SARS-CoV-2 Final Project-The Wanderers

James Medwid, Gary Peng, Neel Srejan, Ishan Ranjan
BENG/CSE/BIMM 182
Dr. Vineet Bafna

Meet The Wanderers

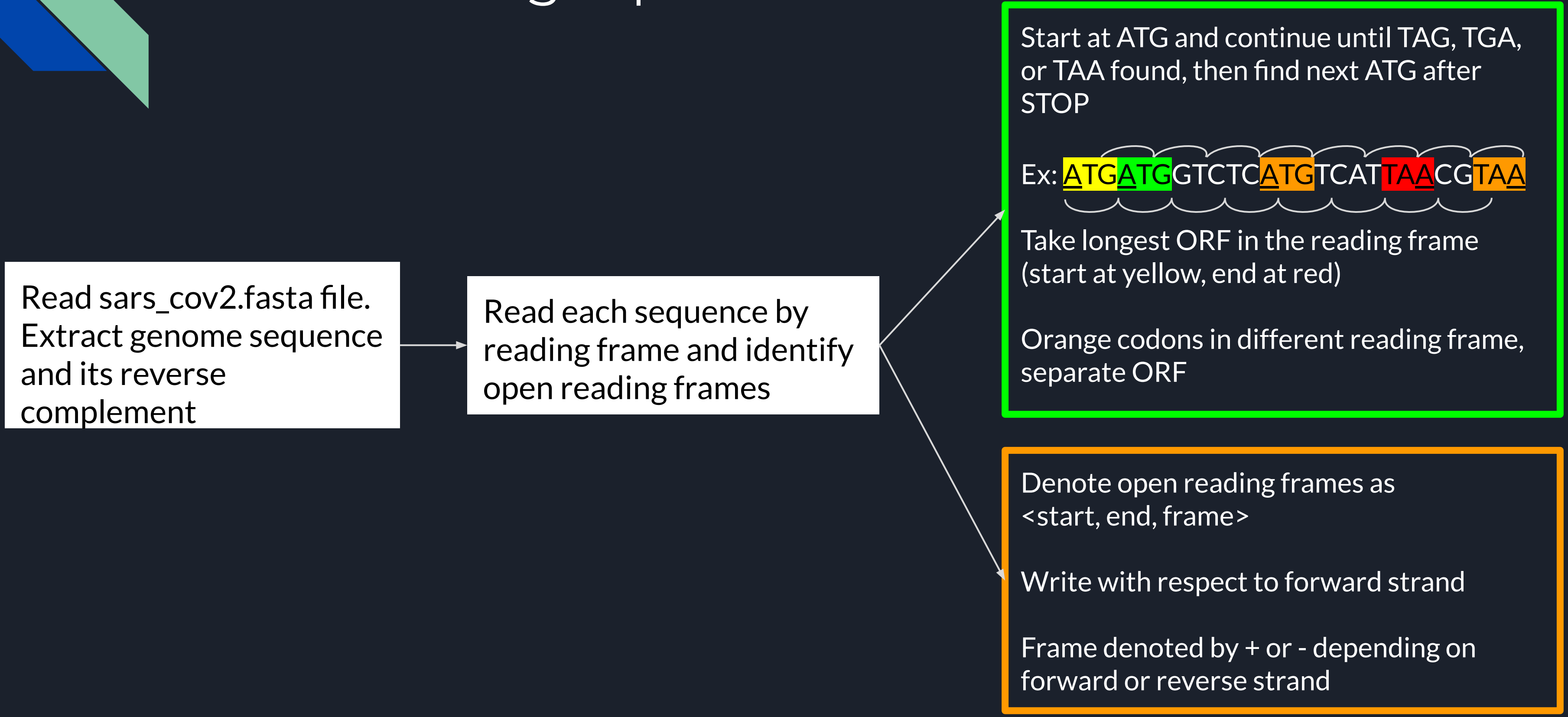
James Medwid
3rd Year
Biology: Bioinformatics

Gary Peng
3rd Year
Biology: Bioinformatics

Ishan Ranjan
3rd Year
Bioengineering: Bioinformatics

Neel Srejan
4th Year
Biology: Bioinformatics

ORF Finding Pipeline



Read sars_cov2.fasta file. Extract genome sequence and its reverse complement

Read each sequence by reading frame and identify open reading frames

Start at ATG and continue until TAG, TGA, or TAA found, then find next ATG after STOP

Ex: ATGATGGTCTCATGTCATTAACGTAA

Take longest ORF in the reading frame (start at yellow, end at red)

Orange codons in different reading frame, separate ORF

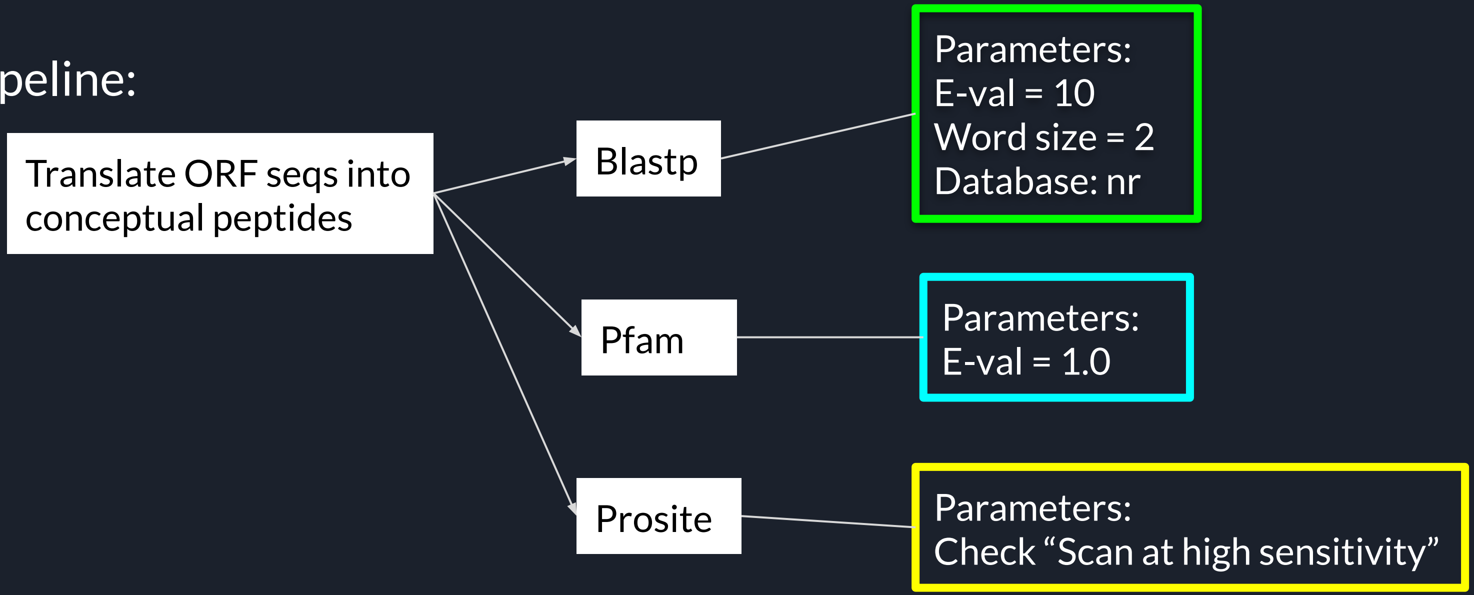
Denote open reading frames as <start, end, frame>

