SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes

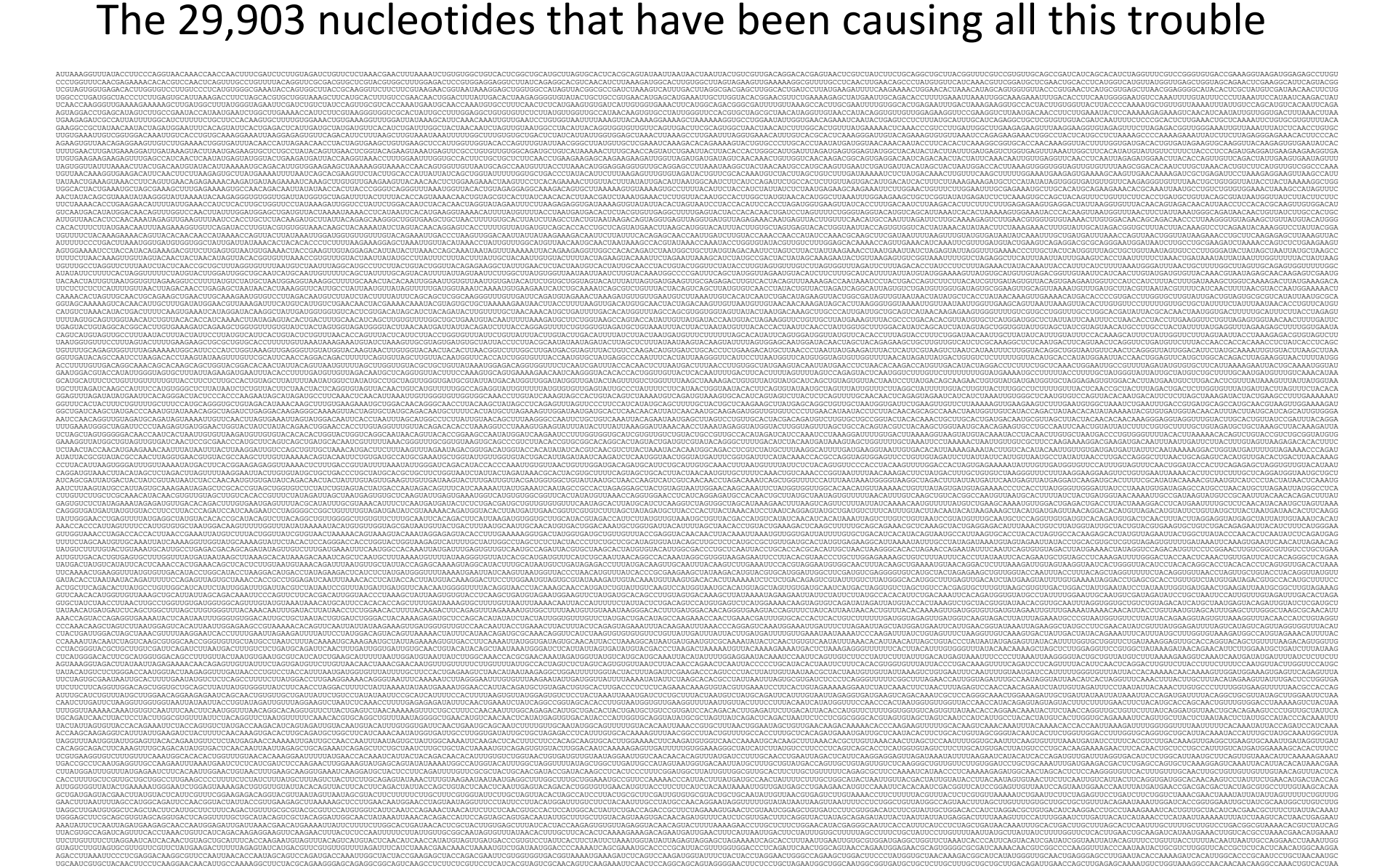
Irwin Jungreis, Rachel Sealfon,
and Manolis Kellis

2021-01-14



Massachusetts
Institute of
Technology

The 29,903 nucleotides that have been causing all this trouble



What can evolution tell us about this?

Outline

- SARS-CoV-2 basics
- Evolutionary signatures
- SARS-CoV-2 gene content
- Mutation impact

Outline

- SARS-CoV-2 basics
- Evolutionary signatures
- SARS-CoV-2 gene content
- Mutation impact

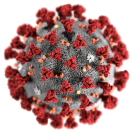
Molecular Biology Review

Central Dogma of Molecular Biology

DNA (long term information storage)
Four nucleotides A, C, G, T

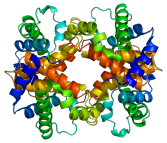
transcription

mRNA (short term intermediary)
Four nucleotides A, C, G, U (U≈T)



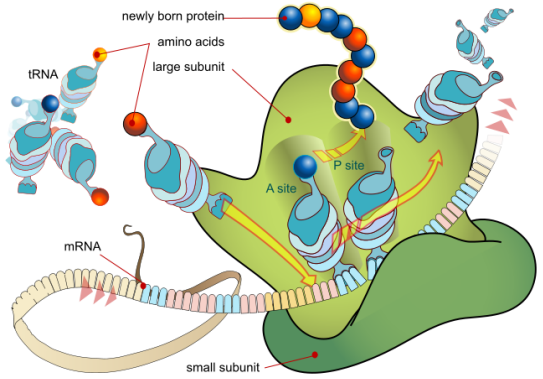
ssRNA virus pretends to be mRNA

translation



protein (does the work)
20 amino acids (A, R, D, …)

Translation



- Ribosome finds AUG (usually first) "start codon"
- Moves along mRNA 3 nucleotides at a time
- 3 nucleotides = 1 codon → 1 amino acid
- Stop at UAA, UAG, or UGA "stop codon"
- One mRNA -> one kind of protein
- Interval from AUG to stop codon with no stops in between = Open Reading Frame = ORF

Genetic Code

First Base	Second Base U		Second Base C		Second Base A		Second Base G		Third Base
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP CODON	UGA	STOP CODON	A
	UUG		UCG		UAG		UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met (START CODON)	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

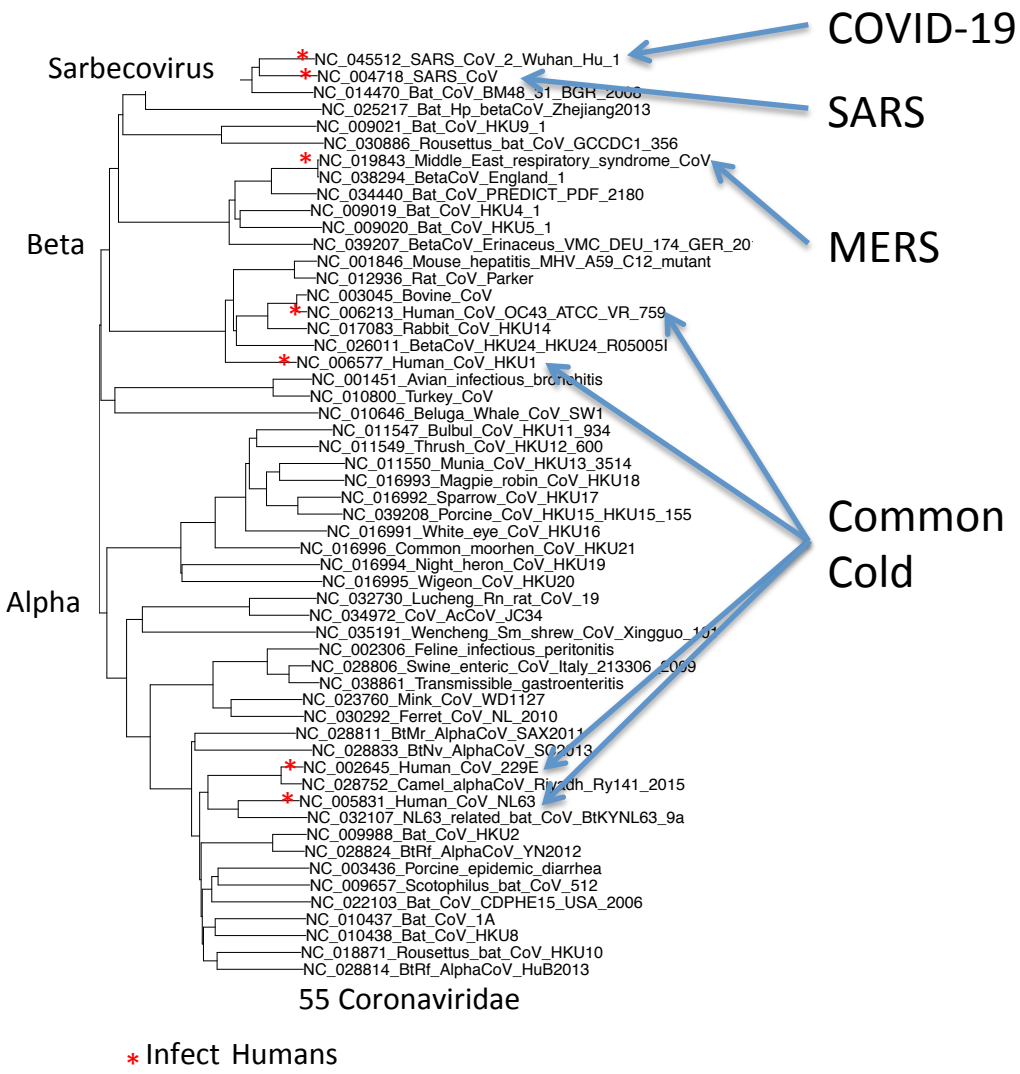
Synonymous Mutations

- 64 codons -> 20 amino acids
- Genetic code is redundant
- Different codons can encode the same amino acid
- ~~Silent~~ Synonymous mutations preserve amino acid

SARS-CoV-2 Phylogeny

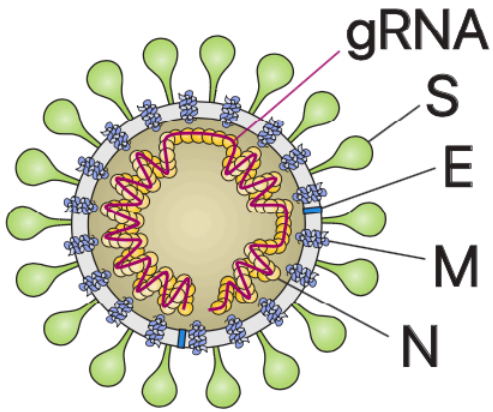
SARS-CoV-2: positive-stranded RNA virus responsible for COVID-19 pandemic

Realm:	Riboviria
Kingdom:	Orthornavirae
Phylum:	Pisuviricota
Class:	Pisoniviricetes
Order:	Nidovirales
Family:	Coronaviridae
Genus:	Betacoronavirus
Subgenus:	**Sarbecovirus**
Species:	*Severe acute respiratory syndrome-related coronavirus*
Strain:	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

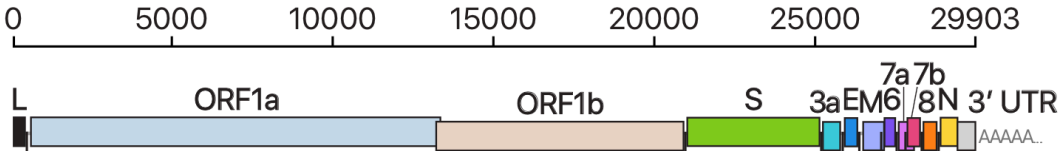


Coronaviruses

- Positive strand RNA virus
- Enclosed in roughly spherical lipid bilayer envelope
- Surface projections “spikes”
- Four structural proteins
  - spike protein (S),
    - S1: viral attachment to host-cell ACE2 receptor
    - S2: membrane fusion, viral entry
  - envelope protein (E),
  - membrane protein (M),
  - nucleocapsid protein (N)
  - Common to all coronaviruses
  - These are packaged with genomic RNA to make new virions

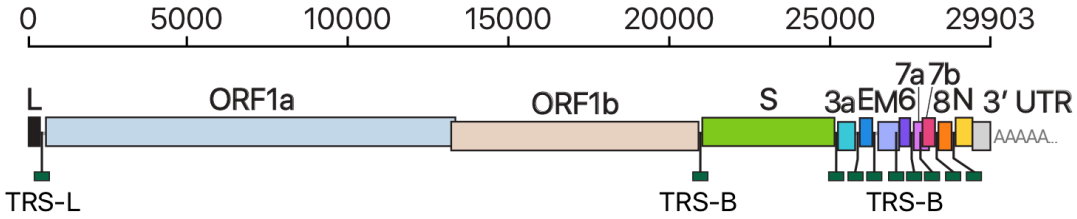


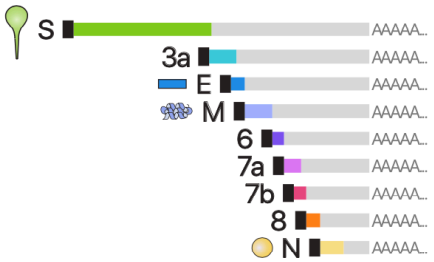
- Genome is ~30,000 nucleotides, very long for an RNA virus
- Genome looks like an mRNA
  - 5’ cap
  - Poly-A tail
- Encodes around 30 proteins
- Questions:
  - How does one “mRNA” encode so many proteins (usually only 1)?
  - How does viral RNA genome get copied (usually RNA is copied from DNA)?



