Methodology

Our first challenge when analyzing this data set was to determine where each gene started and ended. This was the challenge we spent the most time brainstorming how to solve, as it wasn't immediately clear. Ultimately, we ended up determining the positions where the noncoding mutations were located, since these mutations only occur at positions outside of a gene. However, we quickly ran into another problem, based on the noncoding mutations, we were detecting over 50 genes in the simulation, which we believed based on the context was too many. So, we tweaked our gene detection script to combine certain genes that were a certain length apart. Using this method, we were able to tweak the script to detect a more reasonable amount of genes, settling on finding eight. Once we were satisfied with this result, we then ran several DN/DS and PN/PS tests, one on the sample data as a whole, and then a DN/DS and a PN/PS test for each individual gene. For each test, we actually ran twice, swapping the two populations for which one was the sample and which one was the divergence. For several individual genes, we got conflicting results between our DN/DS and PN/PS tests, which we noted in the initial data report.

Actual genes (16 total):

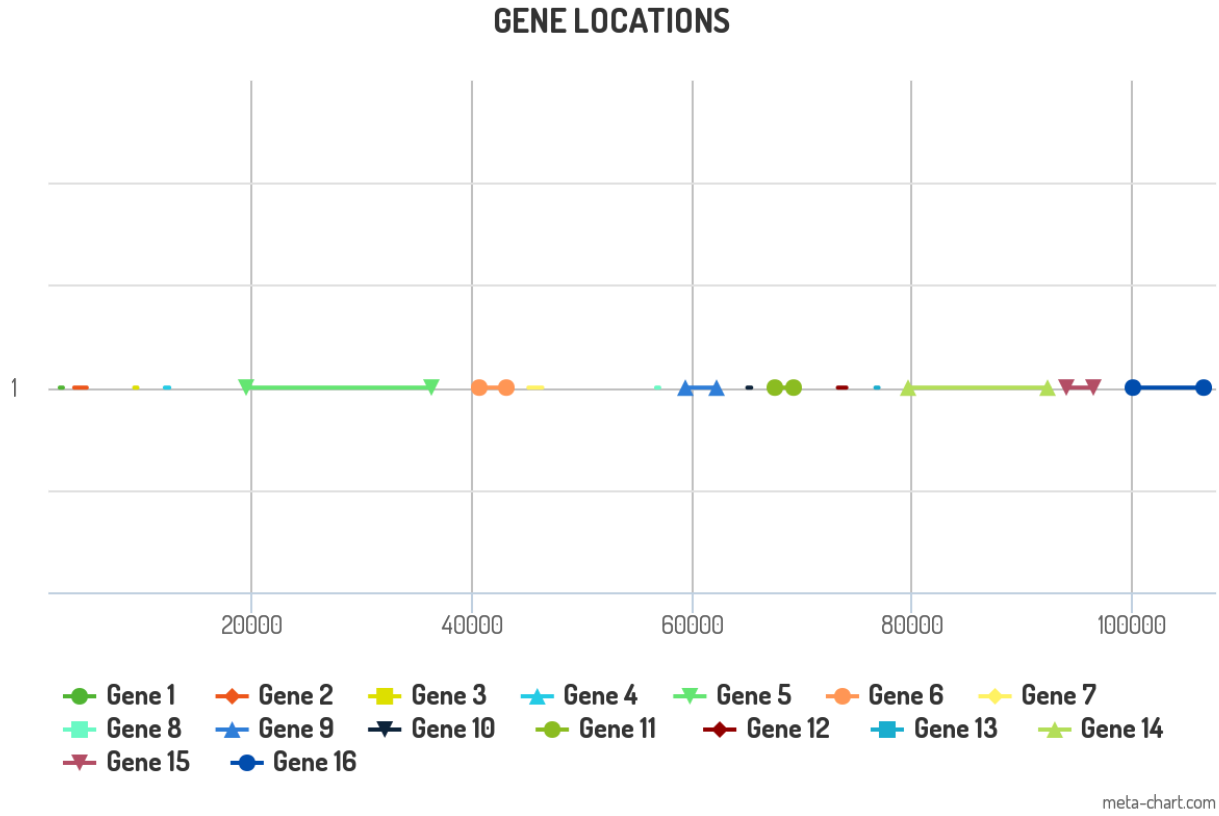
Start	End	Positive
1014	7141	0
8753	10296	0
12192	14590	0
19094	24449	1
24974	36217	1
40582	**47007**	**0**
50878	59674	0
60030	62240	0
64370	65203	0
67658	69085	1
73302	75535	0
76080	76762	1
79775	92240	1
94185	96317	1
98421	99115	0
99979	**106540**	**0**

Predicted gene ranges:

Gene 1: from 2526 to 2800
Gene 2: from 3837 to 4966
Gene 3: from 9289 to 12462
Gene 4: from 19625 to 36291-- multiple genes combined
Gene 5: from **40732 to 46330**
Gene 6: from 56673 to 69194
Gene 7: from 73173 to 96442 -- multiple genes combined
Gene 8: from **100044 to 106478**

After adapting our script to have a lower range in which to combine what we initially determined to be genes:
Gene 1: from 2526 to 2800
Gene 2: from 3837 to 4966
Gene 3: from 9289 to 9563
Gene 4: from 12065 to 12462
Gene 5: from 19625 to 36291
Gene 6: from **40732** to 43087
Gene 7: from 45077 to **46330**
Gene 8: from 56673 to 56947
Gene 9: from **59364 to 62233**
Gene 10: from **64953 to 65299**
Gene 11: from **67528 to 69194**
Gene 12: from 73173 to 73935
Gene 13: from 76585 to 76859
Gene 14: from **79653 to 92355**
Gene 15: from **94048 to 96442**

Gene 16: from **100044 to 106478**



The positions in bold were all close to the actual gene positions, usually within +/- 500. There were multiple genes that we had found which turned out to actually be a single gene. Additionally, even though we had the correct quantity of genes now, there were still some in the actual data that we missed entirely in our prediction. An example of this would be the positions from 5000 to 7000 where there was a gene present but we did not find it. For some reason, our prediction seems to also hold up better towards the middle and end of the genome as opposed to the start. For the ones we did get correct they were all fairly close to the actual data provided.