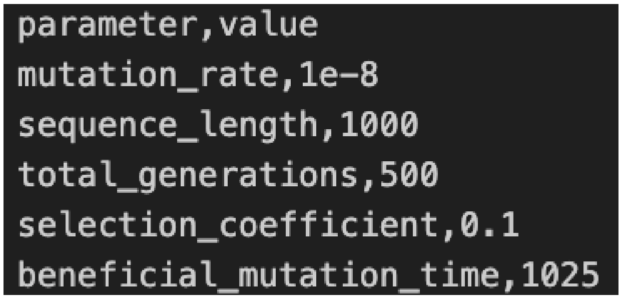
Aaron Morales
BME 205
Fall 2025
Assignment 8

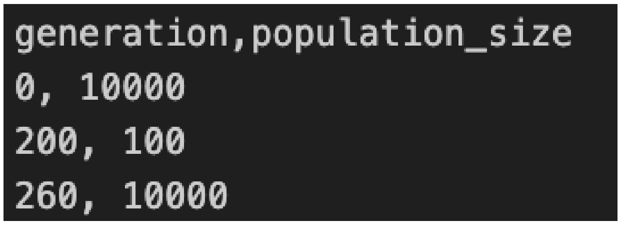
BME 205 Assignment 8 Explanation

Wright-Fisher

I found the best way to determine the efficacy of my Wright-Fisher(WF) algorithm and to verify the associated outputs of my program, it was best to run a smaller simulation. In the hopes that tuning the config and demography parameters might help prove the correctness and biological realism of the WF model.

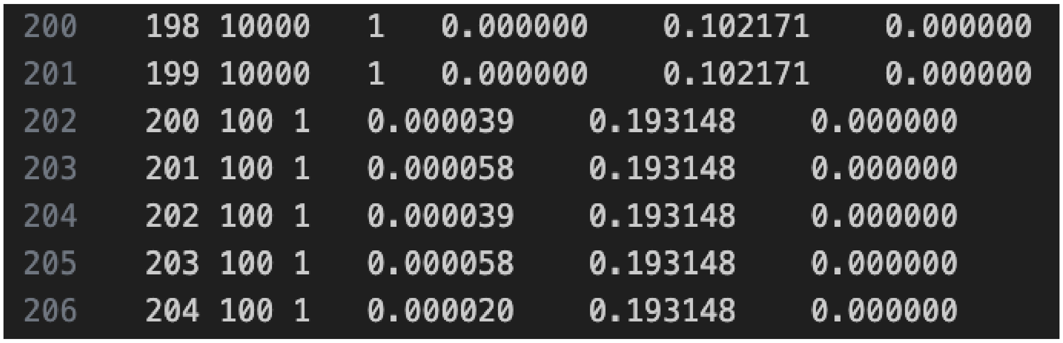


```
parameter,value
mutation_rate,1e-8
sequence_length,1000
total_generations,500
selection_coefficient,0.1
beneficial_mutation_time,1025
```



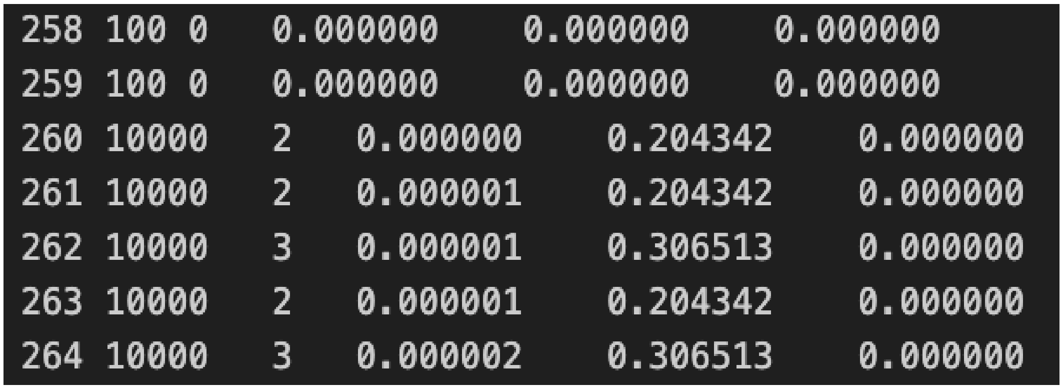
```
generation,population_size
0, 10000
200, 100
260, 10000
```

The results and respective parameters clearly demonstrate, at the onset of the bottleneck, a sudden and significant increase in the number of segregating mutations, nucleotide diversity, and theta-Watterson estimate. This is likely due to a small subset of previously rare alleles being selected by pure chance.