Kim et al. bioRxiv 2020 https://doi.org/10.1101/2020.03.12.988865

Expression and replication





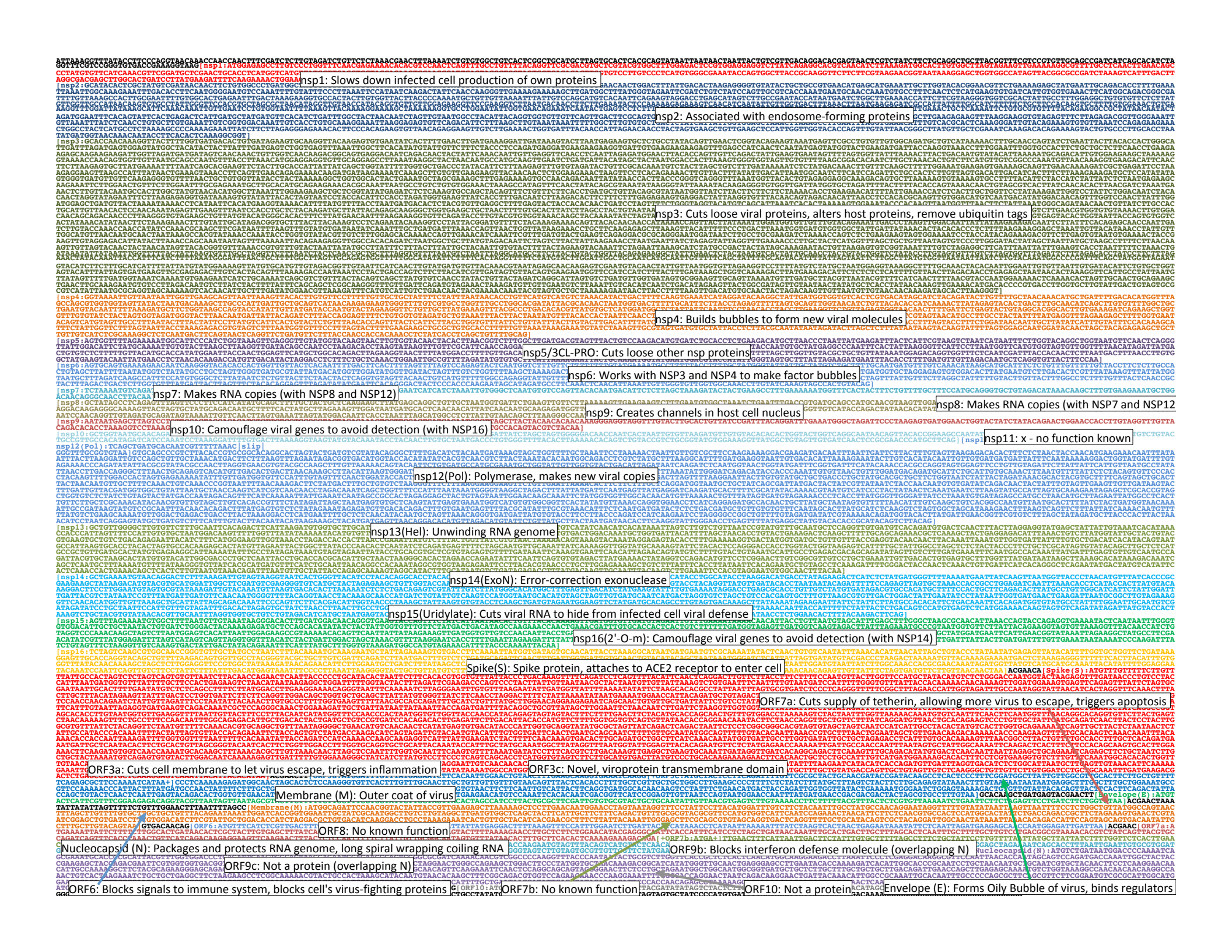
- First, ORF1a is translated
- Cellular proteins cleave ORF1a into several proteins
  - One of them is $3CL^{pro}$ which cleaves more.
  - ORF1a is cleaved into 11 "non-structural proteins" (nsp1-nsp11)
- Sometimes, programmed frameshift occurs during ORF1a translation
- ORF1a + ORF1b is translated, cleaved -> nsp1-10 + nsp12-16
- nsp12 = RNA-dependent RNA polymerase (RdRp) which copies RNA
- nsp14 = ExoN does proofreading → low coronavirus mutation rate
- Other nsps involved in:
  - helicase
  - assembly
  - suppression of host cell and immune system function

Programmed Frameshift

…UUU UUA AAC GGG UUU GCG GUG UAA (end of ORF1a)
AAA CGG GUU UGC GGT GUA AGU GCA…

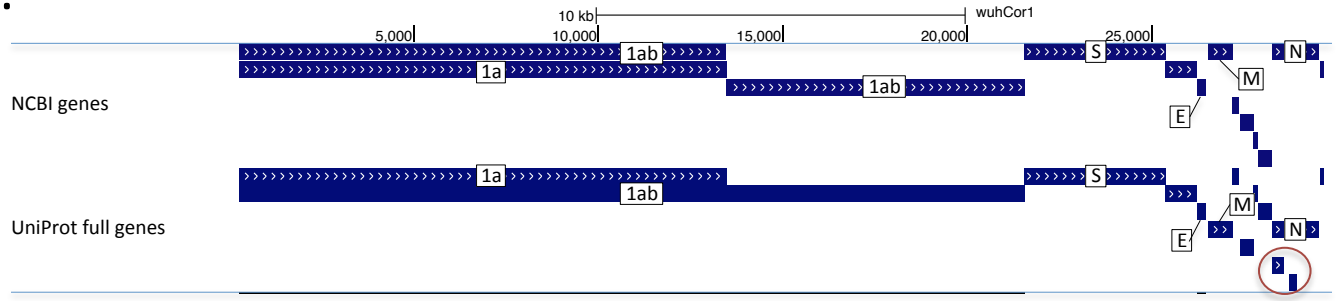
- Bunched RNA slows down ribosome
- Ribosome sometimes moves back one nucleotide
- Continues translating in a different reading frame
- Bypasses the stop codon
- Five codons translated in two different reading frames

- How are proteins in the rest of genome translated?
- RdRp transcribes genomic RNA to negative strand RNAs
  - Some are full length. These are then transcribed back to positive strand RNA to replicate genomic RNA
  - Others stop transcription at transcription-regulating sequences (TRS) and join to TRS in leader for leader-body fusion
    - Transcribed back to a nested set of positive strand **subgenomic RNAs** (sgRNAs) for translation of other genes
- subgenomic RNAs are translated to create:
  - Structural proteins: spike protein (S), envelope protein (E), membrane protein (M), and nucleocapsid protein (N)
  - Also several "accessory proteins" with less well characterized function (ORFs 3a, 6, etc.)
- One other bit of coronavirus biology:
  - Recombination is common in coronaviruses when two viruses infect the same cell


nsp1: Slows down infected cell production of own proteins
nsp2: Associated with endosome-forming proteins
nsp3: Cuts loose viral proteins, alters host proteins, remove ubiquitin tags
nsp4: Builds bubbles to form new viral molecules
nsp5/3CL-PRO: Cuts loose other nsp proteins
nsp6: Works with NSP3 and NSP4 to make factor bubbles
nsp7: Makes RNA copies (with NSP8 and NSP12)
nsp8: Makes RNA copies (with NSP7 and NSP12
nsp9: Creates channels in host cell nucleus
nsp10: Camouflage viral genes to avoid detection (with NSP16)
nsp11: x - no function known
nsp12(Pol): Polymerase, makes new viral copies
nsp13(Hel): Unwinding RNA genome
nsp14(ExoN): Error-correction exonuclease
nsp15(Uridylate): Cuts viral RNA to hide from infected cell viral defense
nsp16(2'-O-m): Camouflage viral genes to avoid detection (with NSP14)
Spike(S): Spike protein, attaches to ACE2 receptor to enter cell
ORF7a: Cuts supply of tetherin, allowing more virus to escape, triggers apoptosis
ORF3a: Cuts cell membrane to let virus escape, triggers inflammation
ORF3c: Novel, viroprotein transmembrane domain
Membrane (M): Outer coat of virus
Nucleocapsid (N): Packages and protects RNA genome, long spiral wrapping coiling RNA
ORF9b: Blocks interferon defense molecule (overlapping N)
ORF9c: Not a protein (overlapping N)
ORF6: Blocks signals to immune system, blocks cell's virus-fighting proteins
ORF7b: No known function
ORF10: Not a protein
Envelope (E): Forms Oily Bubble of virus, binds regulators
ORF8: No known function

Improving SARS-CoV-2 gene annotations

- But, wait. It's a tiny viral genome. Surely all the genes are known. Is there room for improvement?
- Yes!



- 5 named genes in all coronaviruses
- SARS-COV-2 (NC_045512) accessory genes
  - NCBI annotations has 6 unnamed genes:
    - ORF3a, ORF6, ORF7a, ORF7b, ORF8, ORF10
  - UniProt annotations has 2 more:
    - ORF9b, ORF9c
    - Both overlap nucleocapsid in a different frame
  - Other proposed genes
    - ORF3b, ORF3c, ORF3d, ORF3d-2, ORF2b, …

No agreement on gene list or names:

- Wu et al. 2020. Nature.
  - UniProt + ORF3b
  - Refers to 9b as 9a; 9c as 9b
  - Most recent GenBank record lacks 7b, 9b, 9c
- Lu et al. 2020. Lancet.
  - Missing 10 but includes 9c.
  - Refers to 3a as 3, 6 as 7, 7a as 8, 7b as 9, 8 as 10b, 9b as 13.
- …

Techniques for distinguishing protein-coding regions…

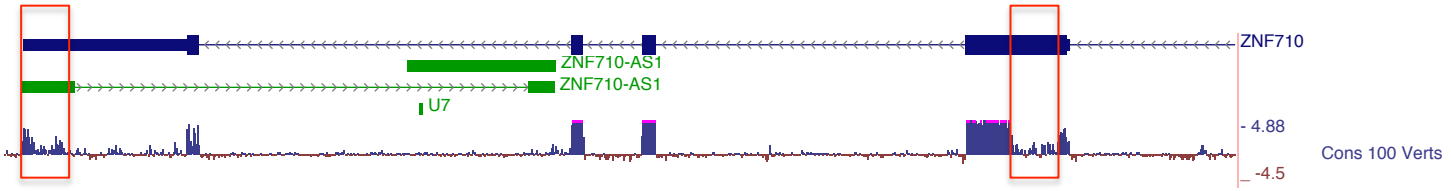
- Nanopore direct RNA sequences to find sgRNAs
  - Evidence of transcription but not translation
- Ribosome profiling (Ribo-Seq) detects RNA being translated
  - Cycloheximide immobilizes ribosomes
  - Use ribonucleases to digest RNA
  - RNA inside ribosome is protected and not digested
  - Digest proteins of ribosome
  - Sequence the RNA fragments that were protected by the ribosome
  - Ribosome binding does not guarantee stable protein creation
- Proteomics / Mass spectrometry
  - Imperfect sensitivity
  - Does not indicate protein is functional
- Comparative genomics/Evolutionary signatures
  - Evidence of *functional* proteins
  - But, cannot detect recently evolved proteins or recent loss
  - Doesn't tell us when translation occurs or why

Outline

- SARS-CoV-2 basics
- Evolutionary signatures
- SARS-CoV-2 gene content
- Mutation impact

