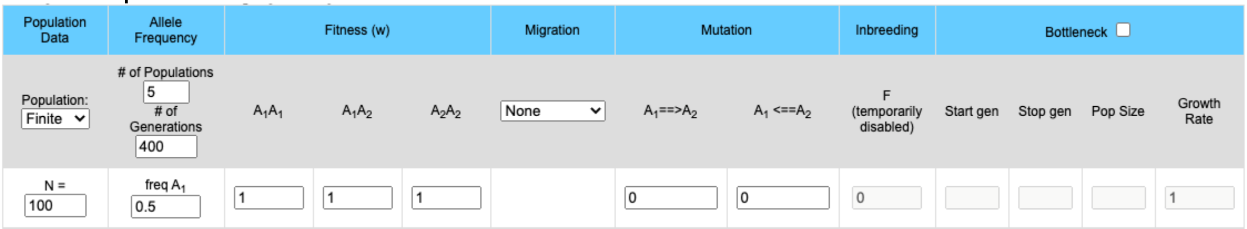
MCBL 127 MODULE #02, 21-Jan DUE 28-Jan before class

Name _____________________ (Name _____________________)
NetID _____________________

Web Popgen
https://www.radford.edu/~rsheehy/Gen_flash/popgen/
This is an online population genetics simulator. It allows users track allele frequencies in up to five independent populations, while changing the population size, starting frequency and selective force. ***The web browser you use will have to have Flash enabled.***

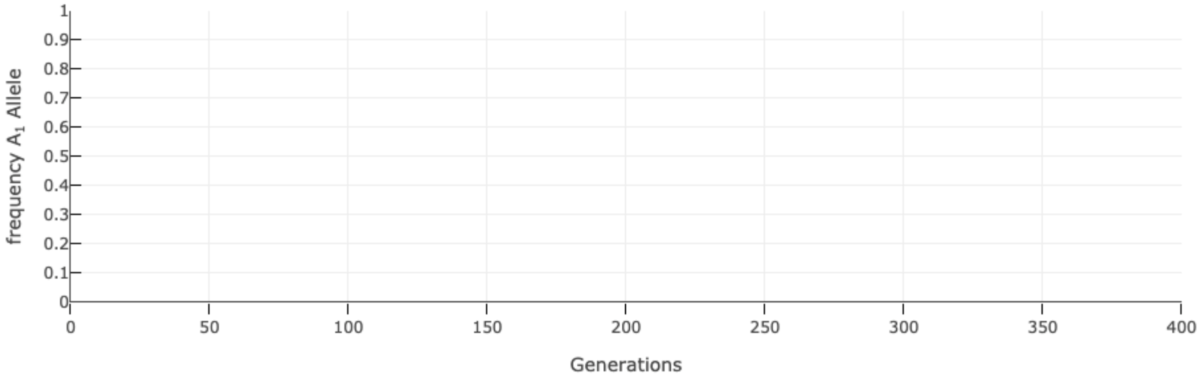
1. Open a web browser with FLASH enabled and navigate to the Web Popgen website. ***Note:*** If the browser on your computer does not work Chrome browser in Apporto does.
2. The program includes the following initial parameters set in the menu bar at the top:



Population Data	Allele Frequency	Fitness (w)			Migration	Mutation		Inbreeding	Bottleneck □			
Population: Finite	# of Populations 5 # of Generations 400	A_1A_1	A_1A_2	A_2A_2	None	A_1==>A_2	A_1 <==A_2	F (temporarily disabled)	Start gen	Stop gen	Pop Size	Growth Rate
N = 100	freq A_1 0.5	1	1	1		0	0	0				1

