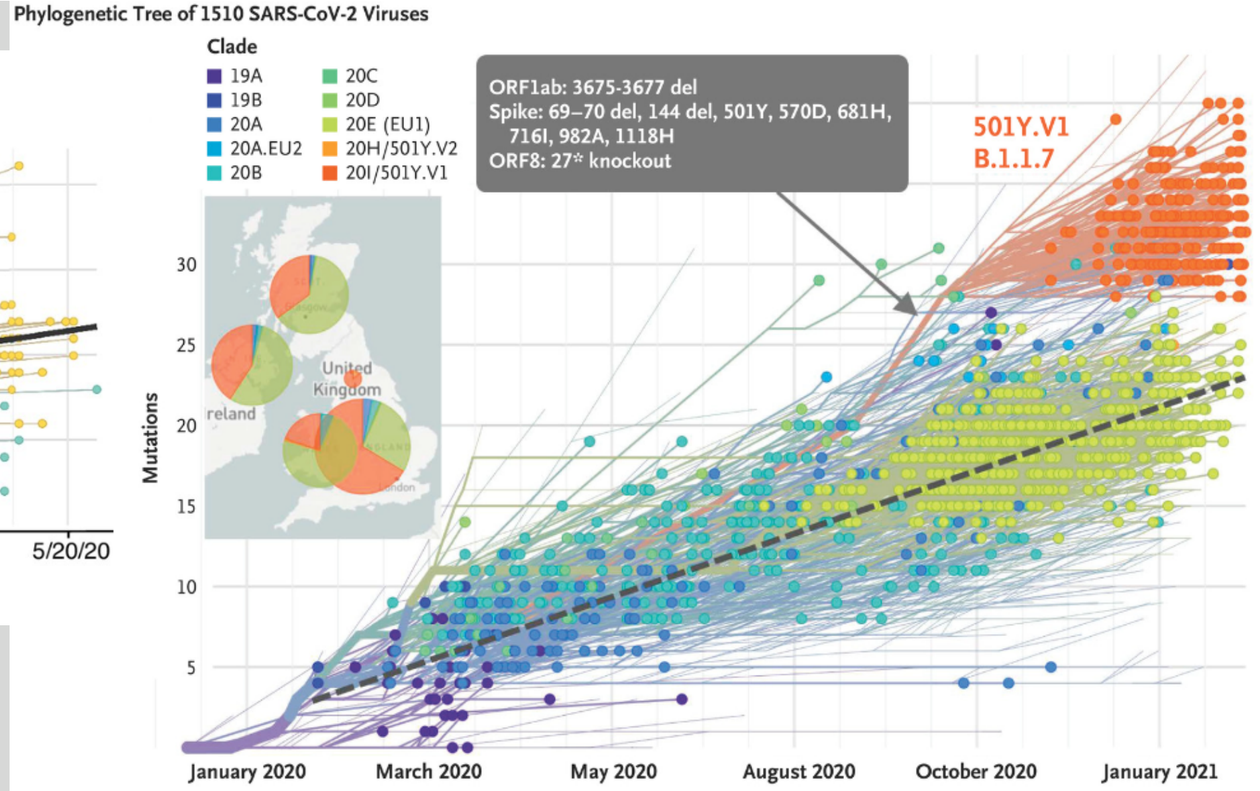
Low early mutation rate, indicative of pre-adaptation to humans