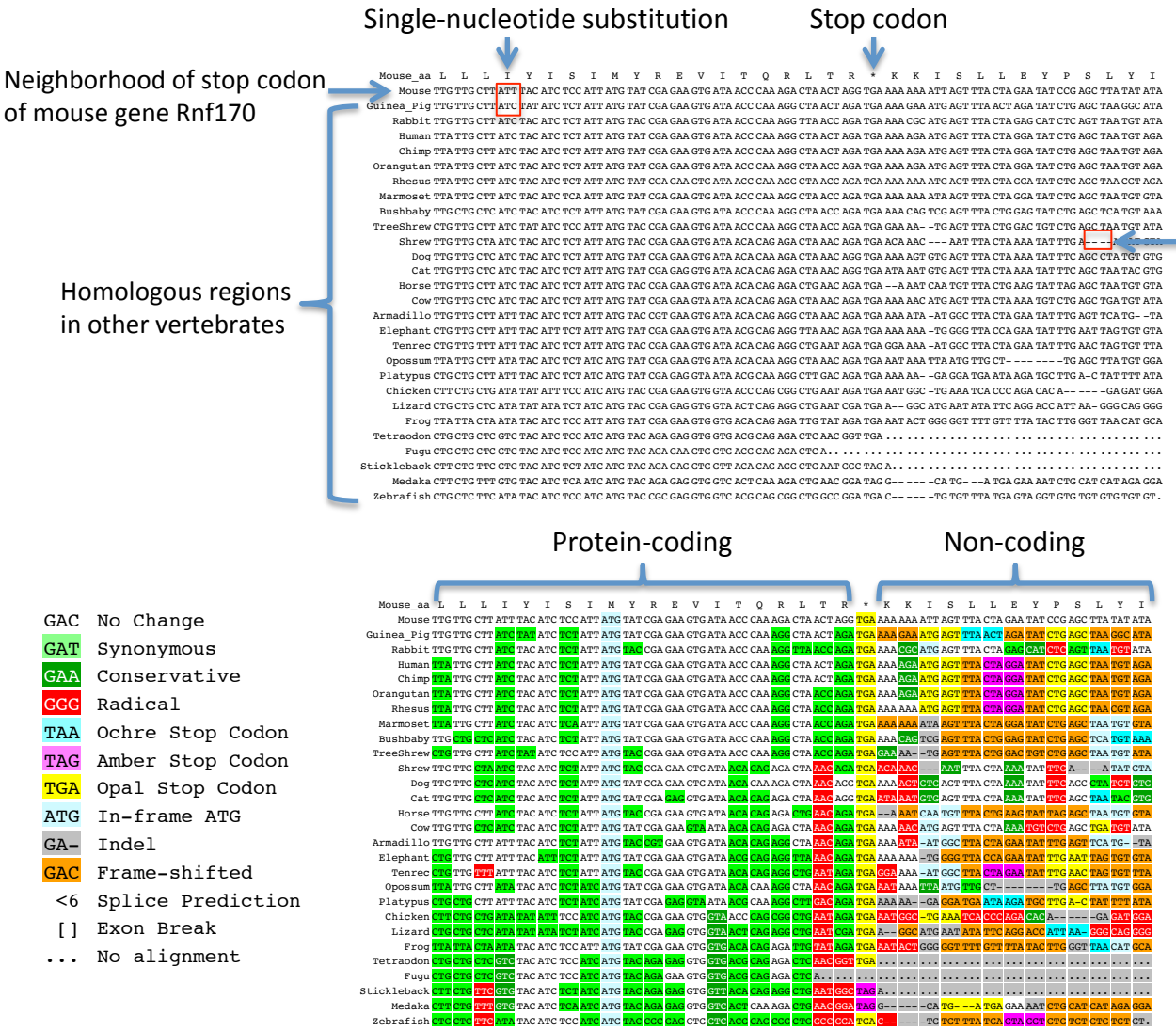
Evolutionary Signatures

- Basic idea: Conserved = Functional
- Step 1: Align homologous regions of genomes
  (homologous = descended from same region of ancestor genome)
- Step 2: See which nucleotides changed less than expected
  - Changes to these nucleotides must break something
  - Organisms with changes in those nucleotides were less likely to survive



- Example: human gene ZNF710
- Blue rectangles are exons
  - Thick parts are protein-coding open reading frame
- Green rectangles are non-protein-coding gene
- Bottom graph: how well each nucleotide is conserved among 100 vertebrates
- Conservation matches protein-coding pretty well, but not perfect
  - Some protein-coding regions are not well conserved
  - Some well conserved regions are not protein-coding
- Can we find signature that is specifically for protein-coding?

Is there an evolutionary signature specific to protein-coding?



Codon coloring by
CodAlignView
the Codon Alignment Viewer

- Protein-coding region (before stop codon)
  - Most substitutions are synonymous (light green)
  - Avoids frame-shifting insertions and deletions (orange)
- Non-coding region (after stop codon)
  - Many amino acid changing substitutions (red and dark green)
  - Frame-shifting insertions and deletions (orange)
  - Stop codons (yellow, cyan, magenta)

Coding and non-coding models of evolution

- Given a species tree and an alignment of homologous sequences, distinguish protein-coding regions
- dN/dS test compares rates of synonymous and non-synonymous changes
- We can do better:
  - Amino acid changes that preserve biochemical properties are more likely
  - Among synonymous codons, some are more common
- Step 1: Make protein-coding and non-coding models of of codon evolution:
  - Rate at which any codon changes to any other
  - Likelihood of any particular codon in the ancestor
  - Train on all known protein-coding or non-coding codons in human genome
  - Does not explicitly consider synonymous vs. non-synonymous
  - Implicitly distinguishes conservative vs. radical amino acid changes, codon frequencies, …
- Step 2: Compute likelihood of an alignment under each models
  - Log likelihood ratio tells us if alignment is more likely under coding or non-coding model of evolution
  - PhyloCSF – Phylogenetic Codon Substitution Frequencies

Symmetric 64x64 matrix with substitution rate for every pair of codons

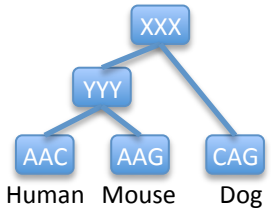
	AAA	AAC	AAG	…
AAA		0.857	21.4	
AAC	0.857		0.243	
AAG	21.4	0.243		
…				

64-vector with ancestral codon frequencies

AAA	AAC	AAG	AAT	…
0.018	0.024	0.037	0.023	

Example

```
Human AAC
Mouse AAG
Dog   CAG
```

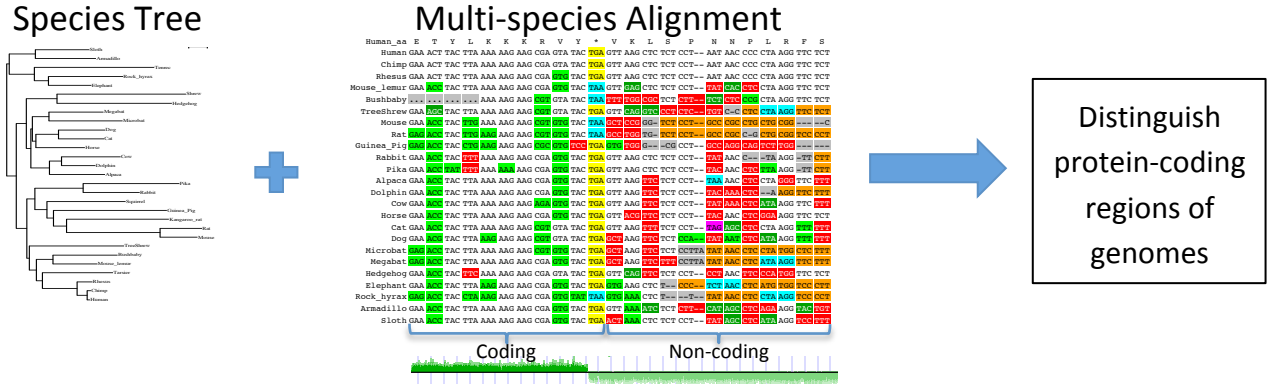


- We know codons at leaves
- We don't know codons at ancestral nodes.

$$Likelihood = \sum_{XXX} Freq(XXX) \sum_{YYY} \Pr ob_1(XXX, YYY) \Pr ob_2(YYY, AAC) \Pr ob_3(YYY, AAG) \Pr ob_4(XXX, CAG)$$

- Where:
  - Sums are over all possible codons XXX and YYY
  - Freq(XXX) is the ancestral frequency of codon XXX
  - $Prob_i$(UUU,VVV) is the probability that UUU changes to VVV in the i'th branch
    - (Depends on branch length)
- Compute using Felsenstein algorithm
  - Dynamic programming
  - Like Forward algorithm for HMMs but for trees

PhyloCSF – Phylogenetic Codon Substitution Frequencies



PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions

Michael F. Lin[1,2,*], Irwin Jungreis[1,2] and Manolis Kellis[1,2,*]

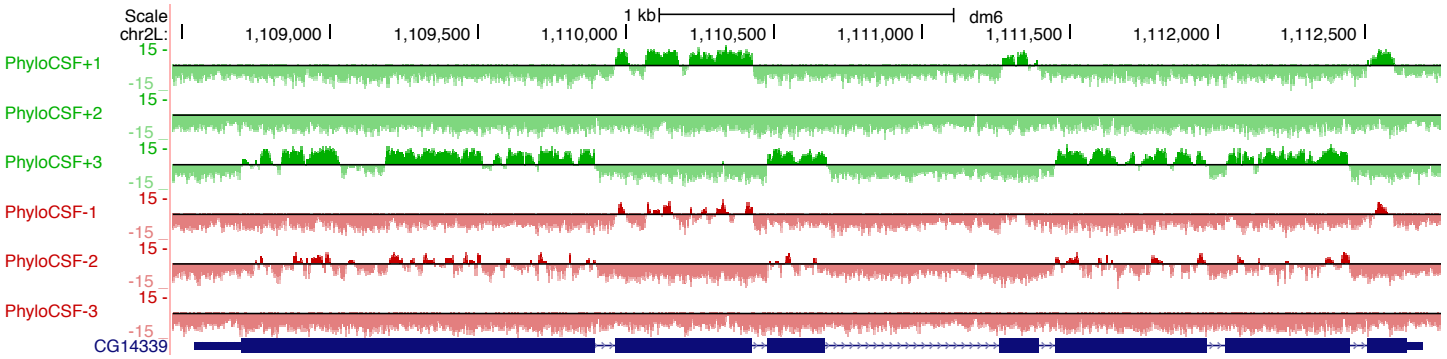
Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci

Jonathan M. Mudge,[1,9] Irwin Jungreis,[2,3,9] Toby Hunt,[1] Jose Manuel Gonzalez,[1] James C. Wright,[4] Mike Kay,[1] Claire Davidson,[1] Stephen Fitzgerald,[5] Ruth Seal,[1,6] Susan Tweedie,[1] Liang He,[2,3] Robert M. Waterhouse,[7,8] Yue Li,[2,3] Elspeth Bruford,[1,6] Jyoti S. Choudhary,[4] Adam Frankish,[1] and Manolis Kellis[2,3]

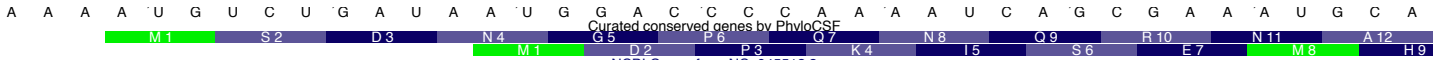
Evolutionary Dynamics of Abundant Stop Codon Readthrough

Irwin Jungreis,*[,1,1,2] Clara S. Chan,[1,1,2] Robert M. Waterhouse,[1,2,3,4] Gabriel Fields,[5] Michael F. Lin,[6] and Manolis Kellis*[1,2]

- Depends on the *kind* of substitutions rather than the *number*
  - Neutral to overall level of nucleotide conservation
- For SARS-CoV-2 we used human-trained matrices
  - Not perfect, but result is not very sensitive to species
- Score individual Open Reading Frames, or
- Visualize protein-coding constraint in genome browser
  - Browser tracks score **every codon in 6 reading frames**
    - Smoothed with HMM
  - Public "PhyloCSF" track hub in UCSC or Ensembl Genome browsers



FRESCo: Finding Regions of Excess Synonymous Constraint



- Viral (and other) genomes sometimes have overlapping genes
- These translate the same region in two different reading frames
- This imposes constraint different from PhyloCSF model
- Conservation of one amino acid sequence suppresses synonymous substitutions in other frame
- Lower synonymous substitution rate can also result from other overlapping function:
  - RNA structures, binding sites, translation regulation, …


Sealfon et al. Genome Biology (2015) 16:38
DOI 10.1186/s13059-015-0603-7

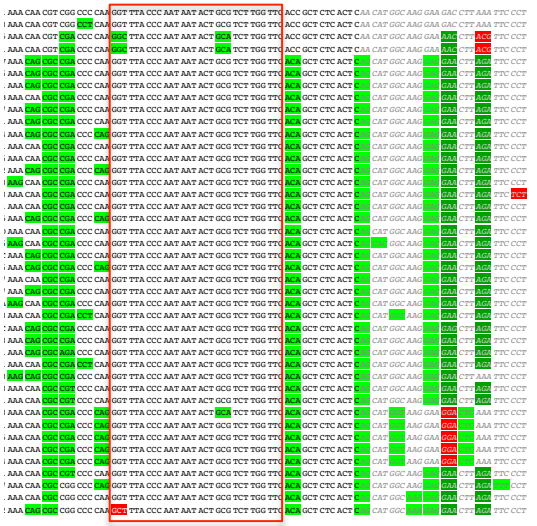
Genome Biology

METHOD Open Access

FRESCo: finding regions of excess synonymous constraint in diverse viruses

Rachel S Sealfon[1,2*], Michael F Lin[3], Irwin Jungreis[1,2], Maxim Y Wolf[1,2], Manolis Kellis[1,2] and Pardis C Sabeti[2,4*]

Synonymous Constraint



- FRESCO finds regions of significantly fewer synonymous substitutions
  - Designed specifically for individual viral genes with limited training data
  - Relative to average per gene
  - Input is alignment and tree topology
  - Estimates maximum likelihood branch lengths from alignment
  - Fit Muse and Gaut model of codon evolution to whole gene

    Probability of single-nucleotide substitution from codon i to codon j is:

    $$q_{ij} = \pi_x^P \times (\kappa \text{ if transition o.w. } 1) \times (R \text{ if nonsynonymous o.w. } 1)$$

    where π is frequency of target base at the changed codon position
  - Null model allow nonsynonymous rate different from gene rate
  - Alternate model allows different nonsynonymous and synonymous rates
  - Likelihood ratio test finds windows where alternate model is significantly better
- Use on 9-codon windows to find Synonynous Constraint Elements (SCEs)
- Use on 1-codon windows to evaluate Single Nucleotide Variants (SNVs)