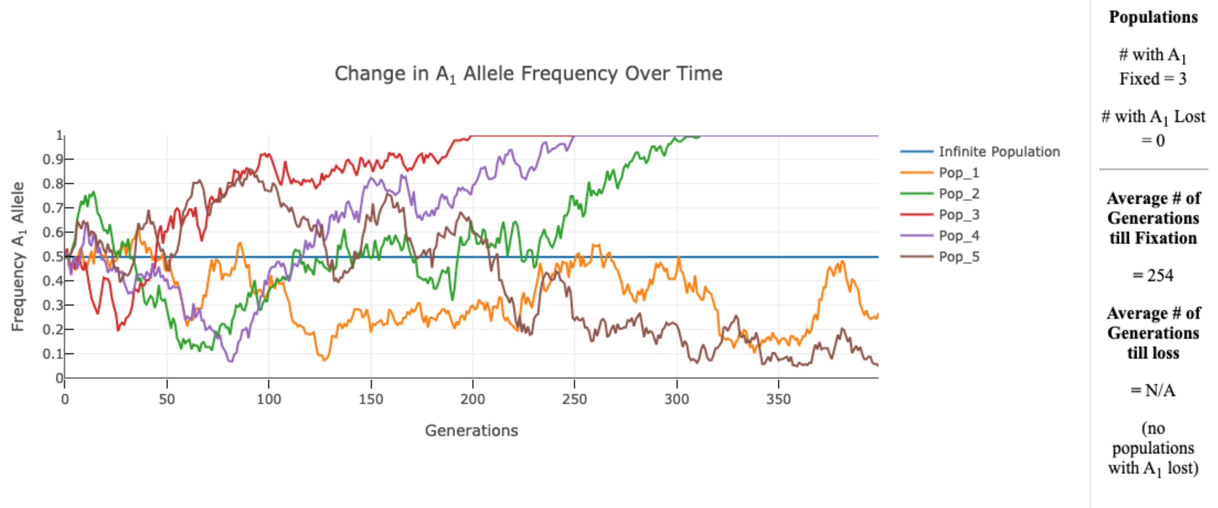
Population size = 100 individuals
Initial Frequency of A_1 allele = 0.5 (50%)
# of replicate populations = 5
Number of Generations = 400
Fitness of A_1 and A_2 are equal

3. There are 2 graphs shown that have will track the allele frequency of A_1 and A_2 for each of the replicate populations



4. Underneath the parameter settings is the "Go" Button. Select it now to run the simulator using the default parameters.

