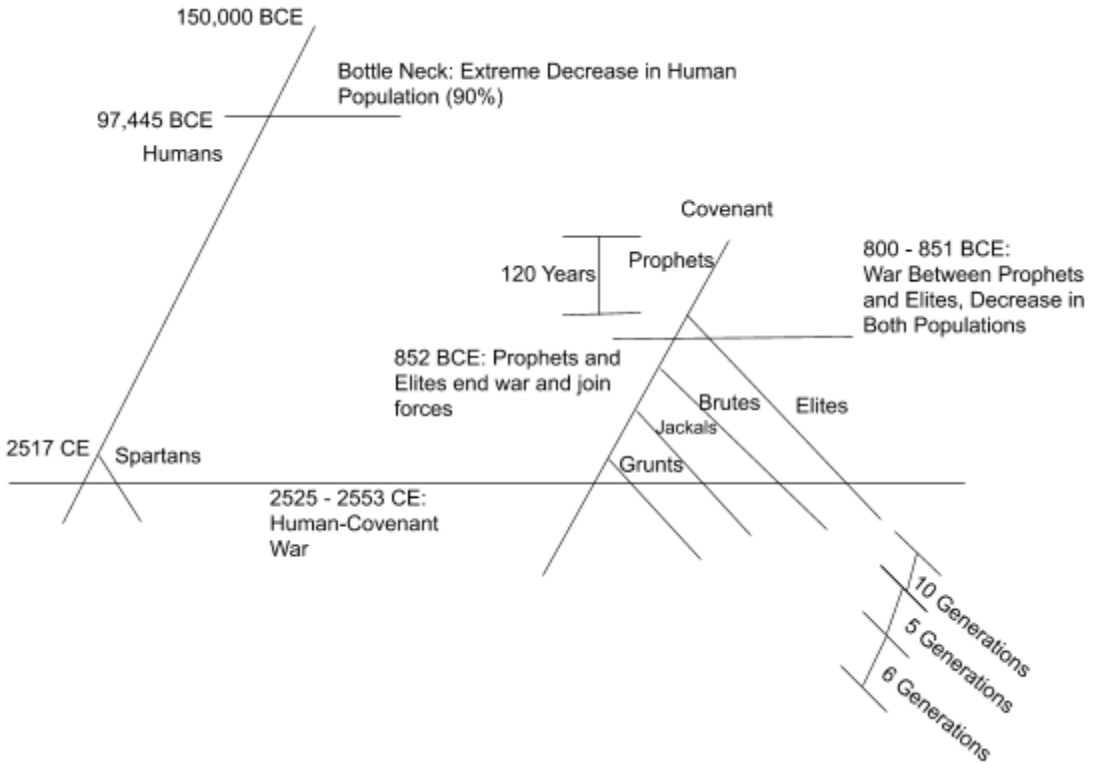
Universe Description



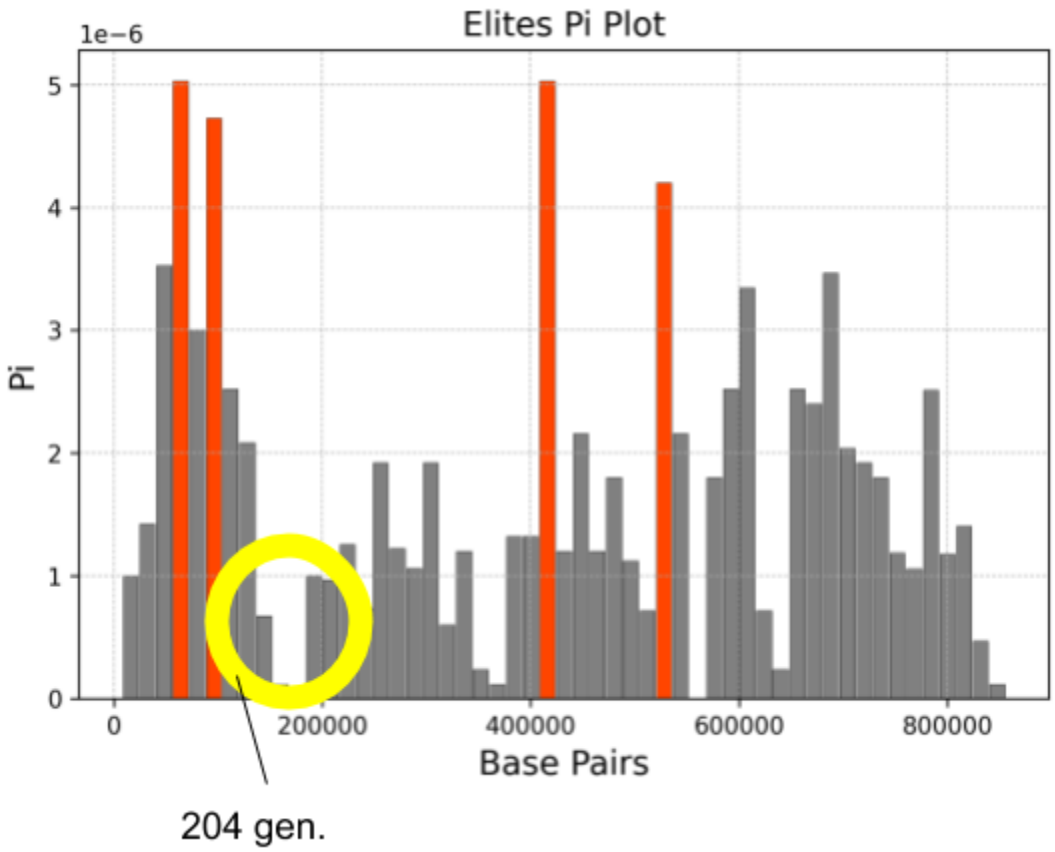
Our Universe takes place in the Halo Universe where there are two different species at war, the Humans and the Covenant. To respond to the threat of the Covenant, the Human