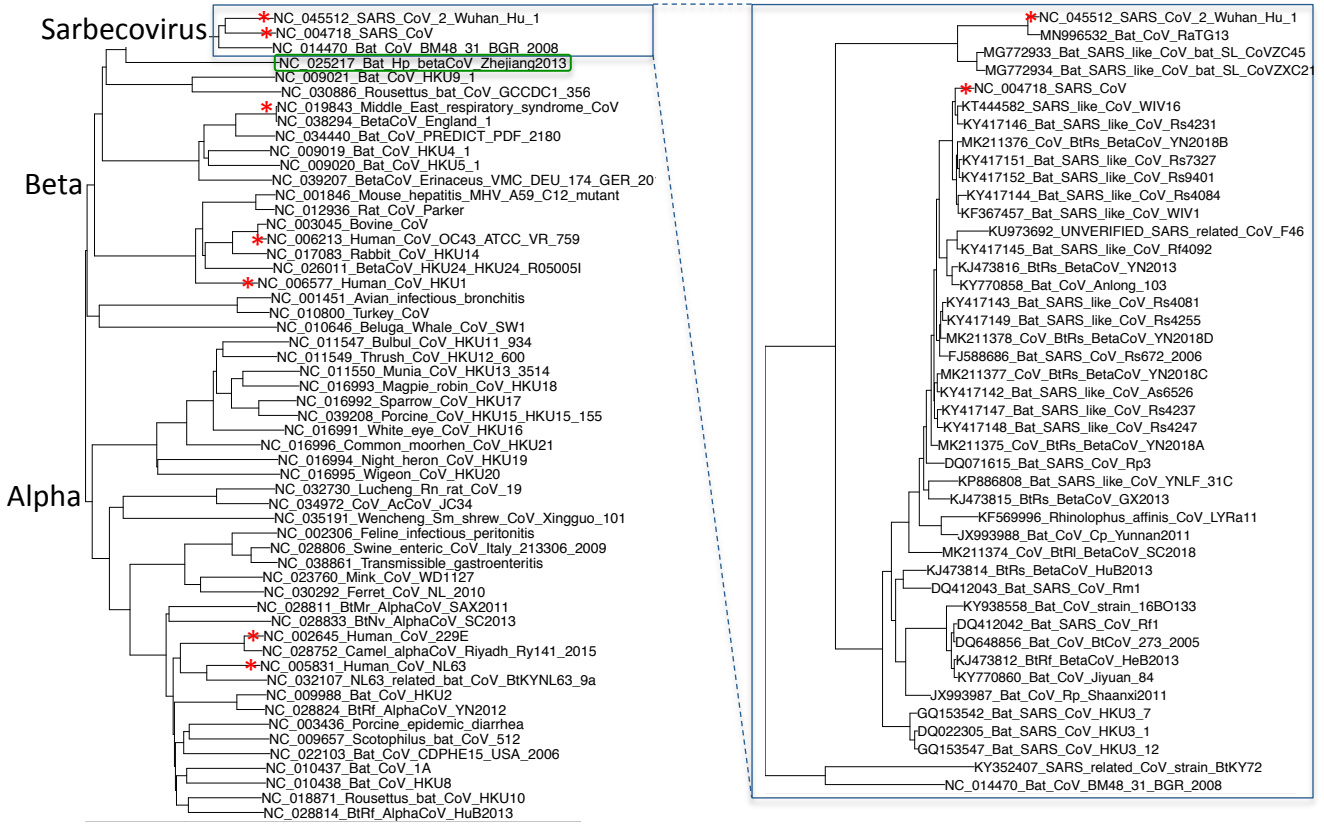
Outline

- SARS-CoV-2 basics
- Evolutionary signatures
- SARS-CoV-2 gene content
- Mutation impact

Choosing an appropriate set of genomes

- Need set of genomes that:
  - Have enough substitutions to be informative
  - Are close enough to have conserved function
    - Much less than 1 substitution per branch
  - No near duplicates in which not enough time to suppress slightly deleterious mutations
- Goldilocks clade:
  - SARS-CoV-2 sequences are too close
  - Betacoronaviruses are too far
  - Sarbecovirus is just right
- 44 Sarbecovirus complete genomes from NCBI
  - About 3 substitutions per neutral site
  - Comparable to 12 flies and 29 mammals
  - Actually, all are in the same species: *SARSr-coronavirus*
    - Don’t know how results will generalize to rest of Sarbecovirus subgenus

Phylogenetic Trees



55 Coronaviridae

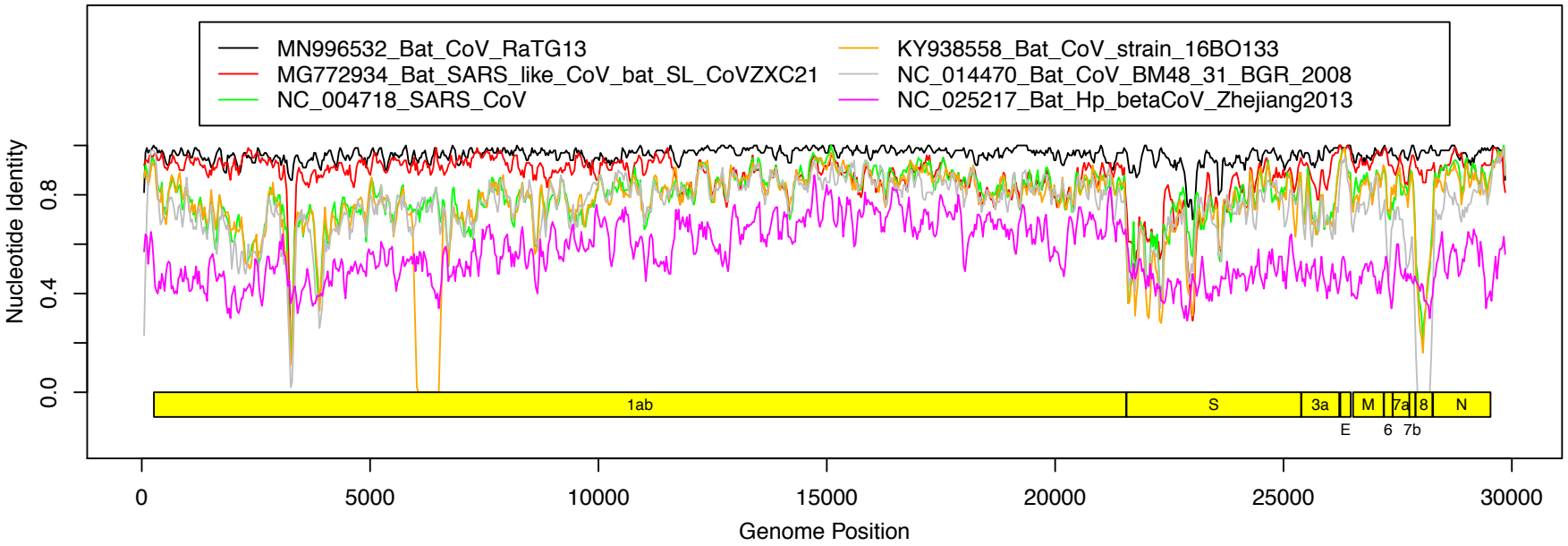
44 Sarbecoviruses
SARS-CoV-2 + SARS-CoV + 42 bat viruses

* Infect Humans

- Even Bat Hp betacoronavirus is too far
  - No homology for ORF 6, 7a, 7b, 8

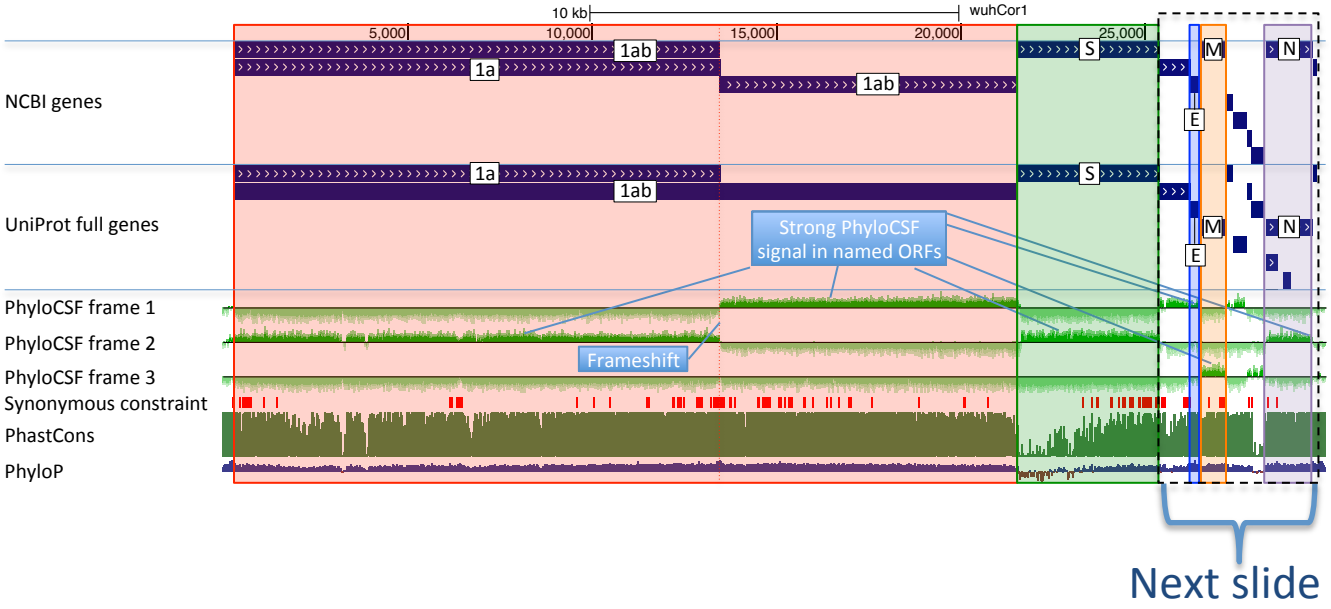
Note: Trees are based on whole-genome alignments. Some regions could be different.

How much do these genomes diverge?



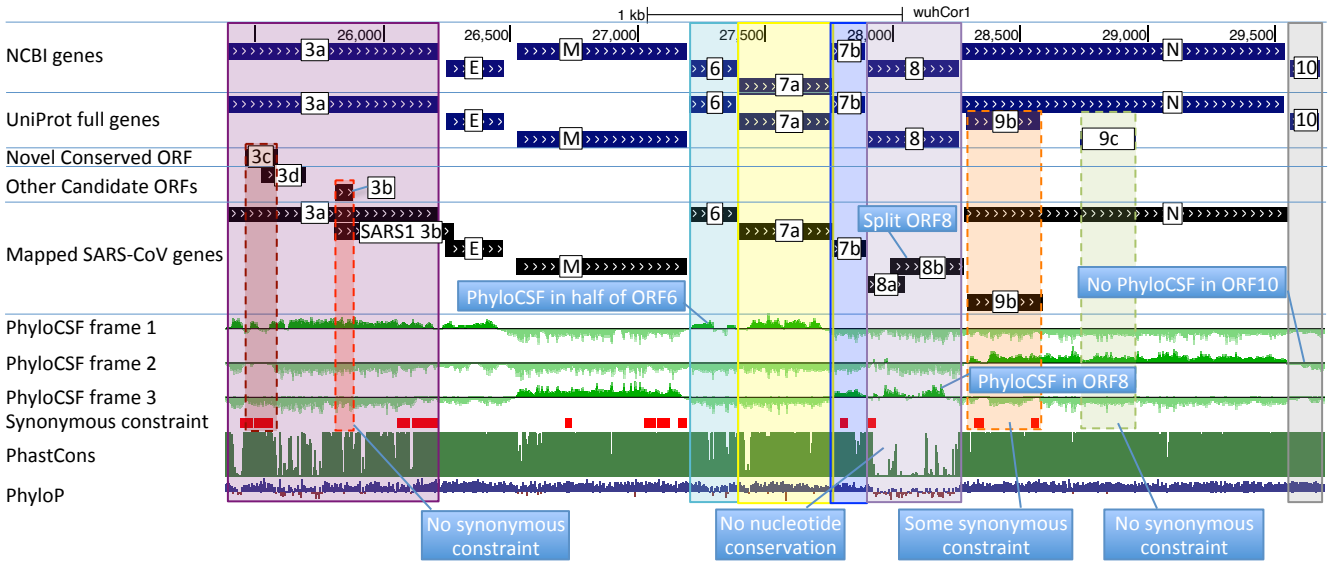
- Fraction of nucleotides that differ from SARS-CoV-2
  - 100-nucleotide windows with 30-nucleotide overlap
- RaTG13 is closest, differs in 4% of positions
- SARS-CoV differs in about 21%
- Outgroup betaCoV-Zhejiang2013 differs in more than half

Strong PhyloCSF signal in named genes



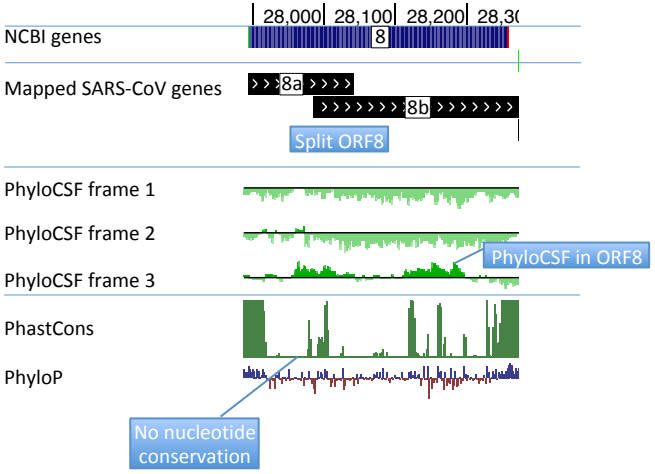
- Strong PhyloCSF Signal in correct frame for all named genes (ORF1ab, S, E, M, N)
- Change of frame at frameshift site
- Signal of fast evolution at start of S could be due to wrong tree