5. This is a stochastic (random) model so everyone's results will vary. Below is one instance.



Populations

with A_1 Fixed = 3

with A_1 Lost = 0

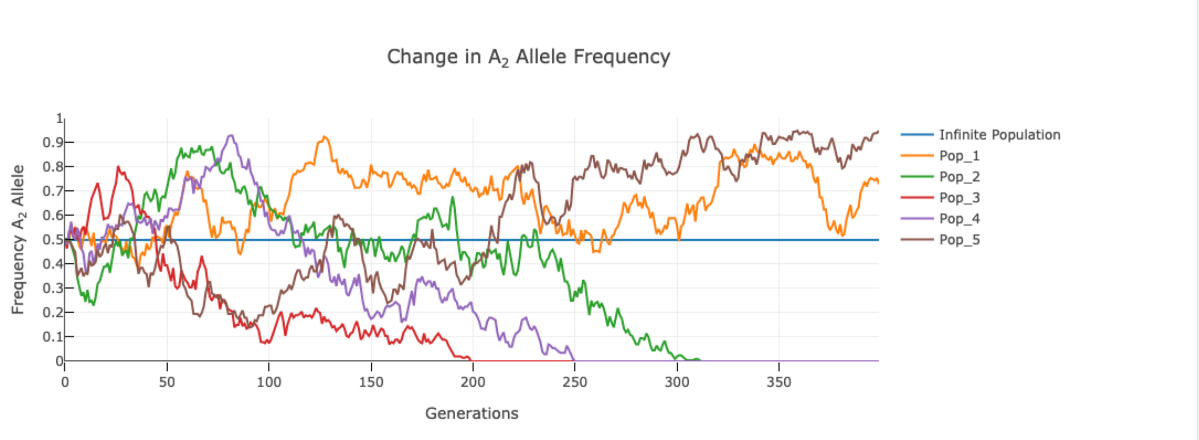
Average # of Generations till Fixation

= 254

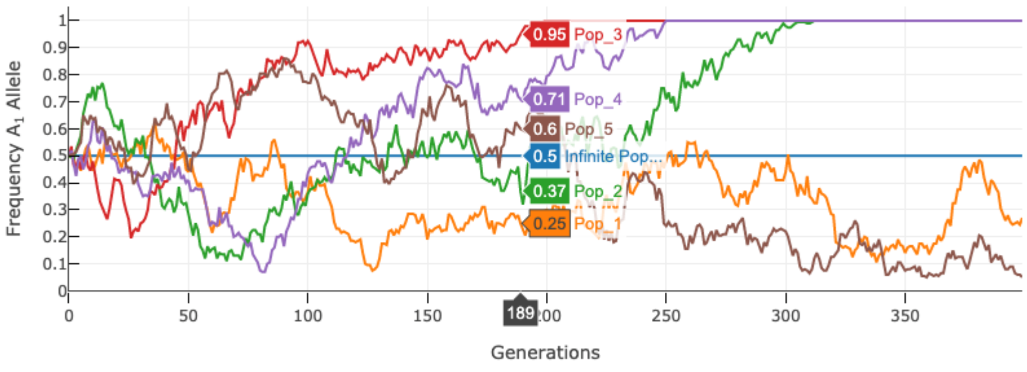
Average # of Generations till loss

= N/A

(no populations with A_1 lost)



6. The program reports the fates for each of the alleles and the mean number of generations for Fixation and Loss.
7. Using your cursor, you can track the allele frequencies of each allele over time.

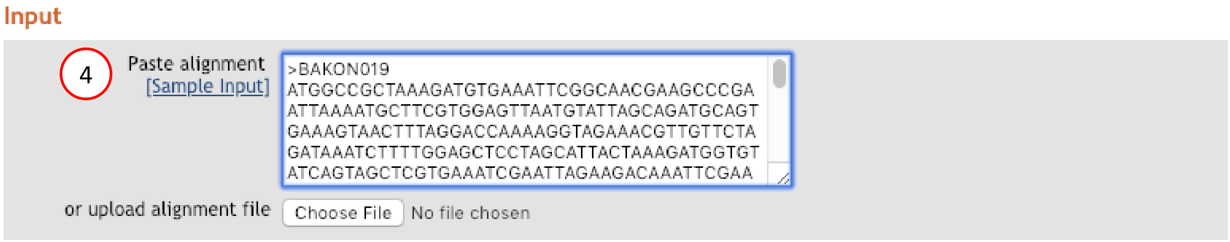


8. To re-run the analysis with the same parameters or new parameters enter them and hit "Go" once again.
9. Use this simulator to answer the questions at the end of the online document.

10. For further details on the program and how it works see the Help/? Webpage: https://www.radford.edu/~rsheehy/Gen_flash/popgen/Popgen_help/index.html

SNAP v2.1.1
Synonymous Non-synonymous Analysis Program
https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html
This program calculates the number of synonymous vs. non-synonymous base substitutions as described in Nei and Gojobori for all pairwise comparisons of sequences in an alignment.

1. From Module #03 folder on Google Drive https://drive.google.com/drive/folders/1myQXFACbmbS2p1oFnijld00cE-RWuusQ?usp=sharing download/open the file groEL.phy
    a. What type of file is it?___
    b. How many sequences?___
    c. How long are they?___
2. Open a web browser and go to the SNAP webpage
3. Cut and paste the entire contents of the groEL.phy file into the "Paste Alignment" window