Write with respect to forward strand

Frame denoted by + or - depending on forward or reverse strand

Functional Analysis Pipeline

Pipeline:



Translate ORF seqs into conceptual peptides

Blastp

Parameters:
E-val = 10
Word size = 2
Database: nr

Pfam

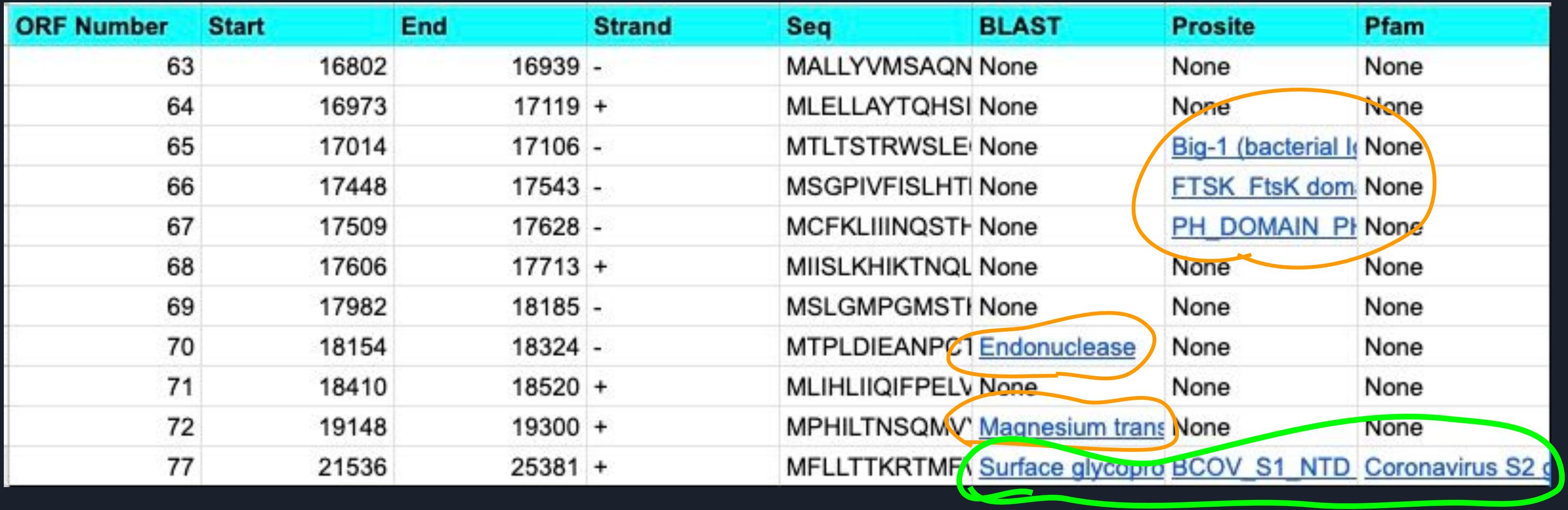
Parameters:
E-val = 1.0

Prosite

Parameters:
Check “Scan at high sensitivity”

Genes Found

Bacterial genes



ORF Number	Start	End	Strand	Seq	BLAST	Prosite	Pfam
63	16802	16939	-	MALLYVMSAQN	None	None	None
64	16973	17119	+	MLELLAYTQHSI	None	None	None
65	17014	17106	-	MTLTSTRWSLE	None	Big-1 (bacterial I	None
66	17448	17543	-	MSGPIVFISLHTI	None	FTSK_FtsK dom	None
67	17509	17628	-	MCFKLIIINQSTH	None	PH_DOMAIN_PH	None
68	17606	17713	+	MIISLKHIKTNQL	None	None	None
69	17982	18185	-	MSLGMPGMSTI	None	None	None
70	18154	18324	-	MTPLDIEANPC1	Endonuclease	None	None
71	18410	18520	+	MLIHLIIQIFPELV	None	None	None
72	19148	19300	+	MPHILTNSQMV'	Magnesium trans	None	None
77	21536	25381	+	MFLLTTKRTMF\	Surface glycopro	BCOV_S1_NTD	Coronavirus S2 g

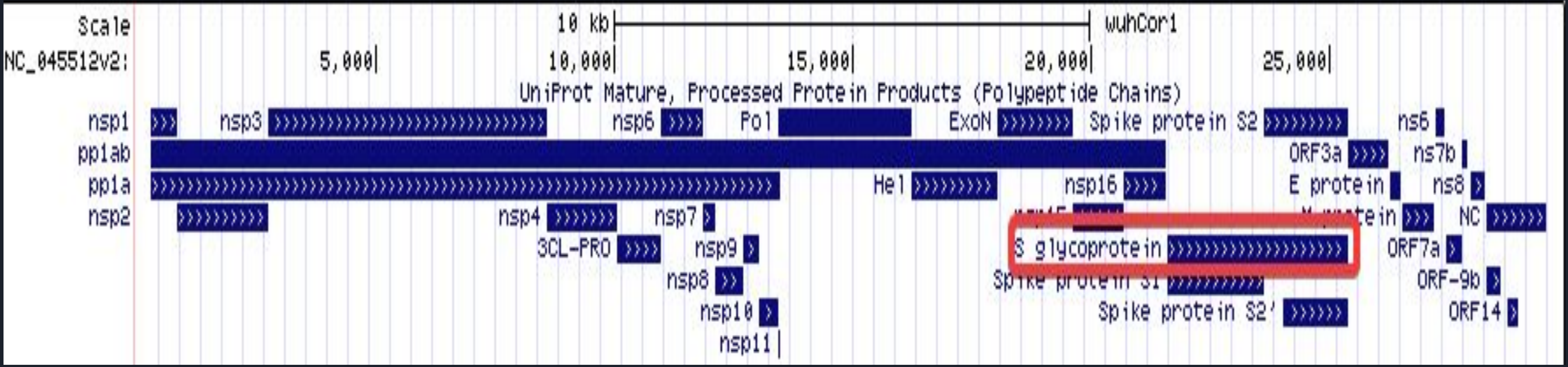
Candidate genes:

SARS-CoV2 gene

- ORF 77 (+)
- ORF 65, 66, 67, 70, and 72 match non-viral (bacterial) genes: not likely true gene.