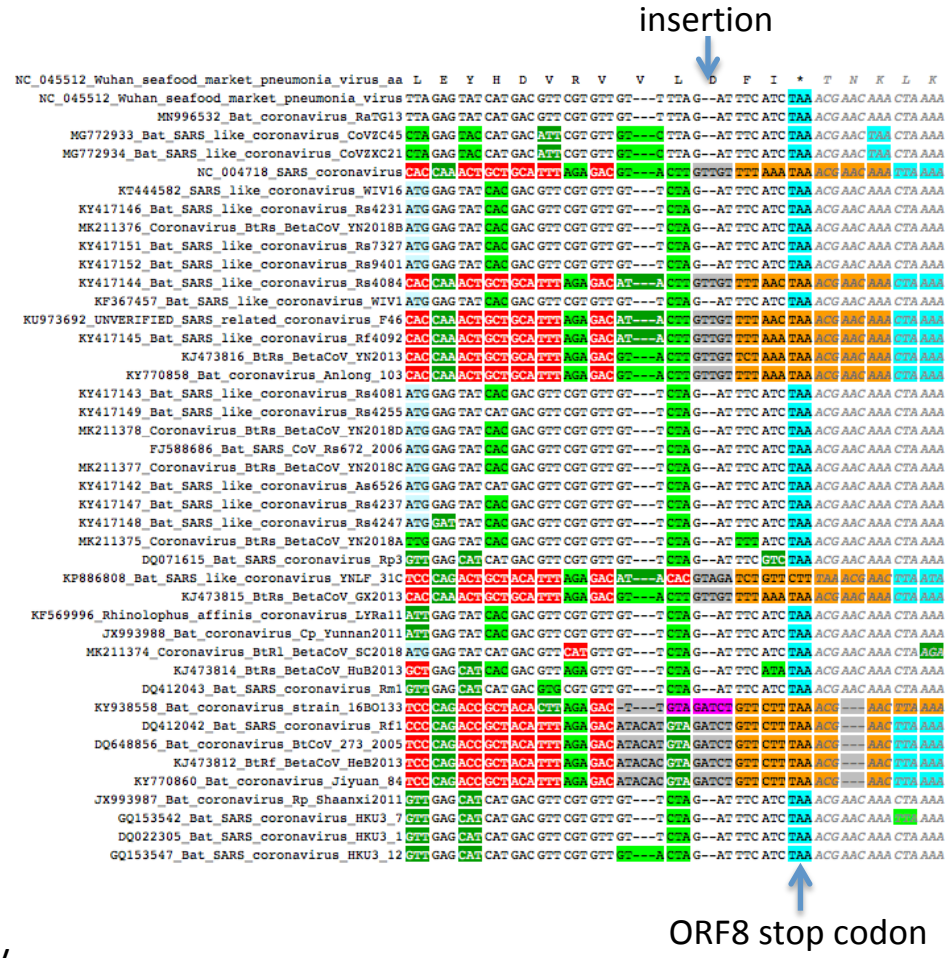
PhyloCSF signal for accessory genes



- Clear PhyloCSF signal for ORFs 3a, 6, 7a, 7b
- Overview of other cases:
  - Adequate PhyloCSF for ORF8 despite no nucleotide conservation
    - Frameshifting indel in ORF8 in some isolates of SARS-CoV makes fragments 8a/8b
  - No evolutionary signature for ORF10 despite high nucleotide conservation
  - No evolutionary signature for dual coding region ORF9c
  - Some evolutionary signature for dual coding region ORF9b
  - No evolutionary signature for ORF3b and ORF3d
  - ORF3c novel conserved protein-coding gene

ORF8 is coding despite low nucleotide-level conservation

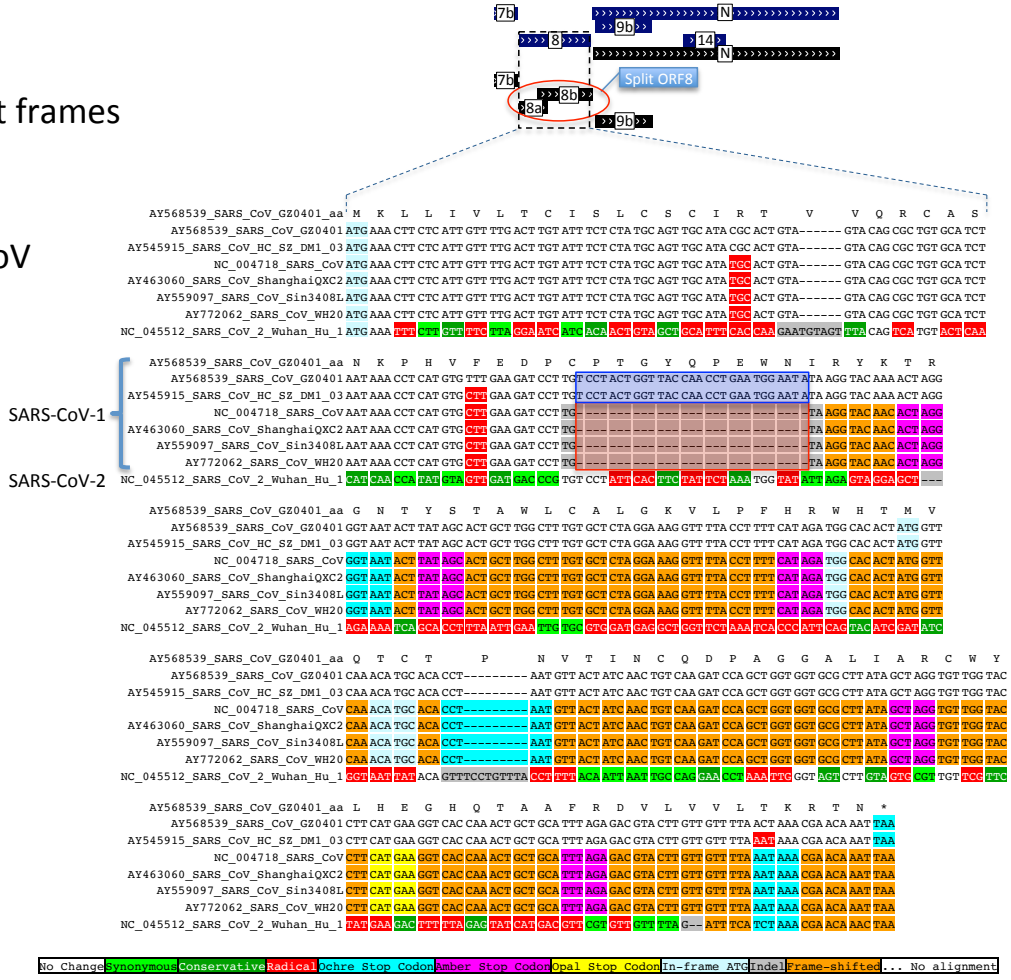




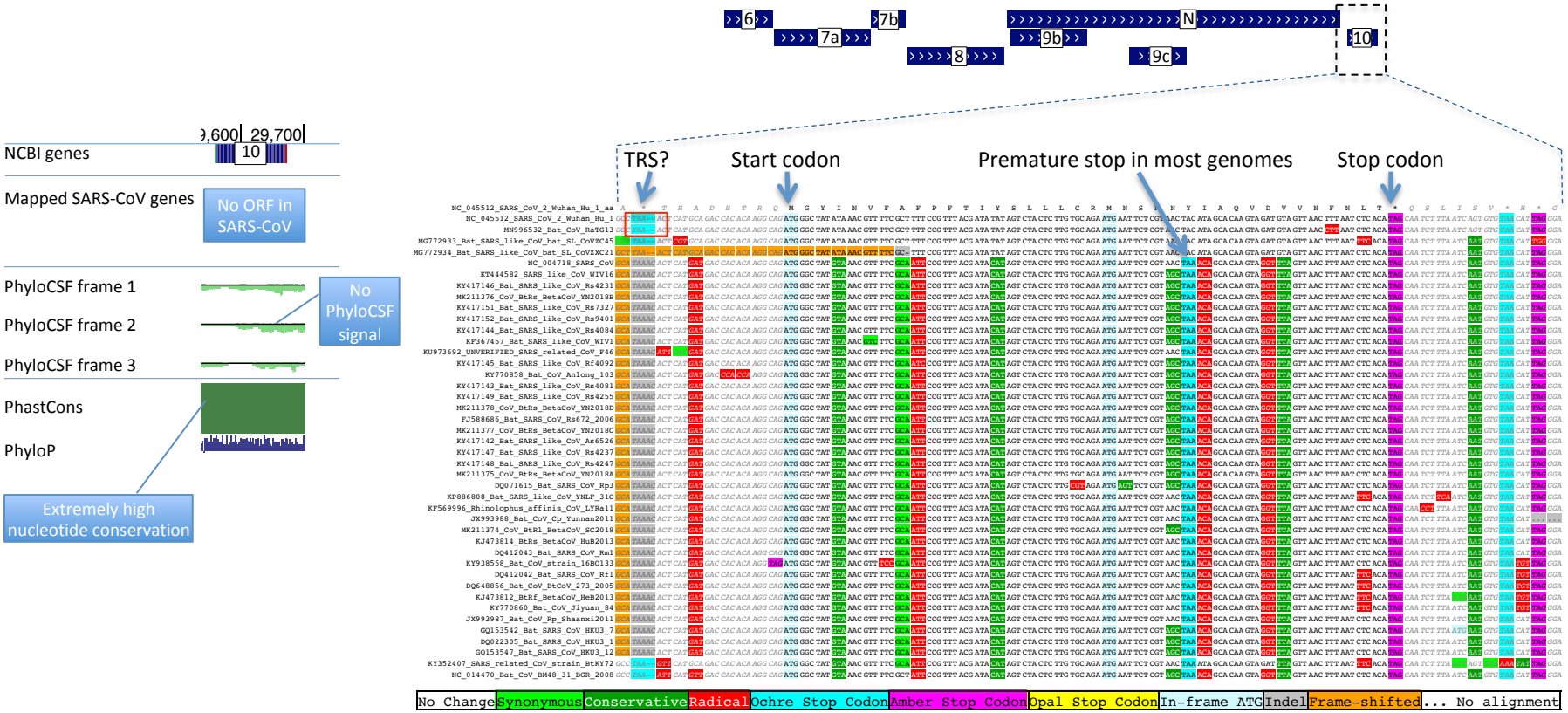
- Note that ORF8 is split in SARS-CoV
- Adequate PhyloCSF signal to show it is a conserved protein-coding region
- Almost no nucleotide-level conservation detected by PhyloP and PhastCons
- Insertion near end shows that local tree is different from species tree
  - Insertion is scattered through tree instead of in neighboring species
- PhyloP and PhastCons are very sensitive to tree, but PhyloCSF is not
- ORF8 has high substitution rate even when using the local tree

ORF8 was pseudogenized within SARS-CoV population

- SARS-CoV has two fragments of ORF8 in different frames
  - ORF8a is 39 amino acids and shares initial ATG
  - ORF8b is 84 amino acids and shares most of 3’ end
- Alignment of SARS-CoV-2 to 7 isolates of SARS-CoV
  - Indel is present in only 5 SARS-CoV lineages
  - Arose within SARS-CoV population



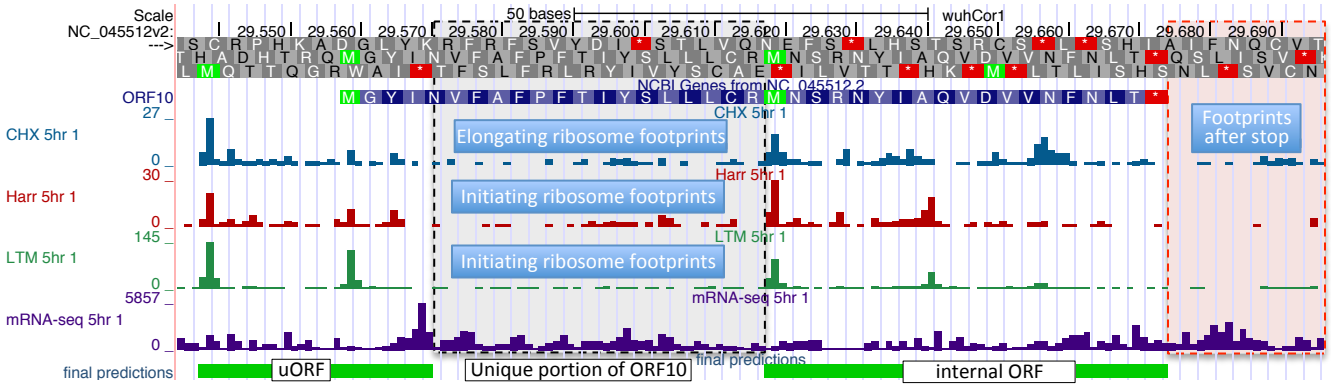
ORF10 is not a conserved protein-coding ORF