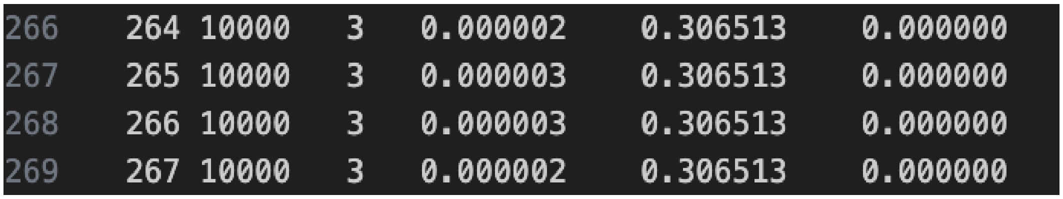
```
198 10000   1   0.000000    0.102171    0.000000
199 10000   1   0.000000    0.102171    0.000000
200 100 1   0.000039    0.193148    0.000000
201 100 1   0.000058    0.193148    0.000000
202 100 1   0.000039    0.193148    0.000000
203 100 1   0.000058    0.193148    0.000000
204 100 1   0.000020    0.193148    0.000000
```

Though as the bottleneck persists, strong genetic drift rapidly eliminates these otherwise fringe alleles slowly eliminating these initially strong variations, this is reflected in the falling Watterson's estimate and nucleotide diversity.



```
258 100 0   0.000000    0.000000    0.000000
259 100 0   0.000000    0.000000    0.000000
260 10000   2   0.000000    0.204342    0.000000
261 10000   2   0.000001    0.204342    0.000000
262 10000   3   0.000001    0.306513    0.000000
263 10000   2   0.000001    0.204342    0.000000
264 10000   3   0.000002    0.306513    0.000000
```

Once the population size returns to ancestral values, the new mutations begin to accumulate and allele frequencies stabilize. Producing a gradual rise in Watterson's, nucleotide diversity, and positive selection of traits.



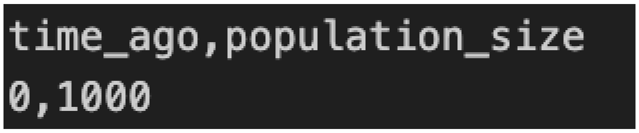
```
264 10000   3   0.000002    0.306513    0.000000
265 10000   3   0.000003    0.306513    0.000000
266 10000   3   0.000003    0.306513    0.000000
267 10000   3   0.000002    0.306513    0.000000
```

This example in particular doesn't show the beneficial allele rising in frequency, and this is likely due to it not surviving the genetic dynamics of the bottleneck(but it can also be expected to rise to prominence post-bottleneck).

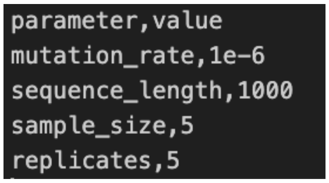
Coalescent Simulation

In order to verify the piecewise-constant coalescent algorithm, I found it helpful to once again develop two baseline tests that can mirror expected results from known config and demography parameters to recognize the differences between constant pop and bottleneck populations.

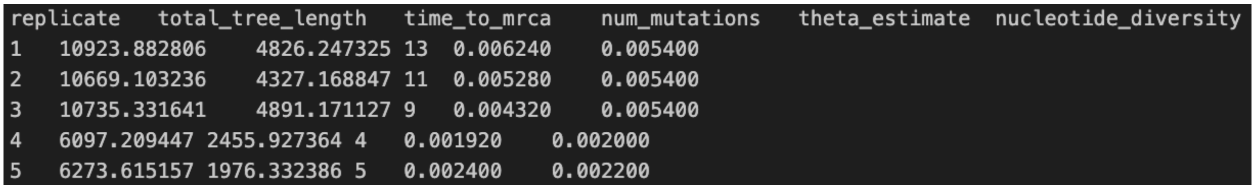
The results for the constant-population test like total tree length, time to mrca, num mutations, theta estimate, and nucleotide diversity all converge around expected values for a stable Wright-Fisher population for expected levels of genetic variation.



```
time_ago,population_size
0,1000
```