How Our Results Match the Literature

ORF 77 (3846 bp) matches the S gene on SARS-Cov2 (3822 bp) [2]

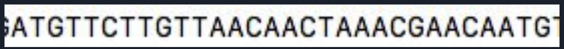


https://genome.ucsc.edu/covid19.html

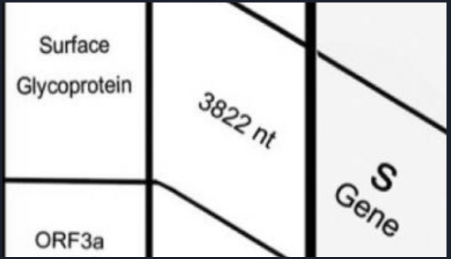
How Our Results Match the Literature

ORF 77 (3846 bp) matches the S gene on SARS-Cov2 (3822 bp) [2]

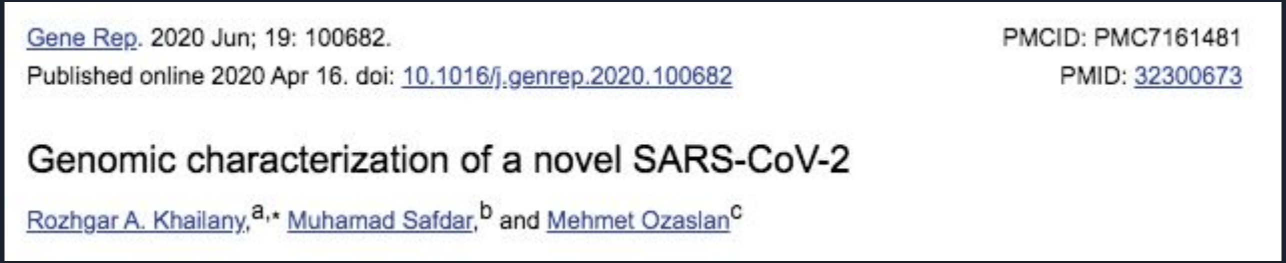
- Discrepancy in length (24 bp) suggests alternative start site:
- There are ATG at position 21536 and at position 21563.



ATGTTCTTGTTAACAACTAAACGAACAATGT



- The paper included the STOP codon in the length of their S gene.

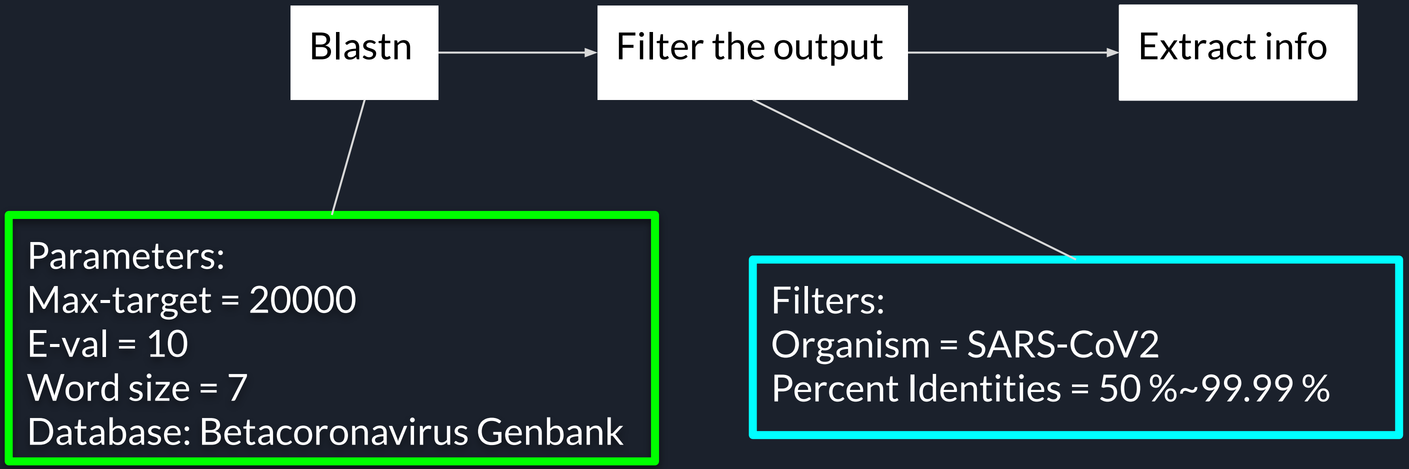


Gene Rep. 2020 Jun; 19: 100682. PMCID: PMC7161481
Published online 2020 Apr 16. doi: 10.1016/j.genrep.2020.100682 PMID: 32300673

Genomic characterization of a novel SARS-CoV-2

Rozhgar A. Khailany,[a,*] Muhamad Safdar,[b] and Mehmet Ozaslan[c]