race creates genetically-enhanced supersoldiers called Spartans. The Covenant contains the Prophet species which are the leaders of the Covenant and also contain the subspecies of Elites, Brutes, Jackals, and Grunts.

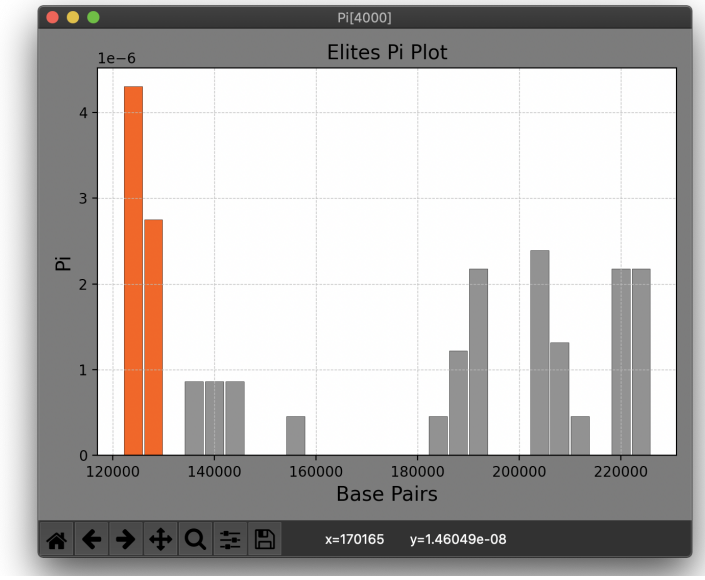
ESTIMATES OF SWEEP:

Elites:

recombination rate: 2*10^-7

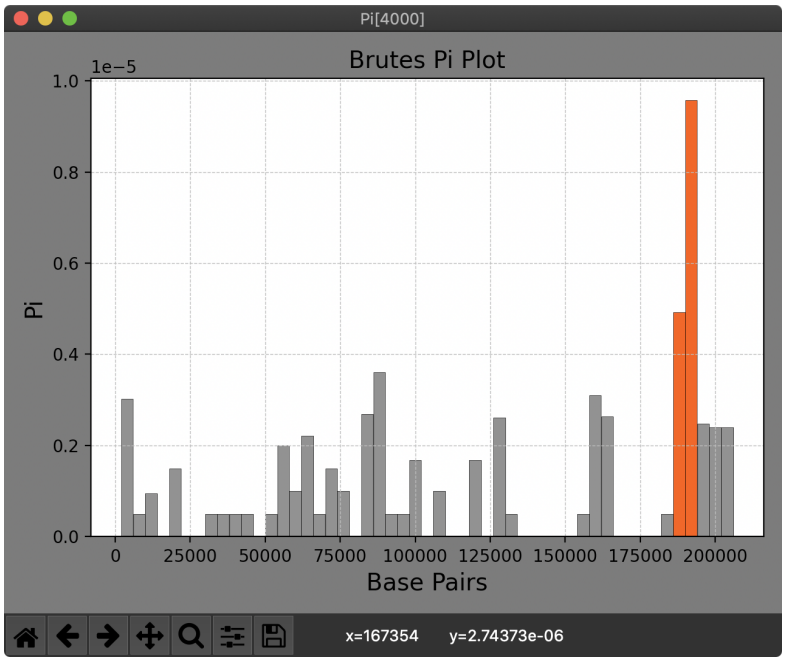
l = 187,000- 170,000 = 17k





This sweep seems to have been a stronger one as it significantly affects the surrounding pi values before they return to their baseline. Our estimate of 204 generations would also fit into this since a newer sweep would mean that the lower diversity would cover a wider area. Looking at Figure 15 which shows the CLR plot for the elites population, there is also a spike at around 180'000 base pairs signifying that a sweep happened in that location. Since the Pi data and the CLR both independently signify a potential sweep around 170'000 to 190'000 base pairs we are fairly certain that a sweep did happen there.

Brutes:
Ne = -ln(0.5)/(recomb*width)



Sweep 1:

$$Width = 144,000 - 132,500 = 11500$$

$$Recombination = 2 \times 10^{-7}$$

$$Ne \approx 346.57 \approx 347\ generations$$

Looking at the surrounding data this sweep does not seem to have been very strong. It seems likely that there were multiple sweeps that caused the repeated dips in the pi value along this stretch of the brutes chromosome. This also seems to be confirmed by our CLR plot, Figure 16, where there is no CLR change in that area, which could mean that a smaller, harder to detect sweep happened in that area. It is also possible that no selective sweep occurred especially since the pi data is somewhat erratic in the area and the CLR analysis does not confirm it either.

ESTIMATES OF TAJIMA'S D:

Pop Name	Pi	Segregating Sites	Tajima's D
Brutes	1.13*10^-4	239	-3.21
Elites	8.4*10^-4	207	-3.192
Grunts	2.63*10^-5	196	-3.27
Humans	8.55*10^-5	289	-2.95
Jackals	8.86*10^-5	214	-3.26
Prophets	3.69*10^-5	172	-3.1648
Spartans	5.47*10^-5	199	-4.989

ESTIMATES OF PI:

The Pi statistic, or more widely known as Nucleotide Diversity, is used to measure the amount of differences in a population. Pi is calculated by finding the average number of nucleotide differences between DNA sequences for all possible pairs. In other words, it is a measure of genetic variation. Through the SLiM program, we were able to look at how many individuals contained a certain gene and then multiplied that number by how many people didn't have that gene. We continued to do this all the way to the end, and then divided the whole number by (N 2). Later on, we split these into windows through the MatplotLib package in python and were able to arrive with Figures 1-7. Looking at Figures 1-7, there are two clear patterns that the Pi figures show. In Group 1, a pattern appears between prophets, humans and Spartans while in Group 2, a different pattern appears between brutes, elites, jackals, and grunts. The common feature in Group 1 is that the data is unimodal and skewed to the right. This indicates that a selective sweep may have wiped out the genes with the exception of the gene in the beginning. A sweep is basically when a beneficial mutation increases in frequency and becomes fixed. This then leads to a lower genetic variation which is what the graph is showing. In Group 2, the blue bars(the bars that are significantly different from the rest) are much more widespread and vary in length. Interestingly, looking at the number of grey bars and the range, the Brutes and Elites have the highest Pi and range in Group 2, and then follows Jackals, and lastly Grunts. From this order, we can draw that Brutes and Elites first arrived in the timeline, then Jackals, and finally Grunts in the overall time period. Also with these two groups, we know that the group's population arrived from a common ancestor and are thus similar in many ways. Especially looking at Humans and Spartans, the Pi plots look exactly the same but the Human Pi plot has a larger range compared to the Spartans and so we can assume that humans came first and then Spartans followed.