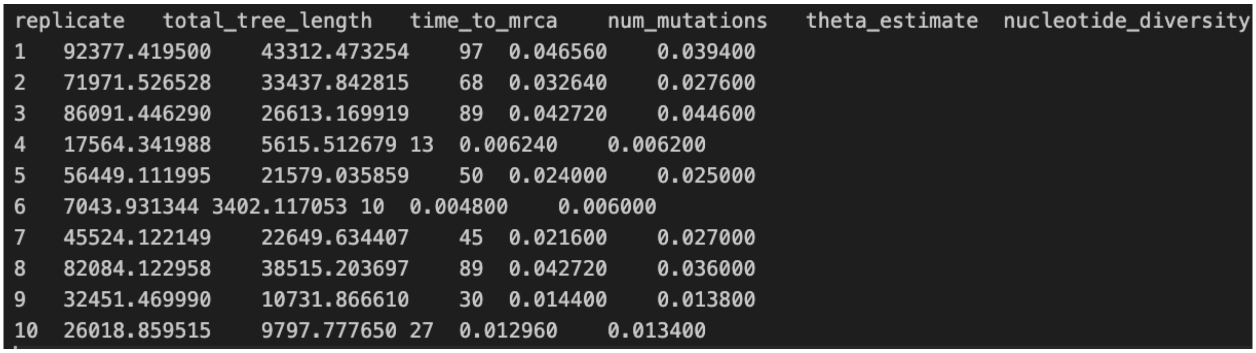
```
parameter,value
mutation_rate,1e-6
sequence_length,1000
sample_size,5
replicates,5
```



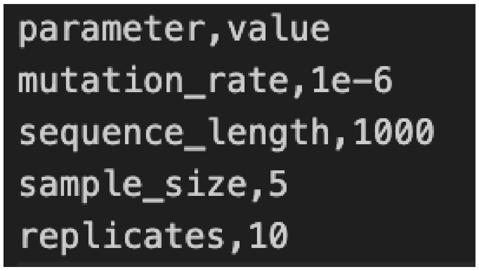
```
replicate   total_tree_length   time_to_mrca    num_mutations   theta_estimate  nucleotide_diversity
1   10923.882806    4826.247325 13  0.006240    0.005400
2   10669.103236    4327.168847 11  0.005280    0.005400
3   10735.331641    4891.171127 9   0.004320    0.005400
4   6097.209447 2455.927364 4   0.001920    0.002000
5   6273.615157 1976.332386 5   0.002400    0.002200
```

Specifically; tree length corresponds to 4 * N * H4(harmonic) ~ 4 * 1000 * 2.083 ~ 8.3k; nucleotide diversity is described by 4 x N x mu ~ 4 * 1000 * 10e-6 ~0.004 (per site); expected time to MRCA = 4N(1 - 1/n) = 4000 * (1-0.2) = 3.2 k generations

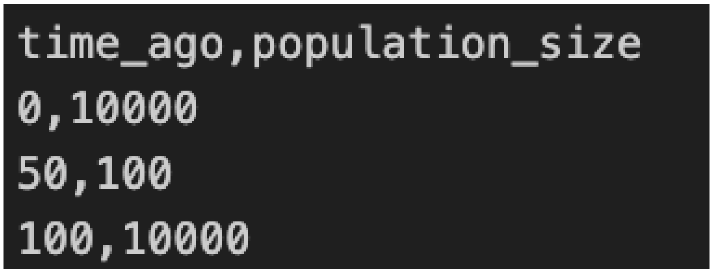
On the other hand, the bottleneck test produces very different signatures. We see tree lengths shrink significantly, MRCA occurs much more recently, mutation counts drop to near zero, and both theta Watterson and nucleotide frequency collapses.



```
replicate   total_tree_length   time_to_mrca    num_mutations   theta_estimate  nucleotide_diversity
1   92377.419500    43312.473254    97  0.046560    0.039400
2   71971.526528    33437.842815    68  0.032640    0.027600
3   86091.446290    26613.169919    89  0.042720    0.044600
4   17564.341988    5615.512679 13  0.006240    0.006200
5   56449.111995    21579.035859    50  0.024000    0.025000
6   7043.931344 3402.117053 10  0.004800    0.006000
7   45524.122149    22649.634407    45  0.021600    0.027000
8   82084.122958    38515.203697    89  0.042720    0.036000
9   32451.469990    10731.866610    30  0.014400    0.013800
10  26018.859515    9797.777650 27  0.012960    0.013400
```



```
parameter,value
mutation_rate,1e-6
sequence_length,1000
sample_size,5
replicates,10
```



```
time_ago,population_size
0,10000
50,100
100,10000
```

Comparing the results from empirical values, tree length ~ 83.3k, num mutations ~ 83.3, nucleotide diversity ~ 0.04 per site. However, in and around bottleneck events, the expected tree length will be much smaller than the empirical 83.3k and this is reflected in replicate 6 wherein the tree length is a mere 7k, the number of mutations is 10, and nucleotide diversity is 0.006!

Each of these tests and results confirms the efficacy of the model in reflecting a severe loss of genetic diversity expected during bottleneck events and the correlation of our output values in response to these changes.