Low early mutation rate

SARS2 mutation rate is fairly constant since the start







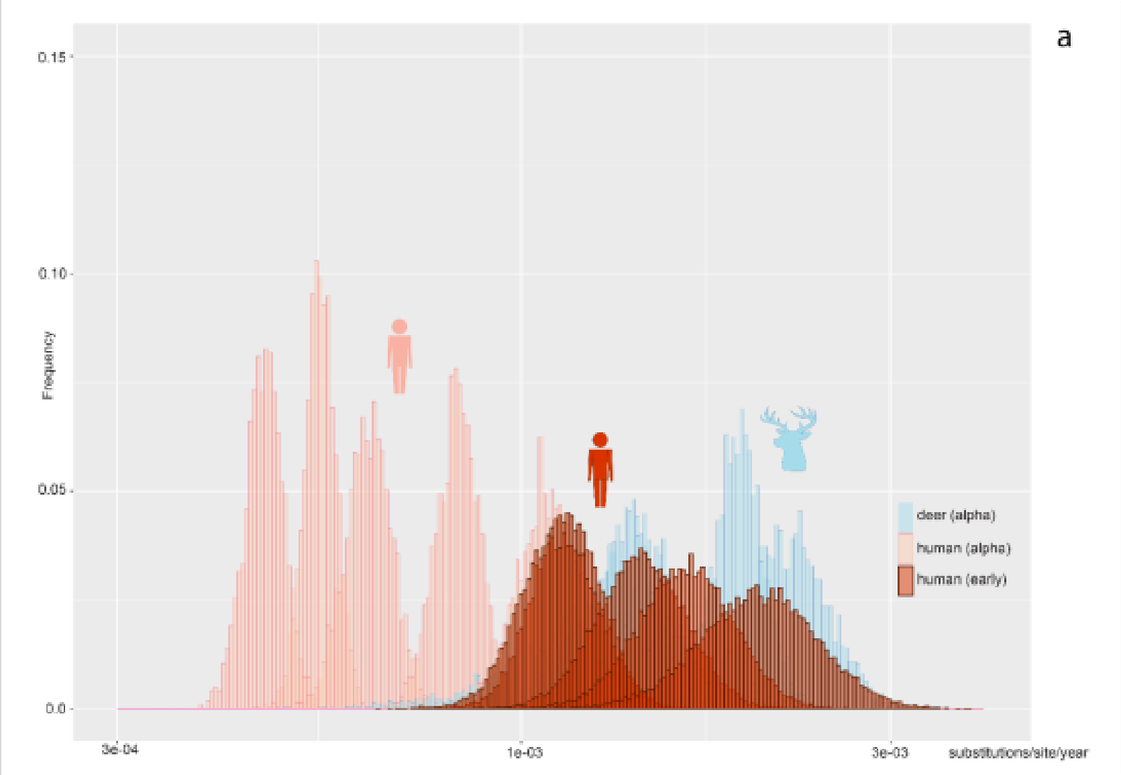
https://www.nejm.org/doi/full/10.1056/NEJMsb2104756
https://www.mdpi.com/1999-4915/14/9/2009
https://elifesciences.org/articles/65365

Low early mutation rate (vs. deer)

Early phase of Covid in humans: 37 mutations per year.
Alpha and delta strain, in humans: 18 mutations per year.

After covid spilled over into mink: 24 mutations per year.
After covid spilled over into deer: 36 mutations per year.

To test this hypothesis, we conducted an additional comparison with early SARS-CoV-2 strains collected in humans during December 2019 to February 2020 that were used in Pekar et al.. The overall rate of SARS-CoV-2 evolution was significantly higher in early human strains (1.3 × 10−3 substitutions/site/year; 95% HPD 1.1–1.6 × 10−3) compared to the alpha and delta strains that emerged in humans later in the pandemic (5.9–6.0 × 10−4), but not as high as the deer rate (1.6–1.8 × 10−3)