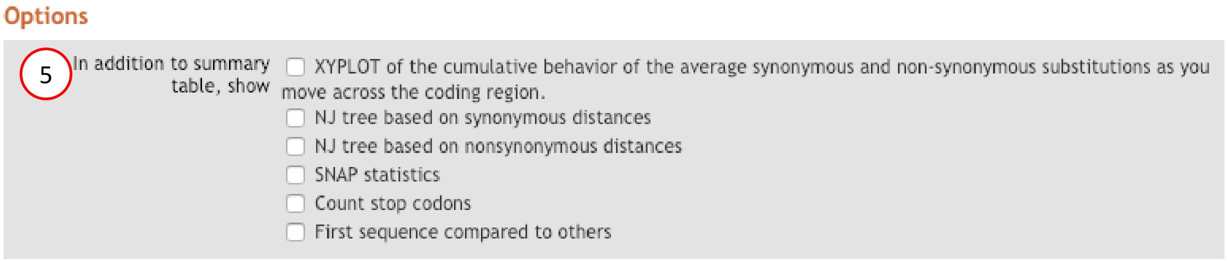
Input

Paste alignment [Sample Input]

```
>BAKON019
ATGGCCGCTAAAGATGTGAAATTCGGCAACGAAGCCCGA
ATTAAAATGCTTCGTGGAGTTAATGTATTAGCAGATGCAGT
GAAAGTAACTTTAGGACCAAAAGGTAGAAACGTTGTTCTA
GATAAATCTTTTGGAGCTCCTAGCATTACTAAAGATGGTGT
ATCAGTAGCTCGTGAAATCGAATTAGAAGACAAATTCGAA
```

or upload alignment file Choose File No file chosen

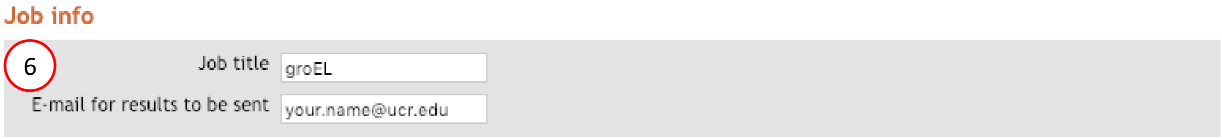
4. For now, options can be left un checked, but feel free to experiment with them on your own time.



Options

In addition to summary table, show
□ XYPLOT of the cumulative behavior of the average synonymous and non-synonymous substitutions as you move across the coding region.
□ NJ tree based on synonymous distances
□ NJ tree based on nonsynonymous distances
□ SNAP statistics
□ Count stop codons
□ First sequence compared to others