Mutation Finding Pipeline



Blastn

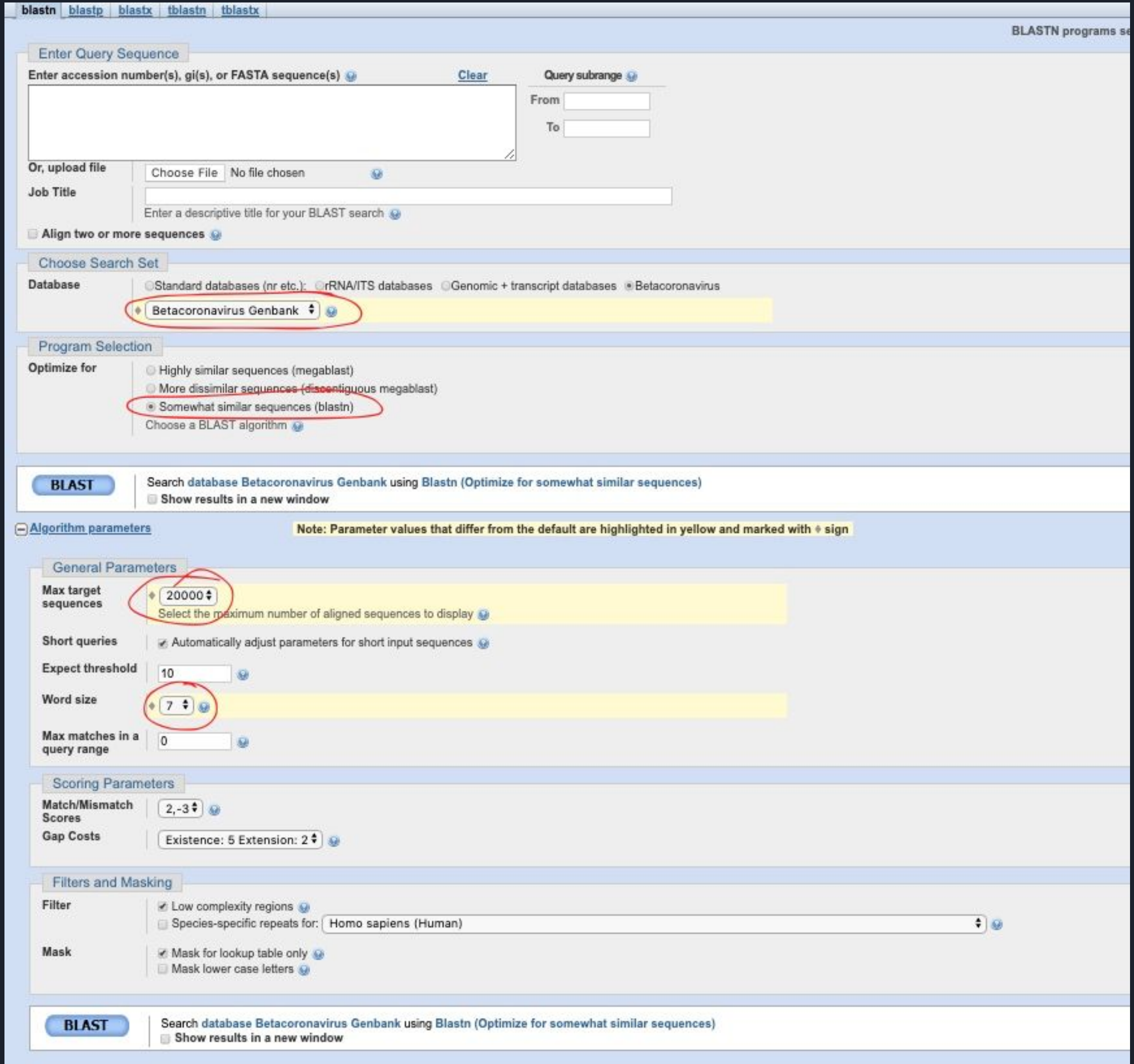
Filter the output

Extract info

Parameters:
Max-target = 20000
E-val = 10
Word size = 7
Database: Betacoronavirus Genbank

Filters:
Organism = SARS-CoV2
Percent Identities = 50 %~99.99 %

BLASTN



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) — Clear

Query subrange — From — To

Or, upload file: Choose File — No file chosen

Job Title: Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: ○ Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ● Betacoronavirus

♦ Betacoronavirus Genbank

Program Selection

Optimize for:
○ Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
● Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST — Search database Betacoronavirus Genbank using Blastn (Optimize for somewhat similar sequences)
☐ Show results in a new window

⊖ Algorithm parameters — Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences: ♦ 20000 — Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: ♦ 7

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

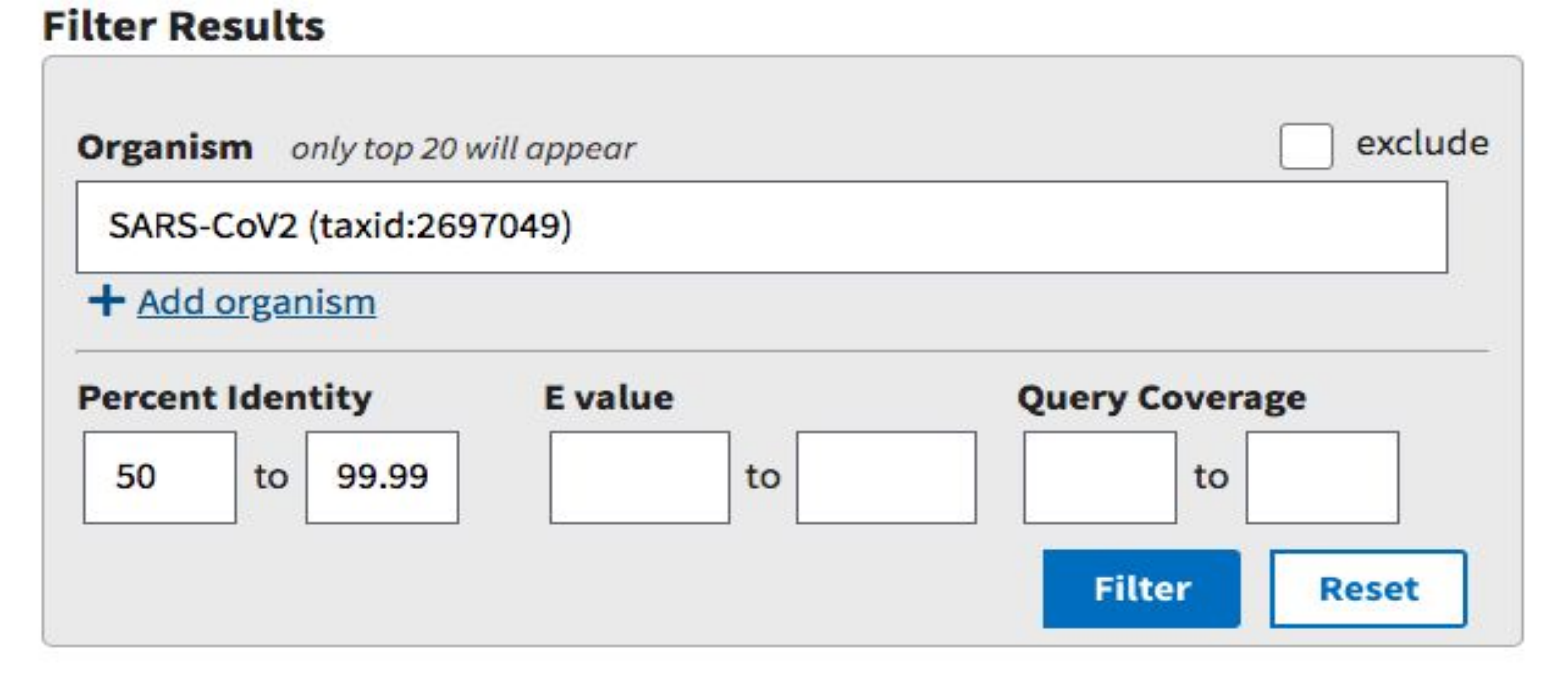
Filters and Masking

Filter: ☑ Low complexity regions
☐ Species-specific repeats for: Homo sapiens (Human)