ESTIMATES OF EFFECTIVE POPULATION SIZE:

Pop Name	Ne
Brutes	285
Elites	223
Grunts	66
Humans	216
Jackals	222
Prophets	91
Spartans	139

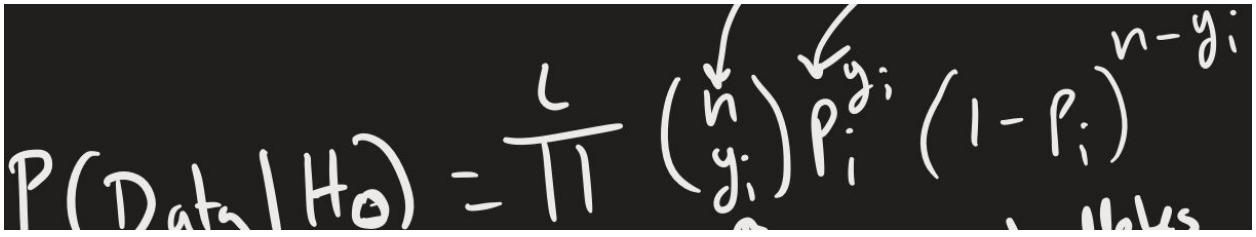
To get these values we first used our previously calculated Pi values across the respective populations entire genomes and then divided that value by 4 times the mutation rate which in for this analysis was a constant we knew to be 10^-7.

Looking at some specific examples, these values seem to make sense. We know that in our simulation, during the latter half where all of the recombination occurred humans and spartans only migrated between each other. Humans had a large population size and spartans by far the smallest which means that the overall pool of genomes would be smaller than for other groups that migrated between each other such as elites and brutes.

ESTIMATES OF DN/DS:

CLR TEST:

The CLR test is a popular genetic tool for detecting genes/ regions of a gene that are subject to either positive or negative sweeps. The formula for calculating CLR is given by this equation.



We then calculated the CLR for a specified base pair window and plotted it by using Python's MatplotLib library. Across all of the subpopulations there is a clear average CLR, but what's more important are the differences. Looking at the data, there are basically two groups that have their own pattern. In Group 1, a similar pattern arises between Spartans, Prophets, Humans, and Elites. In Group 2, a similar pattern arises between Jackals, Grunts and Brutes. With these groupings, it is likely that they each came from the same ancestor. In the CLR plots of Group 1, there is a decrease in CLR towards the beginning of the graph. A consistent decrease supports the assumption made with the PI plots that a selective sweep happened in the population. Looking at Group 1, the Humans and Spartans have very similar CLR plots and the only difference is the range in CLR. The larger CLR in Humans helps support the hypothesis made with the Pi plots that the Spartans arrived after the Humans. A similar hypothesis can be applied with the CLR of Group 2. The number of significant bars and ranges decreases from Elites(having the highest) to Brutes then Jackals and then Grunts. This supports the fact that the Elites first arrived in the timeline and then came Brutes then Jackals and finally Grunts.