5. Enter a job title and your email address



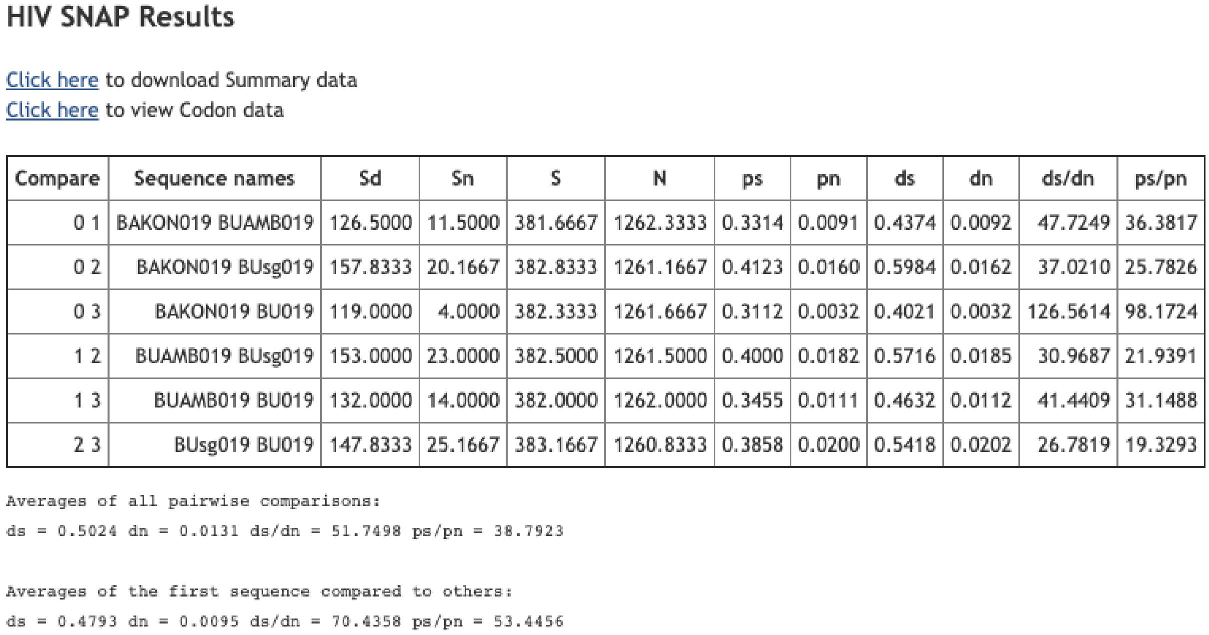
Job info

Job title groEL

E-mail for results to be sent your.name@ucr.edu

6. Click the "Submit" button

7. Results screen should include all possible pairwise comparisons of the sequences in the file:



HIV SNAP Results

Click here to download Summary data
Click here to view Codon data

Compare	Sequence names	Sd	Sn	S	N	ps	pn	ds	dn	ds/dn	ps/pn
0 1	BAKON019 BUAMB019	126.5000	11.5000	381.6667	1262.3333	0.3314	0.0091	0.4374	0.0092	47.7249	36.3817
0 2	BAKON019 BUsg019	157.8333	20.1667	382.8333	1261.1667	0.4123	0.0160	0.5984	0.0162	37.0210	25.7826
0 3	BAKON019 BU019	119.0000	4.0000	382.3333	1261.6667	0.3112	0.0032	0.4021	0.0032	126.5614	98.1724
1 2	BUAMB019 BUsg019	153.0000	23.0000	382.5000	1261.5000	0.4000	0.0182	0.5716	0.0185	30.9687	21.9391
1 3	BUAMB019 BU019	132.0000	14.0000	382.0000	1262.0000	0.3455	0.0111	0.4632	0.0112	41.4409	31.1488
2 3	BUsg019 BU019	147.8333	25.1667	383.1667	1260.8333	0.3858	0.0200	0.5418	0.0202	26.7819	19.3293

```
Averages of all pairwise comparisons:
ds = 0.5024 dn = 0.0131 ds/dn = 51.7498 ps/pn = 38.7923

Averages of the first sequence compared to others:
ds = 0.4793 dn = 0.0095 ds/dn = 70.4358 ps/pn = 53.4456
```

8. This application reports the ratio of dS and dN as dS/dN rather than dN/dS. Using a calculator or Excel dN/dS can be readily calculated.

From the manua,l here are the column contents:

Compare: Lists the two sequences compared, starting with 0 (4 sequences are seqs 0-3)
Sequence_names: The names of the two sequences being compared.
Sd: The number of observed synonymous substitutions
Sn: The number of observed non-synonymous substitutions
S: The number of potential synonymous substitutions (the average for the two compared sequences)
N: The number of potential non-synonymous substitutions (the average for the two compared sequences)
ps: The proportion of observed synonymous substitutions: Sd/S
pn: The proportion of observed non-synonymous substitutions: Sn/N
ds: The Jukes-Cantor correction for multiple hits of ps
dn: The Jukes-Cantor correction for multiple hits of pn
ds/dn: The ratio of synonymous to non-synonymous substitutions

The complete manual is here:
https://www.hiv.lanl.gov/content/sequence/SNAP/help_files/README.html

Note that if **ps** or **pn** has a value >= .75, saturation has been reached and a Jukes-Cantor transformation cannot be done, so the value of NA is returned.

Also, if either **ds** or **dn** is NA or 0, the **ds/dn** ratio is not calculated.

Questions:

1. Complete the following simulations and record the results in the table below. (5)

Pop size	Init. A_1 Freq.	#Pop	# Gen	No. Fixed	No. Lost	Expected Time to Fixation	Initial Prob. Of Loss
1000	0.01	5	200	0	5	4000	.99
1000	0.10	5	200	0	0	4000	.90
1000	0.25	5	200	0	0	4000	.75
1000	0.50	5	200	0	0	4000	.50

Numbers of fixed and lost might vary… However, most should persist for this population size.