Mask: ☑ Mask for lookup table only
☐ Mask lower case letters

BLAST — Search database Betacoronavirus Genbank using Blastn (Optimize for somewhat similar sequences)
☐ Show results in a new window

Filtering Our Results



Filter Results

Organism *only top 20 will appear* □ exclude

SARS-CoV2 (taxid:2697049)

Percent Identity: 50 to 99.99

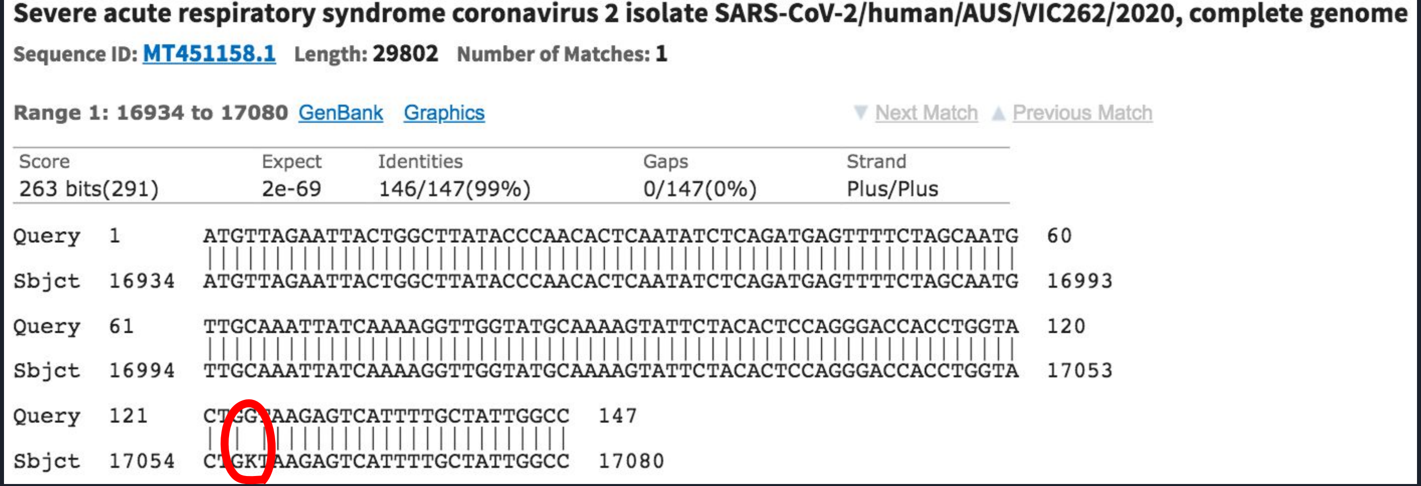
E value: to

Query Coverage: to

Extracting mutations

Read the alignment, and exclude:

- Mismatches due to ambiguous bases in genome (Y, M, K, N, etc)



Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/AUS/VIC262/2020, complete genome

Sequence ID: MT451158.1 **Length:** 29802 **Number of Matches:** 1

Range 1: 16934 to 17080 GenBank Graphics

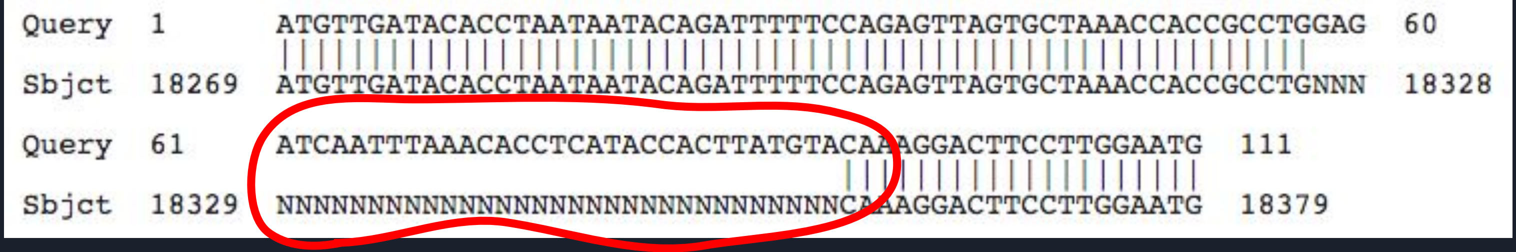
Score	Expect	Identities	Gaps	Strand
263 bits(291)	2e-69	146/147(99%)	0/147(0%)	Plus/Plus

```
Query  1      ATGTTAGAATTACTGGCTTATACCCAACACTCAATATCTCAGATGAGTTTTCTAGCAATG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16934  ATGTTAGAATTACTGGCTTATACCCAACACTCAATATCTCAGATGAGTTTTCTAGCAATG  16993

Query  61     TTGCAAATTATCAAAAGGTTGGTATGCAAAAGTATTCTACACTCCAGGGACCACCTGGTA  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16994  TTGCAAATTATCAAAAGGTTGGTATGCAAAAGTATTCTACACTCCAGGGACCACCTGGTA  17053

Query  121    CTGGTAAGAGTCATTTTGCTATTGGCC  147
              ||| |||||||||||||||||||||||
Sbjct  17054  CTGKTAAGAGTCATTTTGCTATTGGCC  17080
```