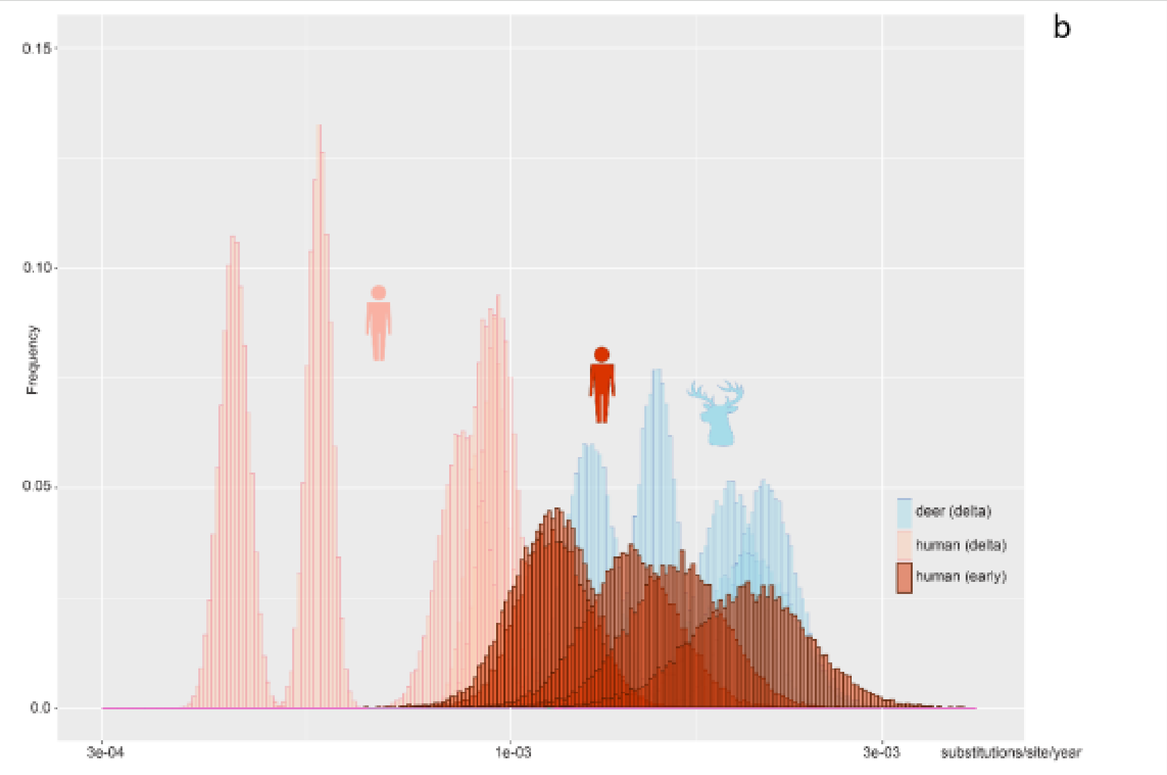


Figure S13. Evolutionary rates during early phase of SARS-CoV-2 outbreak in humans. The posterior distributions of evolutionary rates (substitutions per site per year) for five partitions of the SARSCoV-2 genome (ORF1a, ORF1b, ORF3 – ORF8, spike (S), and nucleocapsid (N) are presented for three datasets: variant in white-tailed deer (blue); variant in humans (pink); and 786 early strains of SARS-CoV- 2 in humans from Pekar et al.1 Alpha is presented above (n = 786) and delta below (n = 1094). Similar plots are available for variant data only (human vs. deer) for the alpha variant (Figure S11) and delta variant (Figure 4C). Mean values and 95% HPD are available for each partition and dataset in Table S5

Low early mutation rate

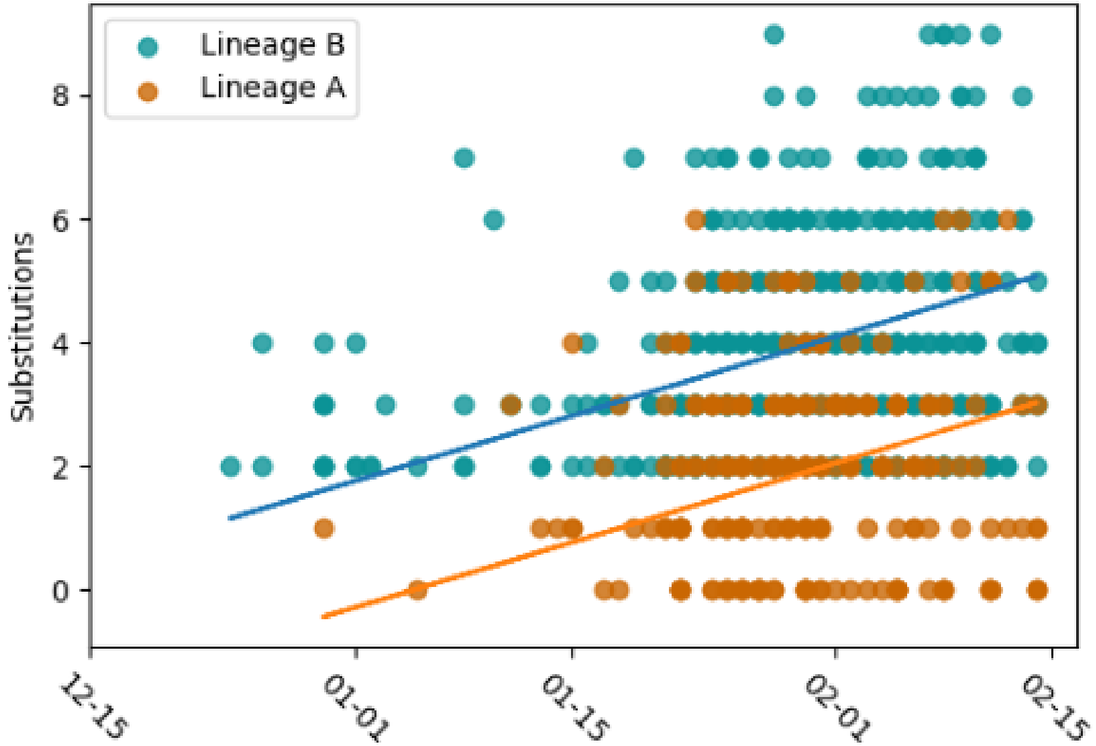
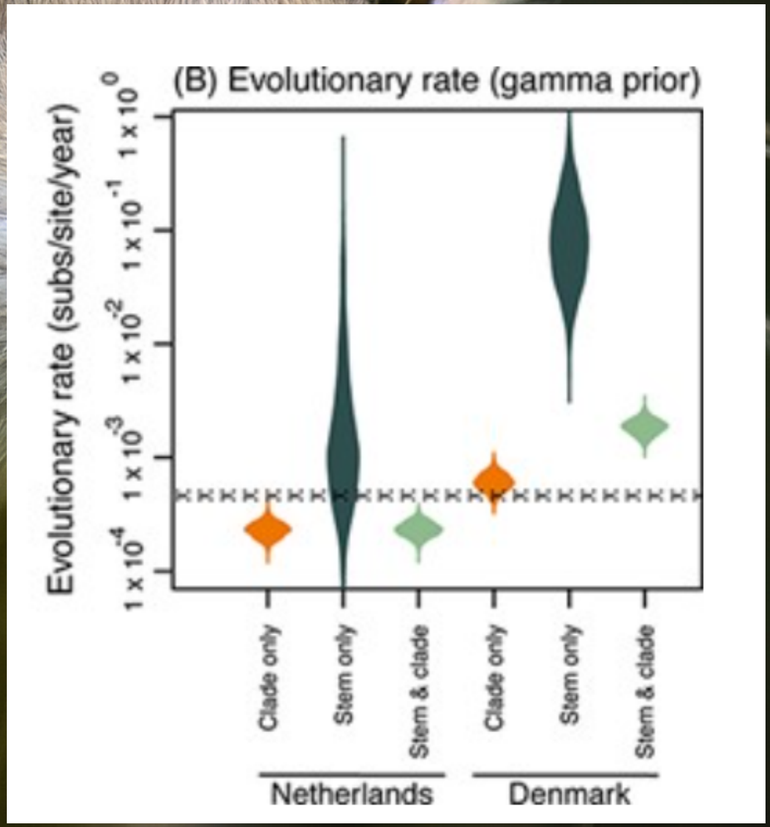