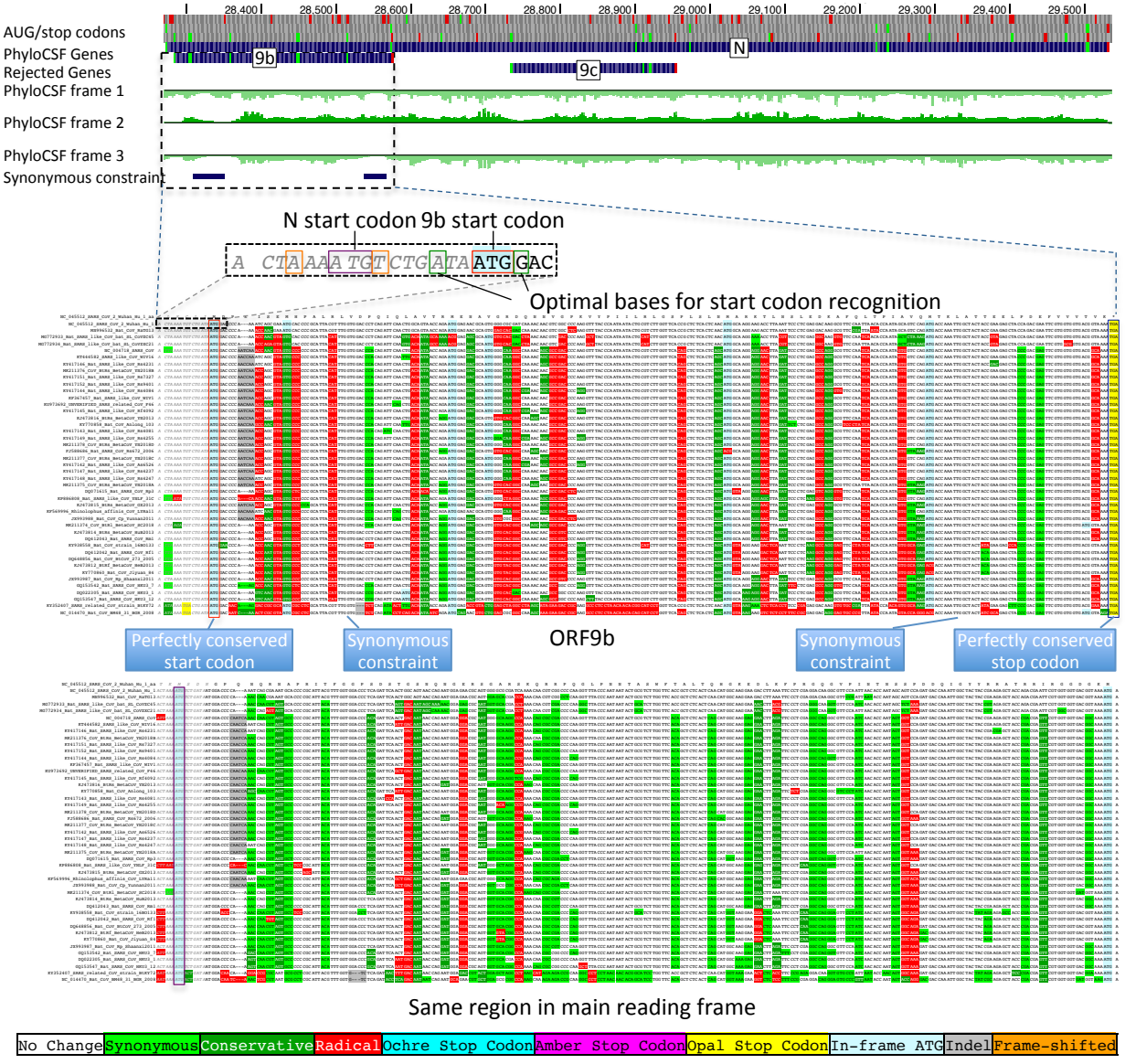
- ORF10 is very highly conserved at nucleotide level
  - It is doing something important (part of 3' UTR RNA structure)
- But it isn't coding for protein
  - No PhyloCSF signal
  - Premature stop in most genomes
  - Almost all substitutions are non-synonymous (red and dark green)
  - Nucleotide-level constraint extends beyond ORF on both sides
- Wu et al 2020 report a partial transcription regulatory sequence (TRS) CUAAAC
  - It is only present in SARS-CoV-2 and Bat RaTG13
- Kim et al. did not find evidence of ORF10 sgRNA in their nanopore sequencing

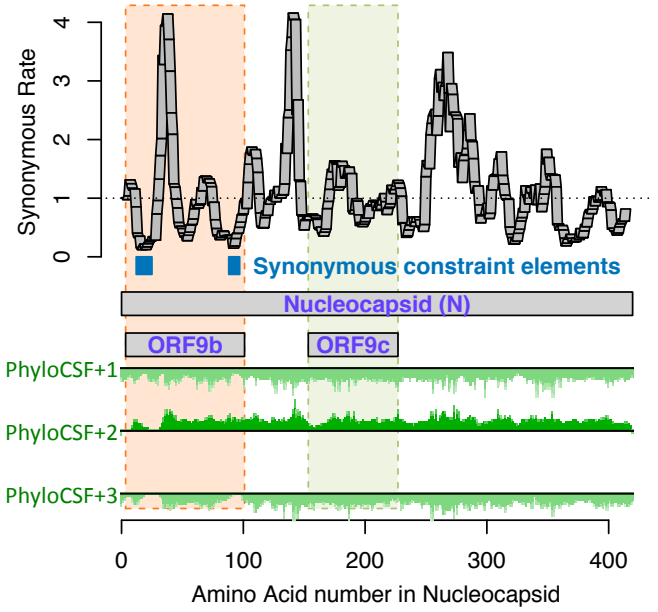
Ribosome profiling evidence for ORF 10?

- Finkel et al. report ribosome footprints in ORF 10
  - But nearly all footprints are either in uORF or downstream ORF (18 codons, 5 in most strains)
  - No more footprints within unique part of ORF 10 than past the end
  - Looks like translational noise



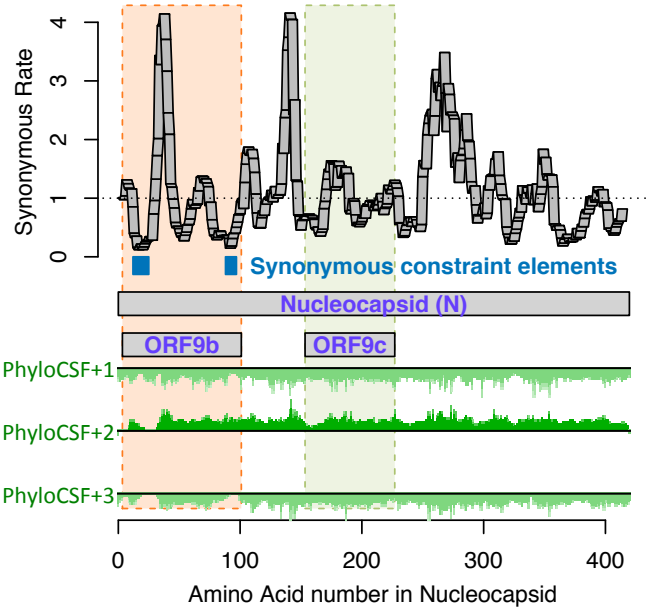
Dual coding regions in N: ORF9b is coding



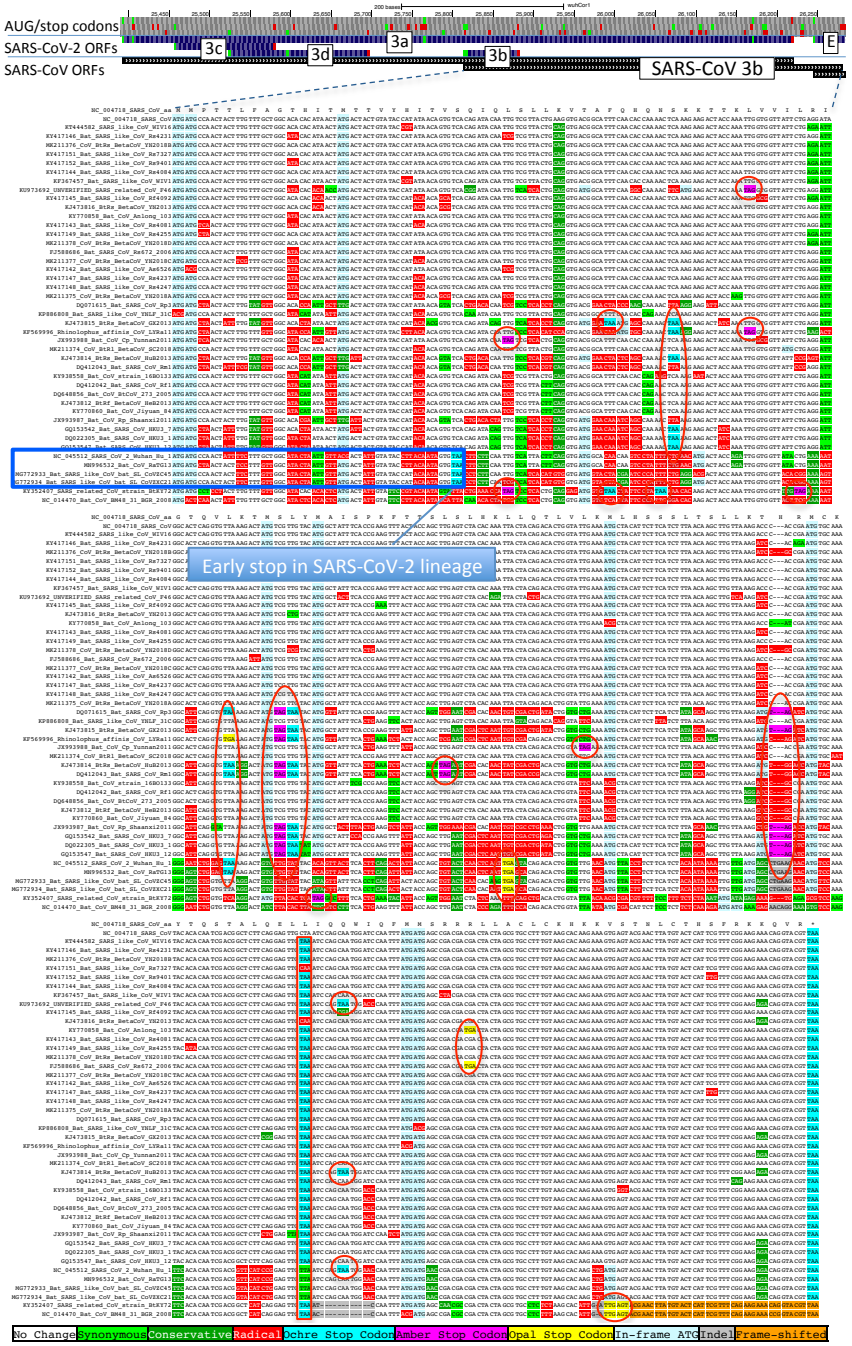


- ORF9b: 97 codons
- Evolutionary evidence is mixed:
  - Has some synonymous constraint
  - But also has regions of accelerated synonymous substitutions
  - Somewhat negative PhyloCSF
  - Perfectly conserved start and stop
- Translation mechanism:
  - Could translate through “leaky scanning”
- There is proteomics evidence
- Experimental evidence in SARS-COV
- Conclusion: likely protein-coding