Extracting mutations

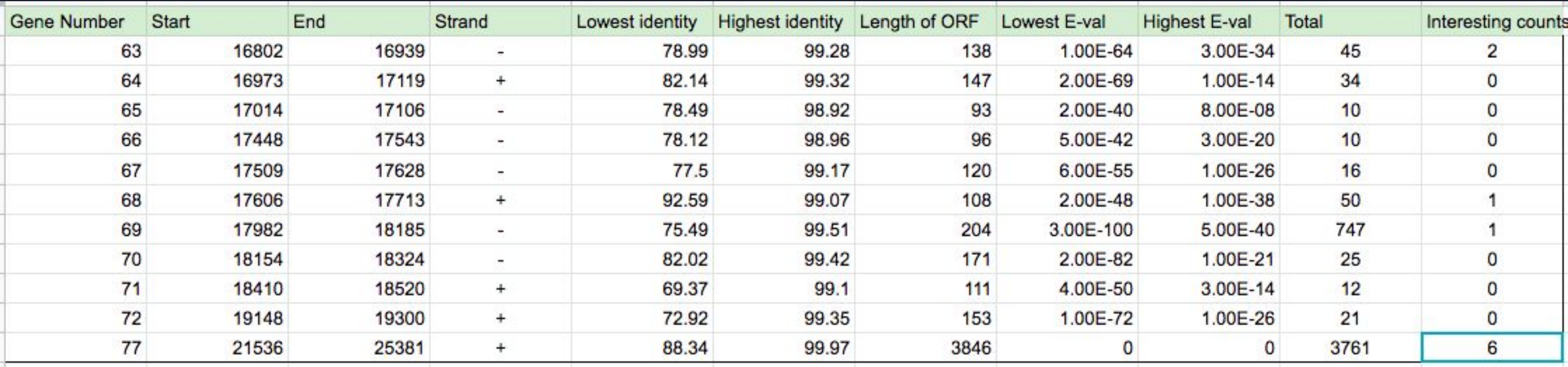
Read the alignment, and exclude:

- Mismatches due to ambiguous based in genome (Y, M, K, N, etc)
- Too many mismatches typically represents gap in the genome, not mutations:



```
Query  1      ATGTTGATACACCTAATAATACAGATTTTTCCAGAGTTAGTGCTAAACCACCGCCTGGAG  60
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||   
Sbjct  18269  ATGTTGATACACCTAATAATACAGATTTTTCCAGAGTTAGTGCTAAACCACCGCCTGNNN  18328

Query  61     ATCAATTTAAACACCTCATACCACTTATGTACAAAGGACTTCCTTGGAATG  111
                                             ||||||||||||||||||||
Sbjct  18329  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAAAGGACTTCCTTGGAATG  18379
```

Mutations!

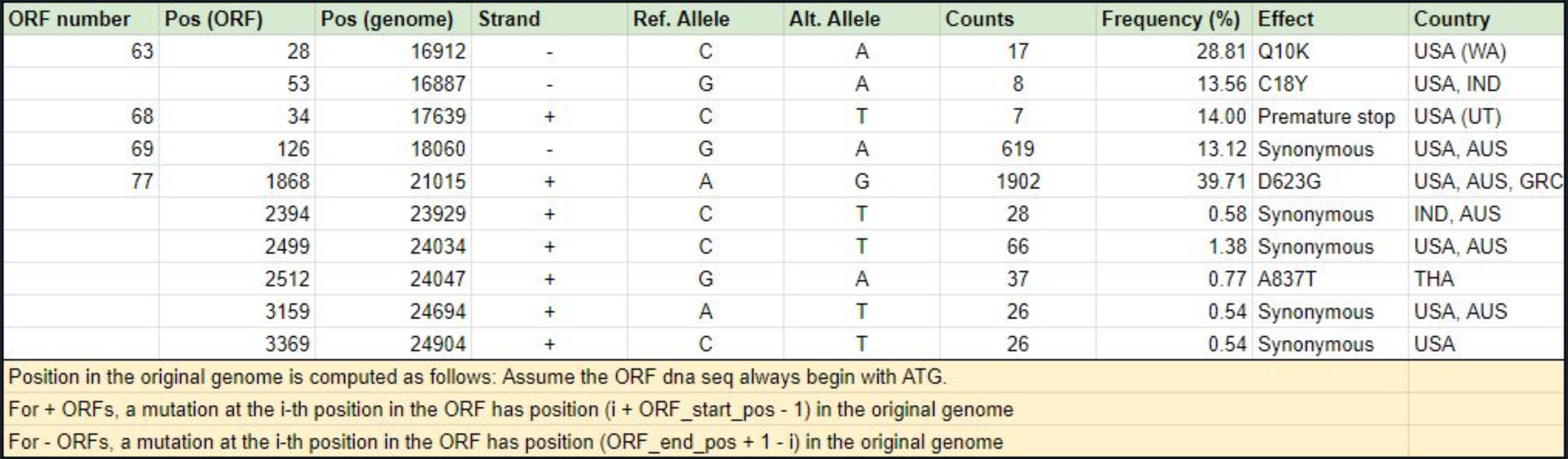


Gene Number	Start	End	Strand	Lowest identity	Highest identity	Length of ORF	Lowest E-val	Highest E-val	Total	Interesting counts
63	16802	16939	-	78.99	99.28	138	1.00E-64	3.00E-34	45	2
64	16973	17119	+	82.14	99.32	147	2.00E-69	1.00E-14	34	0
65	17014	17106	-	78.49	98.92	93	2.00E-40	8.00E-08	10	0
66	17448	17543	-	78.12	98.96	96	5.00E-42	3.00E-20	10	0
67	17509	17628	-	77.5	99.17	120	6.00E-55	1.00E-26	16	0
68	17606	17713	+	92.59	99.07	108	2.00E-48	1.00E-38	50	1
69	17982	18185	-	75.49	99.51	204	3.00E-100	5.00E-40	747	1
70	18154	18324	-	82.02	99.42	171	2.00E-82	1.00E-21	25	0
71	18410	18520	+	69.37	99.1	111	4.00E-50	3.00E-14	12	0
72	19148	19300	+	72.92	99.35	153	1.00E-72	1.00E-26	21	0
77	21536	25381	+	88.34	99.97	3846	0	0	3761	6

Mutations! (cont.)

10 interesting mutations:

- 5 synonymous, 4 results in A.A. change, 1 causes premature STOP (peptide shortened).