Figure S20. Substitution counts of SARS-CoV-2 genomes through 14 February 2020 from the root of the maximum likelihood tree when rooted on lineage A (Fig. S19). The plotted lines have a slope of 27.51 substitutions/year, are fit to their respective lineages, and are separated by 2.04 substitutions, showcasing the greater divergence of lineage B than lineage A when the tree is rooted on lineage A.

Pekar et al. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2

Comparing to SARS2 in Minks

In contrast, SARS2 did have an initial period of 4-13x faster mutation rate when jumping from humans to minks



https://pubmed.ncbi.nlm.nih.gov/36751428/

Low Early Mutation Rate is More Likely for a Lab Leak

Under Lab Leak: The low early mutation rate is expected under DEFUSE style research, which screened for RBDs that match human ACE2.

Under Zoonosis, two options:

- Could happen by chance, if a virus with an RBD perfect for human ACE2 (like BANAL-52) infects an intermediate host which then gets transported to Wuhan via wildlife trade
- Long cryptic transmission during the RBD adaptation - unlikely for a virus with severe symptoms

Summary - Best Explanations

- FCS - Ignored (despite no precedence in sarbecovirus)
- 12nt Clean Insert
  - If basing on frequency of large insertions, probably over 1000x.
    - Similar estimate if looking at the coincidence of the only long insertion happening to be in the most important feature of the virus.
  - Best explanation is there is some unknown reason why an FCS specifically should emerge with a long insertion. Years of discussions have yielded no such suggestion. Estimated at 50x, Low of 20x.
  - CGGCGG - Best explanation is the first CGG is random, and the second was a duplication event (more likely given the insert). 10x.
- Leading Proline
  - Could be inspired by MERS or the PAA sequence in bat coronaviruses.
- Why insert RRA and not RAR (for a more canonical RARR)?
  - Others have done it and they could be testing PAA -> PRA -> PRRA.
- In any case, hard to say any lab action is unreasonable, as it’s hard to cover all the possibilities. (See further discussion in the response deck)

Summary - Updated Probabilities

Genetics			
12 nt clean insert from unknown sour	20	50	100
CGGCGG			
Zoonosis	0.0026	0.0026	0.0026
Lab Leak	0.02777777778	0.02777777778	0.04
Ratio	10.68376068	10.68376068	15.38461538
PRRAR and "out of frame"	0.3	0.4	0.5
Total Genetics	64.1025641	213.6752137	769.2307692
	160.5321377	40117.94037	2496786.051
Updated	**99.48%**	**100.00%**	**100.00%**

Weighted: 99.9%