Hint: Prob. Fixation + Prob. Loss = 1

2. Do the observed numbers of Fixed and Lost alleles in the simulations correspond to your expectations? Explain why or why not. (7)

Not enough time/generations elapses for many to be lost, However, starting with a very low frequency 10 in 1000 individuals -f the allele A1 tends to get lost.

3. If you increase the number of generations to the “expected time to fixation” are you guaranteed to see fixation of an A_1 allele? Explain why or why not. (7)

No, this is an average estimated time and 5 mutations in 5 replicate populations are being examined. More often than not most mutations will be lost because there is a low probability of them ever becoming fixed.

4. Complete the following simulations and record the results in the table below. (5)

Pop size	Init. A_1 Freq.	#Pop	# Gen	No. Fixed	No. Lost	Expected Time to Fixation	Initial Prob. Of Loss
25	0.01	5	200	0	5	100	0.99
25	0.10	5	200	0	5	100	0.90
25	0.25	5	200	0	5	100	0.75
25	0.50	5	200	2	3	100	0.50

5. Do the observed numbers of Fixed and Lost alleles in the simulations correspond to your expectations? Explain why or why not. (7)

None fixed except when starting with 50% in this small population size. Very low frequency alleles have trouble sweeping in population sizes this small

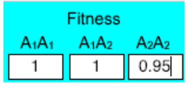
6. Describe how these results compare to the results from Question 1? (7)

Oscillations are much more volatile generation to generation. Larger populations sizes see smaller jumps in the line. Most mutations were lost or fixed in all situations, whereas the mutations were largely maintained in the large population after the allele frequency increased past 10%.

7. Complete the following simulations and record the results in the table below. (5)

Pop size	Init. A_1 Freq.	#Pop	# Gen	A_2A_2 fitness	No. Fixed	No. Lost	Expected Time to Fixation
25	0.01	5	200	0.95	0	5	**156.5**
250	0.01	5	200	0.95	0	4	**248.6**
1000	0.01	5	200	0.95	0 (4 are close)	1	**304.0**
25	0.01	5	200	0.90	2	2	**78.2**
250	0.01	5	200	0.90	1	3	**124.3**
1000	0.01	5	200	0.90	0 (5 are close)	0	**152.0**
25	0.01	5	200	0.50	3	2	**15.6**
250	0.01	5	200	0.50	5	0	**24.9**
1000	0.01	5	200	0.50	5	0	**30.4**

Fill in the A_2A_2 fitness like this:



Fitness

A_1A_1 1

A_1A_2 1

A_2A_2 0.95

8. In each simulation above A_2A_2 had a fitness disadvantage compared to A_1A_1 and A_1A_2. Does having a beneficial allele like A_1 guarantee that it will become fixed? What conditions make it more or less likely? (7)

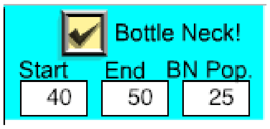
No, there is still no guarantee of fixation. Small population sizes increase risk of loss of the allele. Even in pop of 250 an allele with 10% increase in fitness can be lost. Larger populations and Larger differences in *s* will increase the likelihood of being fixed.

Also, students might note that after A1 reaches≥90% it may take a surprising amount of time to actually reach 100%. This is common. Theoretically the model shown in the infinite population the line asymtopes.

EC #1

Run the following simulations but add a bottleneck at generation #40-50 of 25 and fill in the following table.

Fill in the bottleneck parameters like this:



☑ Bottle Neck!

Start 40

End 50

BN Pop. 25

Pop size	Init. A_1	#Pop	# Gen	A_2A_2 fitness	No. Fixed	No. Lost
1000	0.01	5	200	0.95	0 (3 are ≥60%)	2
1000	0.01	5	200	0.90	0 (5 are close)	0

1000	0.01	5	200	0.50	3 (2 are close)	0

Describe your observations (+2):

Bottleneck disrupts expected increase in frequency. Larger the *s* the more likely it will still reach fixation. Very large allele frequency changes during the 10 generations.