Dual coding regions in N: ORF9c not coding



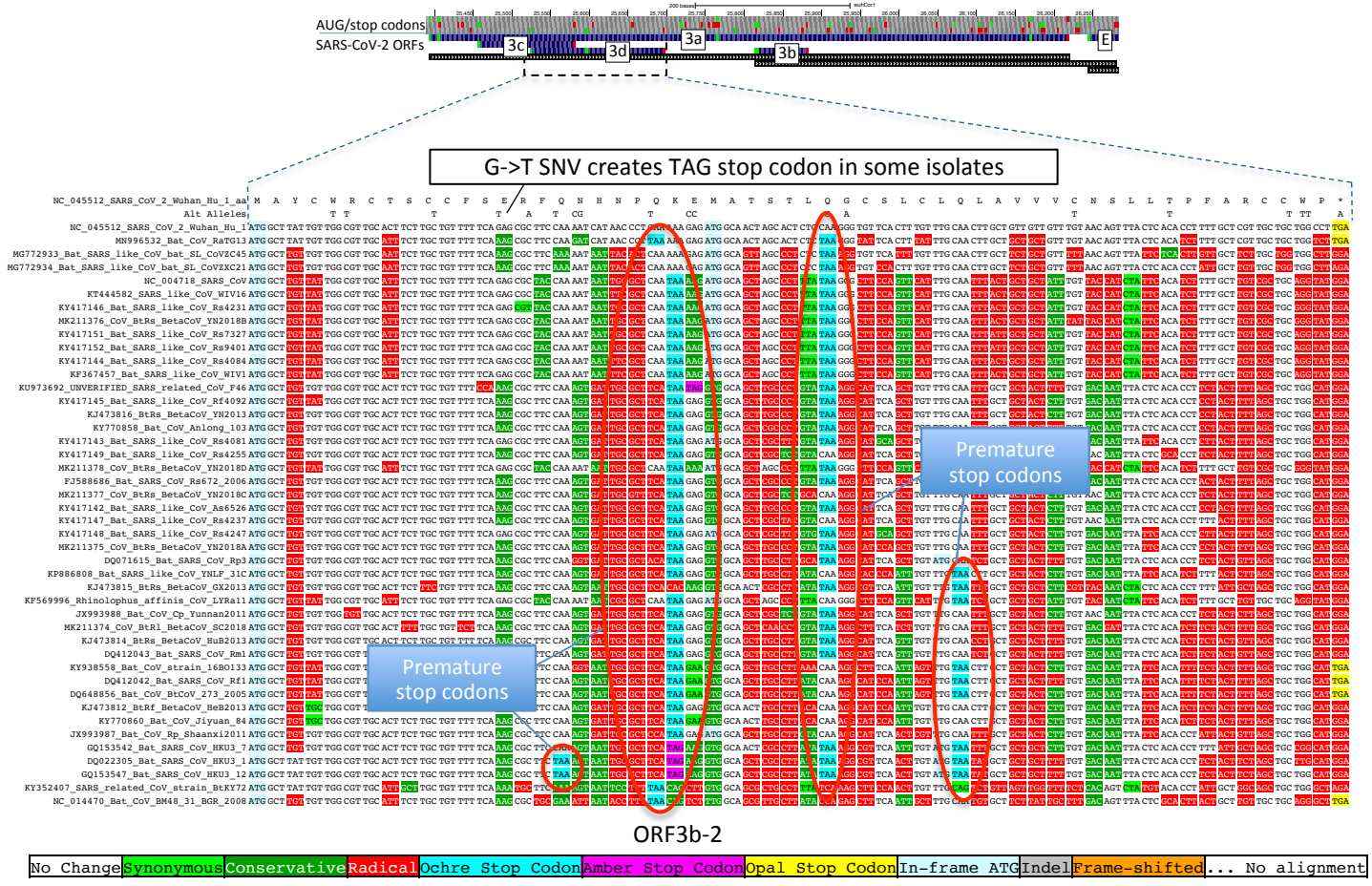
- ORF9c: 73 codons
- Evolutionary evidence is not good
  - No synonymous constraint
  - Has regions of accelerated synonymous substitutions
  - Poor PhyloCSF
  - Imperfectly conserved start and stop
- Translation mechanism:
  - No subgenomic RNA
  - No TRS
  - Unlikely to translate through leaky scanning
- No proteomics evidence
- No experimental evidence in SARS-COV
- Some isolates have internal stop codons
- Conclusion: not protein-coding



ORF3b is not coding

- SARS-CoV has 154-codon ORF3b
- Truncated to 22 in SARS-CoV-2
- Many stops in other strains
- Clearly not conserved coding
- No good way to translate it
- WARNING: Overlapping ORF names are a mess. (We are fixing.)

ORF3d is not coding

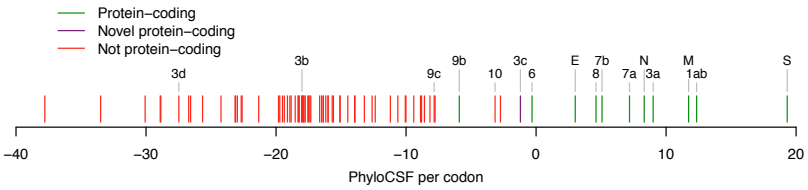


- 57-codon ORF
- Early stops in all other strains so not conserved in Sarbecovirus
- Nonsense SNV in some isolates, so unlikely to be a functional in SARS-CoV-2
- Riboseq did not find initiation at start ATG, but did at downstream ATG (33 codons)
  - Would need to skip four ATGs to translate by leaky scanning

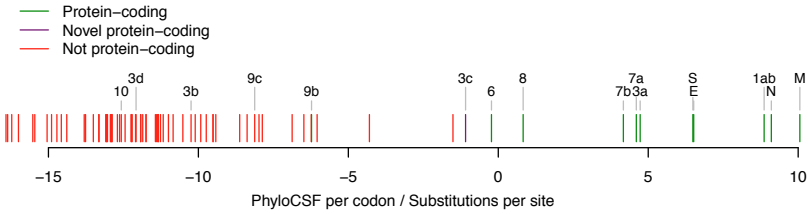
Novel conserved coding ORF3c

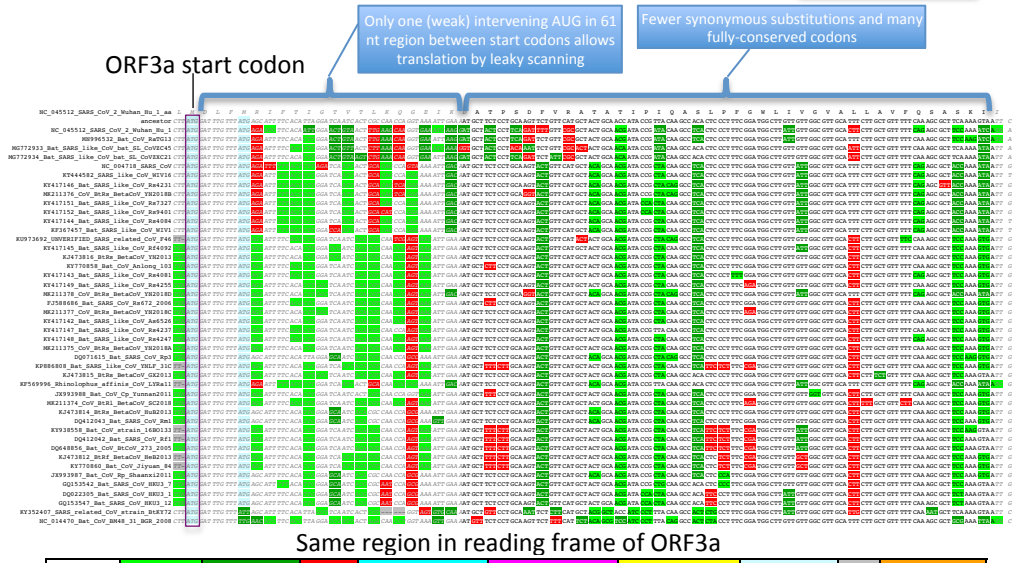
- Ran PhyloCSF on all ORFs >= 10 codons
  - Included near cognate
  - Both strands
- ORFs 10 and 9c artificially high: low rate compresses to 0
- Least negative PhyloCSF is 41-codon “ORF3c”
- Overlaps ORF3a near 5’ end (i.e., start)
- Expect lower score due to overlap
- ORF coincides with Synonymous Constraint Elements
- Most substitutions are synonymous
- ATG conserved except one GTG
- Conserved stop with 1-codon extension in SARS-CoV-2 /RaTG13
- No in-frame premature stop codons in any strain
- Translated by leaky scanning from ORF 3a subgenomic RNA
- Independently discovered by:
  - Cagliani et al. and by Firth using synonymous constraint
  - Finkel et al. using Ribosome profiling

Scores of real & hypothetical AUG ORFs >= 25 codons

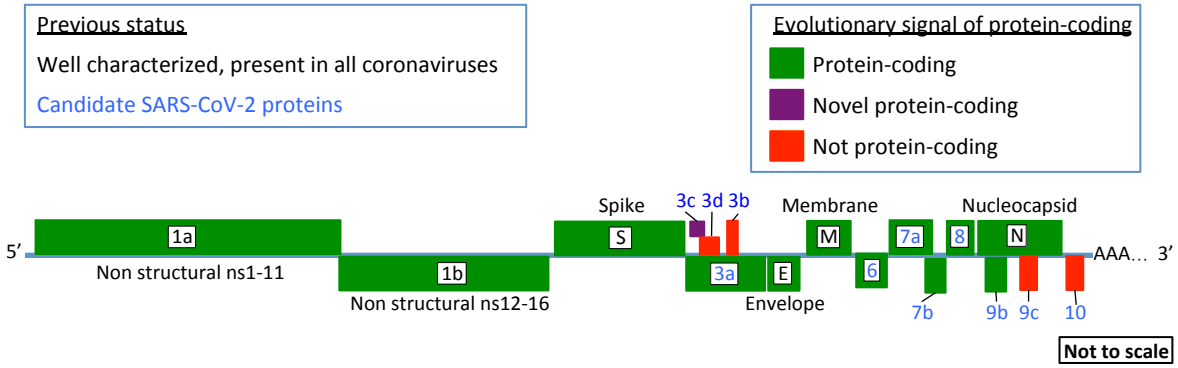


Adjusted for evolutionary rate





A new reference gene set for SARS-CoV-2



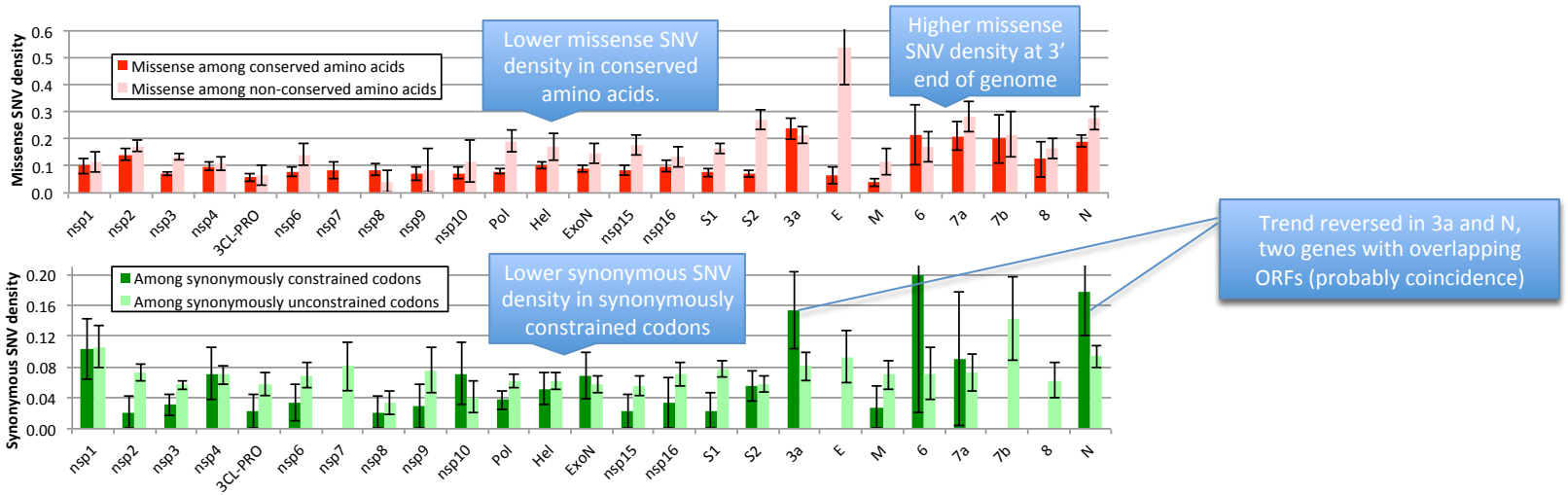
- Conserved protein-coding:
  - ORFs 1ab, S, 3a, 3c, E, M, 6, 7a, 7b, 8, N, and 9b
- Not conserved protein-coding (probably not coding at all)
  - ORFs 2b, 3d, 3d-2, 3b, 9c, 10
- Consistent with experimental evidence
- Gene set affects variant classification

Outline

- SARS-CoV-2 basics
- Evolutionary signatures
- SARS-CoV-2 gene content
- Mutation impact

Can Sarbecovirus conservation inform SARS-CoV-2 variant impact?

- Thousands of Single Nucleotide Variants (SNVs) have been found in SARS-CoV-2
- Which ones are important?
- Guess: ones in Sarbecovirus-conserved position are more likely to be important
- If so, would expect Sarbecovirus-conserved positions to be less likely to have SNVs?
- How to define Sarbecovirus-conserved?
  - Missense: no amino acid changing substitutions
    - Does not depend on tree
  - Synonymous: use FRESCo 1-codon windows
    - synonymous rate < 1.0
    - nominal p-value < 0.034 (FDR 0.125)
    - Note: relative to gene average
  - Noncoding: no substitutions
- Missense variants depleted in conserved amino acids:
  - In conserved amino acids: 607 / 6480, 9.4%
  - In non-conserved amino acids: 535 / 3264, 16.4% ($p<10^{-10}$)
- Synonymous variants depleted in synonymously-constrained codons:
  - In synonymously-constrained codons: 73 / 1394, 5.2%
  - In non-synonymously-constrained codons: 555 / 8350, 6.6% (p=0.029)
- Surprising *excess* of variants in conserved non-coding positions: (but not significant)
  - (17.4% vs. 13.7%, p=0.17).