ORF number	Pos (ORF)	Pos (genome)	Strand	Ref. Allele	Alt. Allele	Counts	Frequency (%)	Effect	Country
63	28	16912	-	C	A	17	28.81	Q10K	USA (WA)
	53	16887	-	G	A	8	13.56	C18Y	USA, IND
68	34	17639	+	C	T	7	14.00	Premature stop	USA (UT)
69	126	18060	-	G	A	619	13.12	Synonymous	USA, AUS
77	1868	21015	+	A	G	1902	39.71	D623G	USA, AUS, GRC
	2394	23929	+	C	T	28	0.58	Synonymous	IND, AUS
	2499	24034	+	C	T	66	1.38	Synonymous	USA, AUS
	2512	24047	+	G	A	37	0.77	A837T	THA
	3159	24694	+	A	T	26	0.54	Synonymous	USA, AUS
	3369	24904	+	C	T	26	0.54	Synonymous	USA

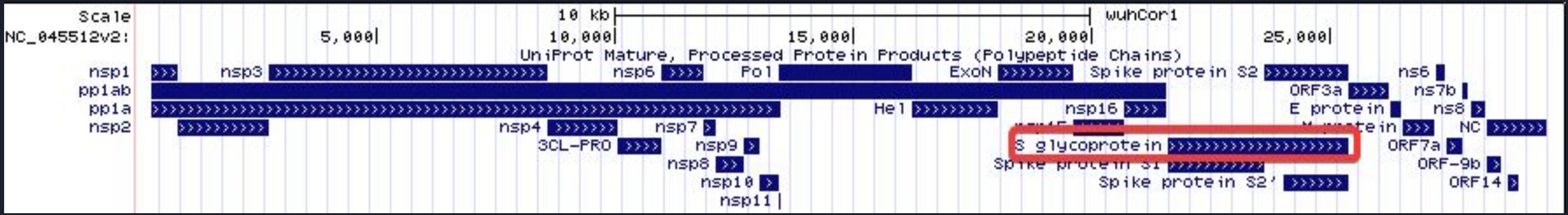
Position in the original genome is computed as follows: Assume the ORF dna seq always begin with ATG.

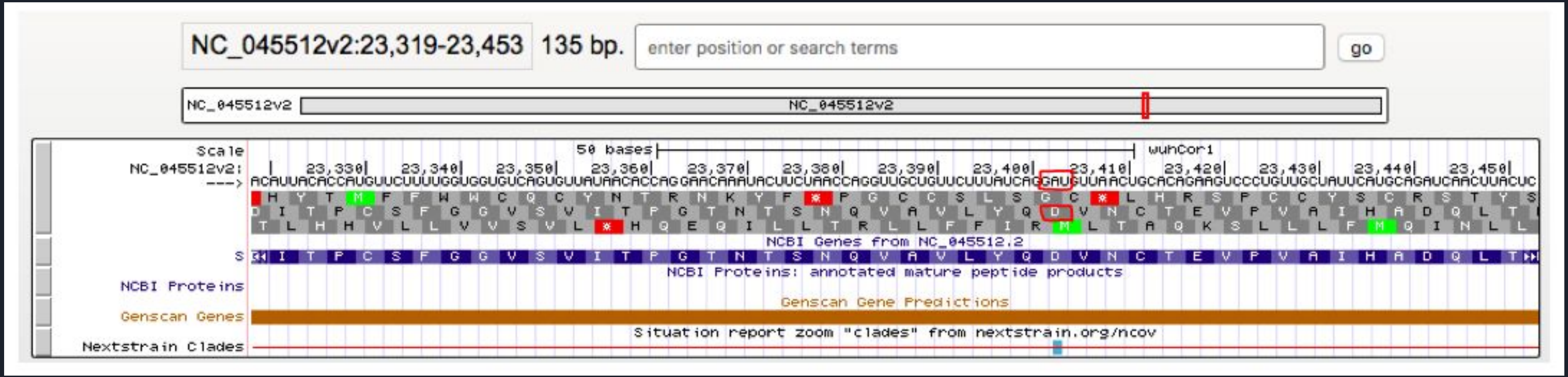
For + ORFs, a mutation at the i-th position in the ORF has position (i + ORF_start_pos - 1) in the original genome

For - ORFs, a mutation at the i-th position in the ORF has position (ORF_end_pos + 1 - i) in the original genome

ORF 77 D623G/D614G Mutation

- The D623G / D614G on S protein is found to be associated with faster virus transmission




NC_045512v2:23,319-23,453 135 bp.

enter position or search terms
go
NC_045512v2
NC_045512v2
Scale
50 bases
wuhCor1
NC_045512v2:
23,330|
23,340|
23,350|
23,360|
23,370|
23,380|
23,390|
23,400|
23,410|
23,420|
23,430|
23,440|
23,450|
--->
ACAUUACACCAUGUUCUUUUUGGUGGUGUCAGUGUUAUAACACCAGGAACAAAUACUUCUAACCAGGUUGCUGUUCUUUAUCAGGAUGUUAACUGCACAGAAGUCCCUGUUGCUAUUCAUGCAGAUCAACUUACUC
NCBI Genes from NC_045512.2
S
NCBI Proteins
NCBI Proteins: annotated mature peptide products
Genscan Genes
Genscan Gene Predictions
Nextstrain Clades
Situation report zoom "clades" from nextstrain.org/ncov

Sources: [1], [3], [4], [5],
https://genome.ucsc.edu/covid19.html

Literature Search References:

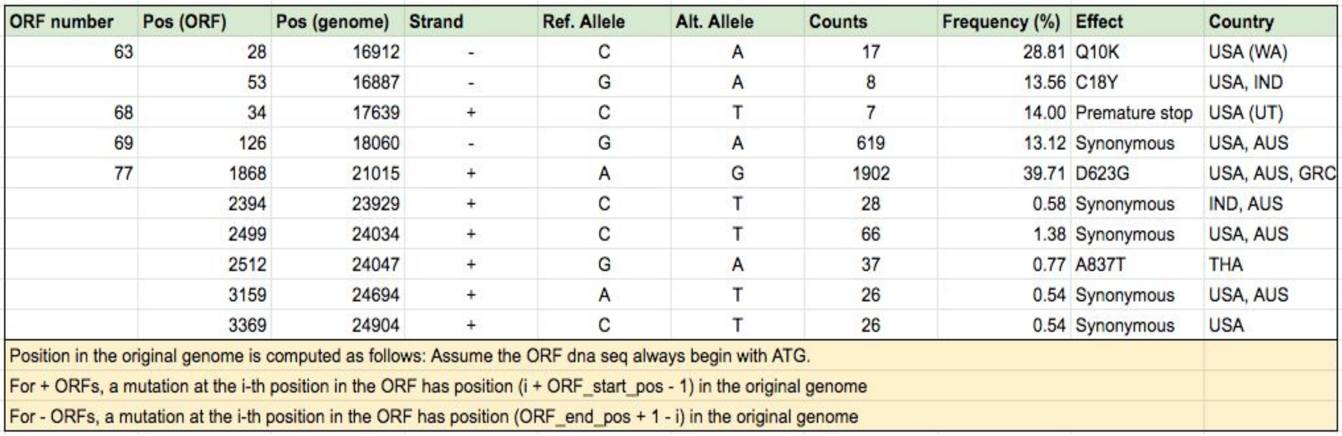
1. Bhattacharyya, C., Das, C., Ghosh, A., Singh, A. K., Mukherjee, S., Majumder, P. P., … Biswas, N. K. (2020). Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of TMPRSS2 and MX1 Genes. *bioRxiv* doi: 10.1101/2020.05.04.075911
2. Khailany, R. A., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. *Gene reports*, *19*, 100682. Advance online publication. https://doi.org/10.1016/j.genrep.2020.100682
3. Korber, B., Fischer, W., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., … Montefiori, D. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *BioRxiv*. doi: 10.1101/2020.04.29.069054
4. Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., Mcguire, A. T., & Veesler, D. (2020). Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. doi: 10.1101/2020.02.19.956581
5. Lowe, D. (2020, May 7). Mutations in the Coronavirus Spike Protein. Retrieved from https://blogs.sciencemag.org/pipeline/archives/2020/05/07/mutations-in-the-coronavirus-spike-protein

THANK YOU!

BIMM 182, Group 9-The Wanderers
James Medwid, Gary Peng, Neel Srejan, Ishan Ranjan

Our data:
https://docs.google.com/spreadsheets/d/1t6EzvHPqy4FHKIYR1vr9SvCjkz6PvoXh1G981aZaU3w/edit?usp=sharing

Updated Mutation Table with general frequencies of mutations:

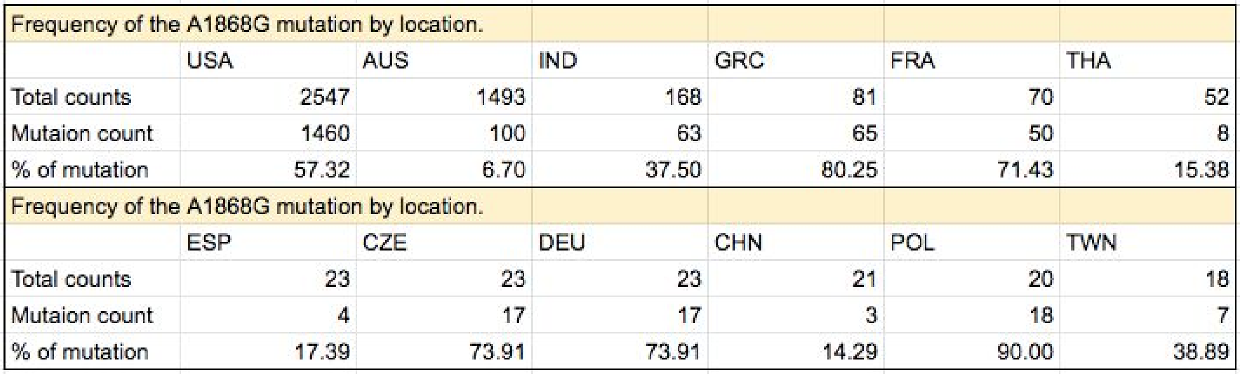


ORF number	Pos (ORF)	Pos (genome)	Strand	Ref. Allele	Alt. Allele	Counts	Frequency (%)	Effect	Country
63	28	16912	-	C	A	17	28.81	Q10K	USA (WA)
	53	16887	-	G	A	8	13.56	C18Y	USA, IND
68	34	17639	+	C	T	7	14.00	Premature stop	USA (UT)
69	126	18060	-	G	A	619	13.12	Synonymous	USA, AUS
77	1868	21015	+	A	G	1902	39.71	D623G	USA, AUS, GRC
	2394	23929	+	C	T	28	0.58	Synonymous	IND, AUS
	2499	24034	+	C	T	66	1.38	Synonymous	USA, AUS
	2512	24047	+	G	A	37	0.77	A837T	THA
	3159	24694	+	A	T	26	0.54	Synonymous	USA, AUS
	3369	24904	+	C	T	26	0.54	Synonymous	USA

Position in the original genome is computed as follows: Assume the ORF dna seq always begin with ATG.
For + ORFs, a mutation at the i-th position in the ORF has position (i + ORF_start_pos - 1) in the original genome
For - ORFs, a mutation at the i-th position in the ORF has position (ORF_end_pos + 1 - i) in the original genome

Additional note on the A1868G mutation in ORF77 (corresponding to the surface glycoprotein gene of SARS-Cov2):

- Causes an amino acid substitution D623G (D614G in literature (*Bhattacharyya et. al, Walls et. al*))
- More frequent in European countries (Greece, France, Czechia, Germany, Poland), less frequent in US, in Australia, and in Asia (India, Taiwan, China, Thailand) [See table↓].



Frequency of the A1868G mutation by location.

	USA	AUS	IND	GRC	FRA	THA
Total counts	2547	1493	168	81	70	52
Mutaion count	1460	100	63	65	50	8
% of mutation	57.32	6.70	37.50	80.25	71.43	15.38

Frequency of the A1868G mutation by location.

	ESP	CZE	DEU	CHN	POL	TWN
Total counts	23	23	23	21	20	18
Mutaion count	4	17	17	3	18	7
% of mutation	17.39	73.91	73.91	14.29	90.00	38.89

BIMM 182, Group 9-The Wanderers
James Medwid, Gary Peng, Neel Srejan, Ishan Ranjan

Literature Search References:

Bhattacharyya, C., Das, C., Ghosh, A., Singh, A. K., Mukherjee, S., Majumder, P. P., … Biswas, N. K. (2020). Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of TMPRSS2 and MX1 Genes. *bioRxiv* doi: 10.1101/2020.05.04.075911

Korber, B., Fischer, W., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., … Montefiori, D. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *BioRxiv*. doi: 10.1101/2020.04.29.069054

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., Mcguire, A. T., & Veesler, D. (2020). Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. doi: 10.1101/2020.02.19.956581

Lowe, D. (2020, May 7). Mutations in the Coronavirus Spike Protein. Retrieved from https://blogs.sciencemag.org/pipeline/archives/2020/05/07/mutations-in-the-coronavirus-spike-protein