9. Open each of the files `orf1b.phy`, `S.phy`, `N.phy` and `orf10.phy` which are available in the Module #03 folder on Google Drive. These are files contain alignments of homologous gene sequences from 4 COVID-19 strains. Fill in the table below with: How many sequences are there? How long is the alignment in NT? How many codons? Are there any gaps? If so, how many? (5)

	No. of Seqs	No. of NT	No. of Codons	Gaps?	How many?*
orf1b	4	8088	2696	No	0
S	4	3831	1277	Yes	OG=1, A=3 D=2, O=3
N	4	1260	420	Yes	O=1
orf10	4	117	39	No	0

*Be flexible here, gap opening I think is most intuitive, # of –/nt should be acceptable, or # of codons.

10. Analyze each file with SNAP using the default parameters. Record the dN, dS and calculate the dN/dS values below for each comparison (10)

	_OG vs _A			_OG vs _D			_OG vs _O		
	dN	dS	dN/dS	dN	dS	dN/dS	dN	dS	dN/dS
orf1b	0.0005	0.0017	0.29	0.0008	0.0006	1.33	0.0003	0.0006	0.50
S	0.0027	0.0012	2.25	0.0034	0.0012	2.83	0.0109	0.0036	3.03
N	0.0057	0.0052	1.10	0.0052	0	n.a.	0.0031	0.007	0.44
orf10	0	0	n.a.	0	0	n.a.	0	0	n.a.

11. List any gene comparisons that have evidence for purifying selection? (6)

Orf1b OG/A, maybe OG/O
N OG/O
Maybe orf10 – zero changes… can't tell reliably using dN/dS because gene is so short (partial credit)

12. List any gene comparisons that have evidence for neutral evolution? List them. (6)

maybe Orf1b OG/D, and or OG/O
N OG/A (1.1 is pretty close to 1)

13. List any gene comparisons that have evidence for positive selection? List them. (6)

S w/ all 3 comparisons
maybe Orf1b OG/D

14. Do any gene comparisons have evidence of being “saturated”? If so list them. (6)

No. Nothing above 0.75

15. Which COVID variants is the most divergent (different)? Alpha (_A) ? Delta (_D)? or Omicron (_O)? Explain your rationale for your decision. (11)

Omicron has greatest average dN and dS
Consider other rational responses.

Note: Below are NCBI accessions of the genome sequences from which nt gene sequences used in this analysis were retrieved.

Suffix	Type	Genome Nucleotide Accession	Geo Location	Collection Date	Isolate Name
_OG	original	MT027064.1	North America; USA: CA	1/29/20	SARS-CoV-2/human/USA/CA-CDC-03040142-001/2020
_A	ALPHA	MZ394583.1	Africa; Djibouti: Camp Lemonnier	1/20/21	SARS-CoV-2/human/DJI/NAMRU3_C681/2021
_D	DELTA	OK457061	USA: New York	9/22/21	SARS-CoV-2/human/USA/NY-CDC-LC0293029/2021
_O	OMICRON	OM212472.1	Asia; Hong Kong	11/14/21	SARS-CoV-2/human/HKG/HKU-691/2021

Extra Credit #2 Refer to the genome sequence from two weeks ago in Module #01, MT027064, for the functional/protein product predictions for these 4 genes.

A. Does the size of the genes have any relationship the predicted type of selection acting on the genes? Explain why or why not this might be the case. (+2 pts)

Orf10’s small size is no doubt influencing the ability to detect mutations. Much smaller mutational target. Small proteins have fewer overall sites and calculations can become somewhat unreliable. Also, validation of small proteins is sometimes suspect – much like the issue last week with Artemis predicting many more small ORFS then were present in the actual annotation.

Other genes unlikely to be affected.

B. Do the protein product predictions have any relationship the predicted type of selection acting on the genes? Explain why or why not this might be the case. (+2 pts)

For orf1B, N, and orf10 # of mutations is pretty low, which can make estimates of dN dS somewhat problematic. Even for orf1B which is pretty big!

Orf10 is strictly hypothetical – no surprise there.
S is the surface glycoprotein – the antigenic target of immune response so yeah – positive selection!
Orf1B is part of core poly protein and N nucleocapsid phosphoprotein – mainly purifying (maybe neutral) is understandable for core functionalities.