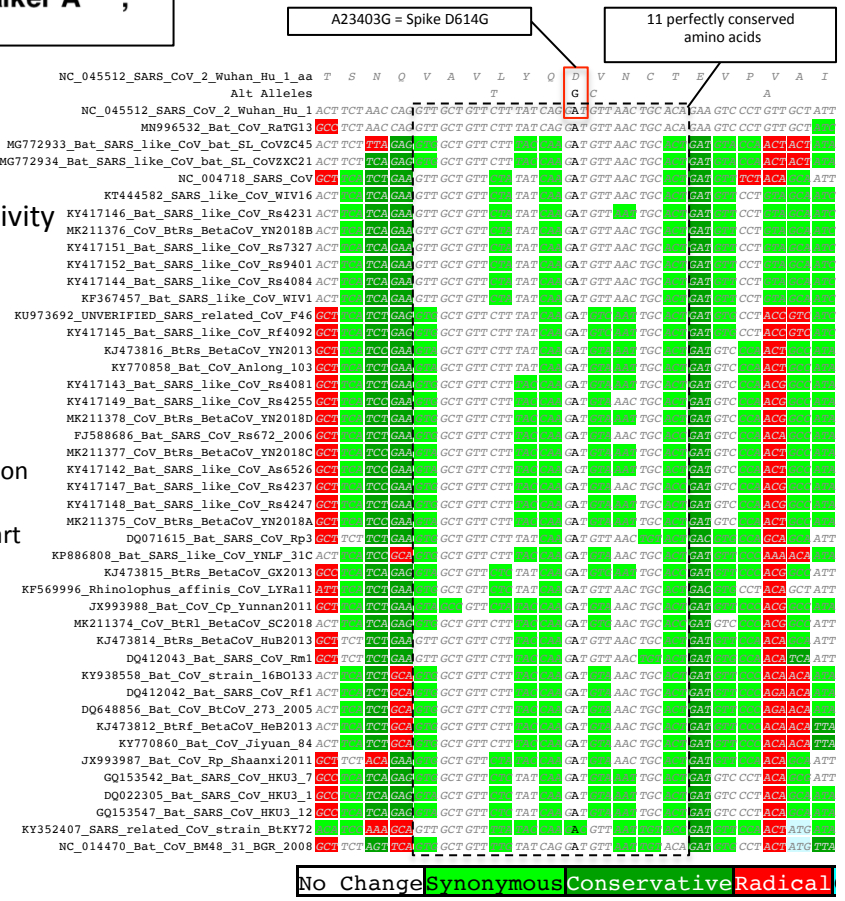


Analysis of Spike D614G and more

Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus

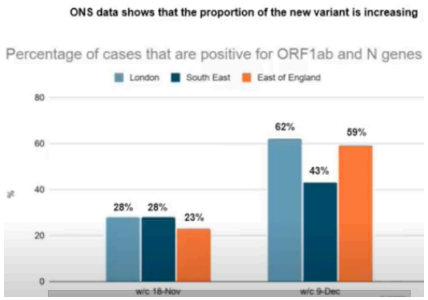
Korber B[1,2], Fischer WM[1], Gnanakaran S[1], Yoon H[1], Theiler J[1], Abfalterer W[1], Hengartner N[1], Giorgi EE[1], Bhattacharya T[1], Foley B[1], Hastie KM[3], Parker MD[4], Partridge DG[5], Evans CM[5], Freeman TM[4], de Silva TI[5,6], on behalf of the Sheffield COVID-19 Genomics Group[#], McDanal C[7], Perez LG[7], Tang H[7], Moon-Walker A[3,8,9], Whelan SP[9], LaBranche CC[7], Saphire EO[3], and Montefiori DC[7]

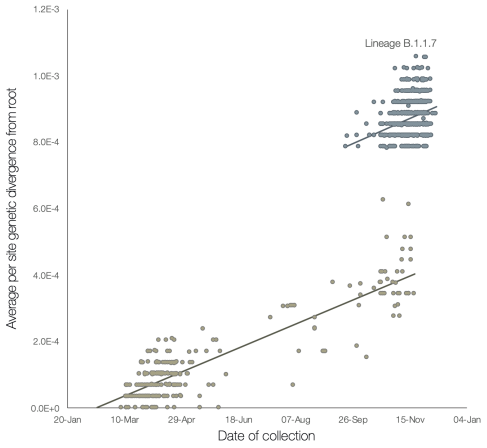
- Korber et al. Cell publication and bioRxiv preprint
- Analyzed spike mutations missense mutations
- D614G
  - Rise in frequency in multiple geographic locations -> increases infectivity
  - Associated with higher viral load but no difference in severity
  - Hou Preprint: "enhances infectivity, replication fitness, and early transmission."
  - **In string of 11 perfectly conserved amino acids**
    - **evidence of conserved function in bat hosts -> human adaptation**
  - Hitchhikers or drivers?
    - Pol P4715L: radical change in conserved amino acid in a highly-conserved region
    - C3037T synonymous mutation in nsp3, not synonymously constrained
    - C241T perfectly-conserved non-coding 5'-UTR position, 25 nt 5' of ORF1ab start
- 15 others
  - V615I/F and P1263L are in perfectly-conserved AAs
  - A831V and A829T/S mostly-conserved in highly-conserved region
  - L5F, L8V/W, R21I/K/T, H49Y, Y145H, Q239K, G476S, V483A
    - in non-conserved amino acids
    - in poorly-conserved regions of the protein
    - less likely to be required for a conserved function
  - V367F, D839Y/N/E, D936Y/H
    - in moderately-conserved contexts
    - ambiguous interpretation



D614G

B.1.1.7, a rapidly spreading lineage





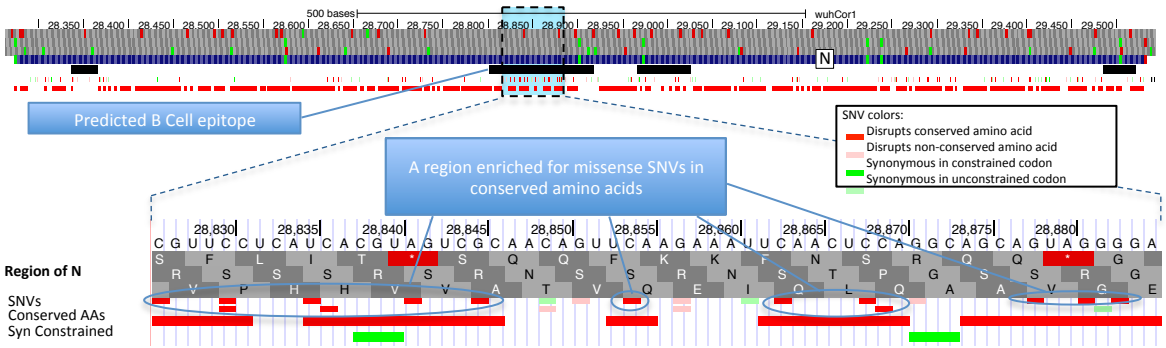
Spike N501Y

- B.1.1.7 lineage appeared in UK in September
  - Also known as VOC 202012/01 − Variant of Concern
  - Also known as UK variant or London variant
- Fairly easy to detect candidates from PCR test
  - PCR test uses primer pairs for ORF1ab, S, and N
  - Variant interferes with pairs for S − "S dropout" results
  - If low Ct value (i.e., lots of RNA), sequence to verify
- Quickly became dominant in UK
  - 28% of new London cases 18-Nov → 62% 9-Dec
  - During UK lockdown 5-Nov -- 12-Dec, B.1.1.7 increased while wildtype decreased in 200/244 UK local authorities [1]
    - Implication: wildtype-stopping lockdown might not stop B.1.1.7
  - Not founder effect: background of 1000s of new cases per day
  - Implication: Variant *causes* faster spread
  - 1.4 − 1.8 times as transmissible
- No effect on virulence
- Probably no effect on vaccine efficacy
- Risk: hospital overload

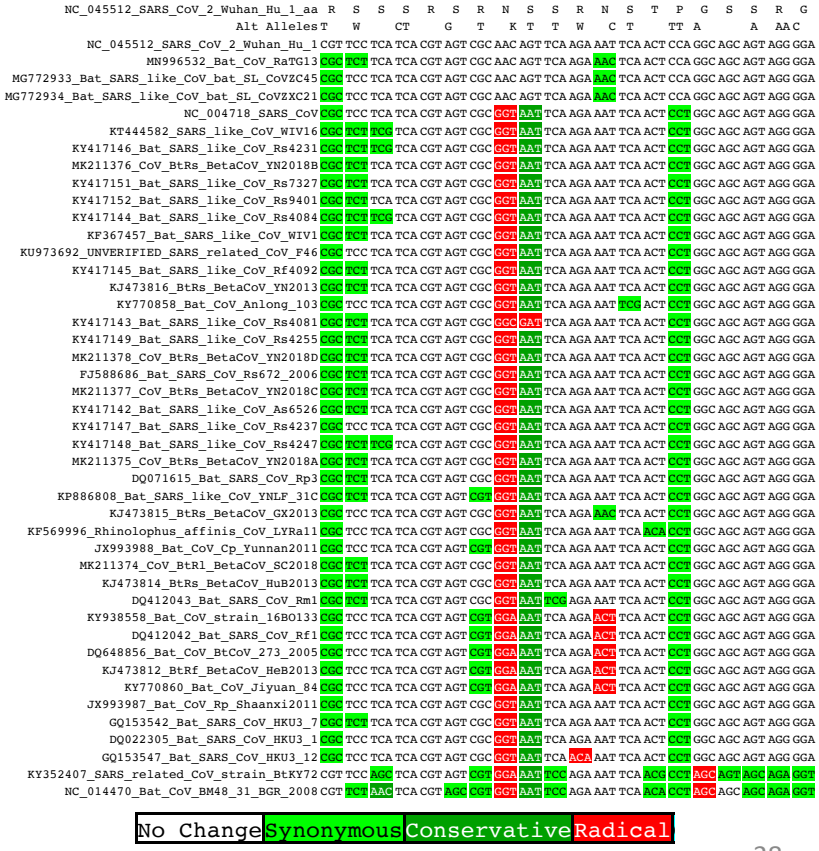
- SARS-CoV-2 mutations usually 1-2 per month
- B.1.1.7 has unprecedented 14 lineage-specific amino-acid changing mutations[2]
- Could have developed within chronically infected immunodeficient SARS-CoV-2 patient
- Includes 8 mutations in Spike
- N501Y
  - alters key contact residue in the receptor binding domain (RBD)
  - experimental data: mutation N501Y increases ACE2 receptor affinity
  - Also independently in South Africa variant
  - Highly variable in Sarbecovirus
    - Could mean unimportant
    - Alternatively, varies based on host ACE2 receptor

[1] Vöhringer et al, "Lineage-specific growth of SARS-CoV-2 B.1.1.7 during the English national lockdown", virological.org
[2] Rambaut et al, "Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations", virological.org

SNV-enriched region in N



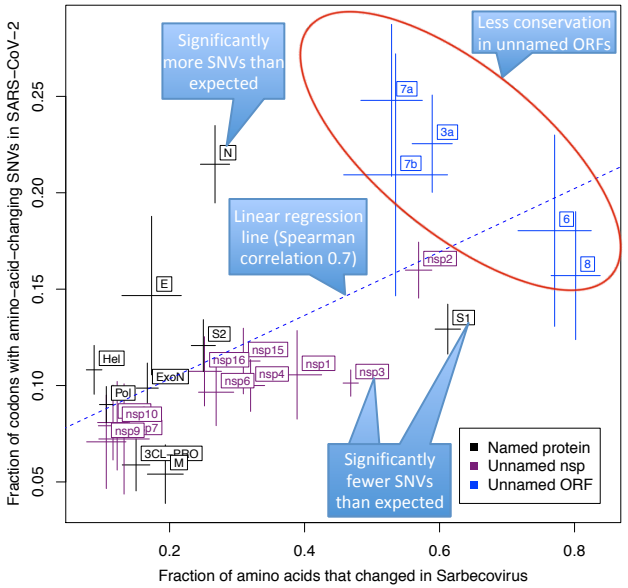
- Nucleocapsid region enriched for SNVs disrupting conserved amino acids
  - 14 variants disrupting conserved residues in 20-amino-acid region
  - Enriched relative to (already high) number for nucleocapsid
  - p<0.012 after conservative genome-wide multiple-hypothesis correction
  - Selection: positive or relaxed purifying
  - Overlaps a predicted B-Cell epitope
    - Positive selection for immune system avoidance?
- No other significantly enriched or depleted regions



Genes with mismatched missense rates

- Faster-evolving proteins across Sarbecoviruses showed more missense mutations in SARS-CoV-2 (Spearman 0.70)
- Significant deviations from trend: N, nsp3, S1

Gene	AAs with SNV	Expected	Excess	Nominal p-value
S1	13%	17%	-28	0.0017
Nsp3	10%	15%	-90	$<10^{-9}$
Nucleocapsid	21%	11%	+42	$<10^{-8}$



- Excess SNVs for N could be positive or relaxed purifying selection
- Some possible explanation for fewer than expected SNVs in nsp3 and S1:
  1. Recent change in mutation rates or selective pressure
  2. Adaptation/expansion cycle in which some genes (S1, nsp3) change earlier than others; most changes would have occurred before reference genome was sampled
  3. Recombination increased tree branch length for these genes
     - Probably would require recombination outside clade
  4. Different distribution of fitness effects for these genes
     - Mildly deleterious mutations are eliminated in Sarbecovirus time frame
     - Only strongly deleterious mutations are prevented within the pandemic
     - Fraction of all possible deleterious mutations that are strongly deleterious could be higher for these genes than other

Acknowledgements

Coauthors

- Rachel Sealfon
- Manolis Kellis

Virology Consultants

- Jeremy Luban
- Bob Garry

UCSC

- Mark Diekhans

SNVs

- Nextstrain
- GISAID
- Many authors who made genomes available pre-publication

Extra Slides

Furin Cleavage Site

- At the junction of S1 and S2
- Has a role in determining viral infectivity
- Not present in closest relatives of SARS-CoV-2
- Other coronaviruses also have furin cleavage sites