FST Data:

Pop Name	**FST**
Brutes	0.5682
Elites	0.6271
Grunts	0.8503
Humans	0.6345
Jackals	0.6281
Prophets	0.8047
Spartans	0.7296

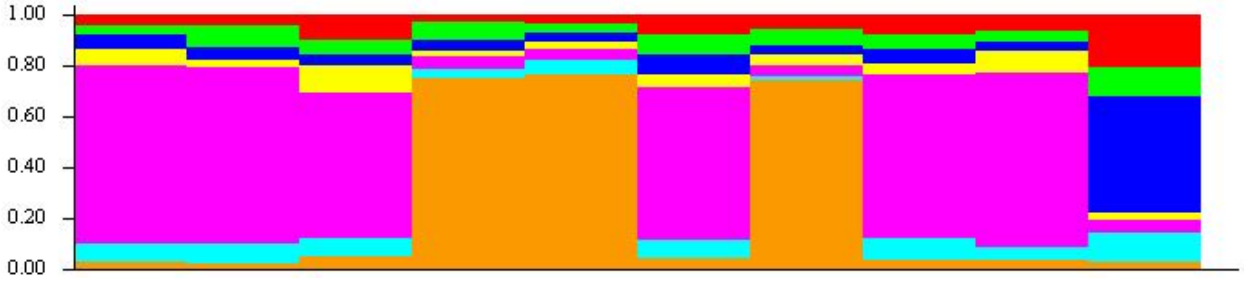
Structure plots:

Spartan population:

Parameter Set	Run Name	K	Ln P(D)	Var[LnP(D)]	α1	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5	Fst_6	Fst_7	Fst_8	Fst_9	Fst_10
p1	p1_run_1	2	-46605.0	91855.2	0.3489	0.6307	0.2521	-	-	-	-	-	-	-	-
p1	p1_run_2	3	-71129.5	141313.0	0.3234	0.7369	0.4816	0.0082	-	-	-	-	-	-	-
p1	p1_run_3	4	-83857.8	166560.3	0.3475	0.0118	0.0194	0.3721	0.6349	-	-	-	-	-	-
p1	p1_run_4	5	-19327.1	36526.4	1.1885	0.0110	0.1500	0.0121	0.0090	0.0073	-	-	-	-	-
p1	p1_run_5	6	-1261.7	233.1	1.7008	0.0099	0.0112	0.0090	0.0134	0.0050	0.0105	-	-	-	-
p1	p1_run_6	7	-88456.7	175723.6	0.2104	0.0095	0.0080	0.0051	0.0114	0.4646	0.0082	0.6017	-	-	-
p1	p1_run_7	8	-28998.1	55991.6	0.5046	0.2424	0.0126	0.0101	0.0113	0.0156	0.0101	0.0066	0.0065	-	-
p1	p1_run_8	9	-1283.0	280.6	1.3512	0.0080	0.0083	0.0187	0.0089	0.0091	0.0082	0.0128	0.0185	0.0180	-
p1	p1_run_9	10	-5188.9	8151.9	0.5584	0.0448	0.0083	0.0068	0.0094	0.0088	0.0134	0.0159	0.0073	0.0107	0.0146

By looking at the data, Ln P(D) is the highest negative number for K=7, meaning structure predicts that K value is most likely. K=4 is the second most likely.

For K=7:

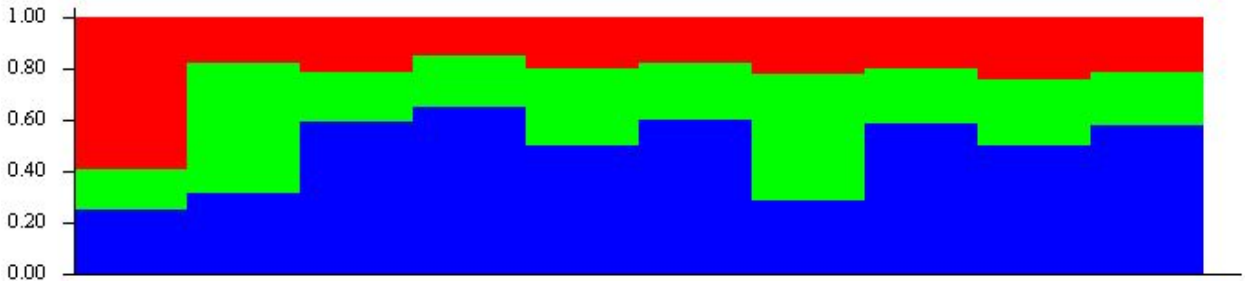


Grunts population:

Parameter Set	Run Name	K	Ln P(D)	Var[LnP(D)]	α1	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5	Fst_6	Fst_7	Fst_8	Fst_9	Fst_10
p1	p1_run_1	2	-9385.2	17216.0	0.3826	0.5331	0.0095	-	-	-	-	-	-	-	-
p1	p1_run_2	3	-11226.8	20748.8	0.5357	0.0118	0.0152	0.3290	-	-	-	-	-	-	-
p1	p1_run_3	4	-1310.0	684.0	0.9925	0.0068	0.0056	0.0134	0.0235	-	-	-	-	-	-
p1	p1_run_4	5	-2104.1	2271.9	1.2231	0.0487	0.0082	0.0096	0.0105	0.0073	-	-	-	-	-
p1	p1_run_5	6	-9632.3	17530.7	0.5614	0.0084	0.0115	0.0134	0.3180	0.0112	0.0071	-	-	-	-
p1	p1_run_6	7	-1061.6	168.0	1.8938	0.0228	0.0330	0.0112	0.0095	0.0407	0.0050	0.0112	-	-	-
p1	p1_run_7	8	-1067.4	179.6	0.9155	0.0144	0.0104	0.0196	0.0077	0.0058	0.0293	0.0093	0.0185	-	-
p1	p1_run_8	9	-1074.4	188.5	0.7640	0.0411	0.0094	0.0113	0.0108	0.0071	0.0167	0.0143	0.0138	0.0094	-
p1	p1_run_9	10	-1100.3	244.6	0.6035	0.0067	0.0071	0.0082	0.0131	0.0106	0.0109	0.0366	0.0074	0.0124	0.0292

By looking at the data, Ln P(D) is the highest negative number for K=3, meaning structure predicts that K value is most likely. K=2 is the second most likely.

For k=2:



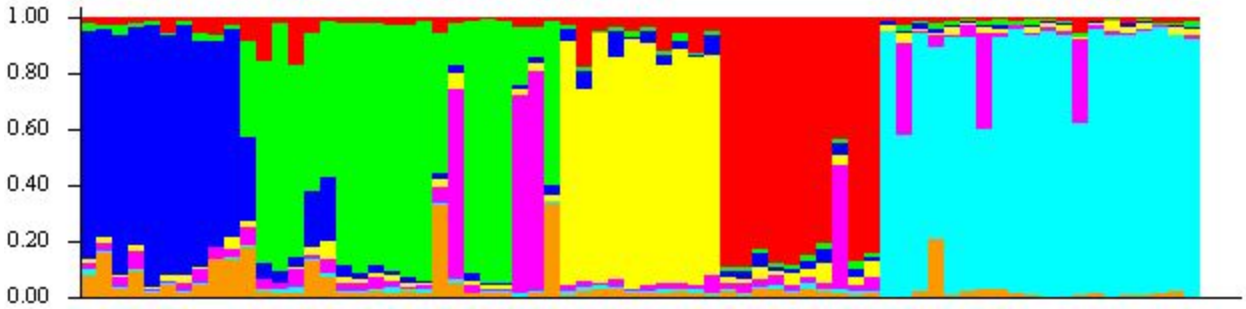
For these two populations, there does seem to be some variance in the data that may suggest subpopulations, especially for the spartan population. This could be a result of having predicted a large K (7), but it could also be because the spartan population was the newest population.

All populations:

Parameter Set	Run Name	K	Ln P(D)	Var[LnP(D)]	α1	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5	Fst_6	Fst_7	Fst_8	Fst_9	Fst_10
p2	p2_run_1	2	-1779278.5	3511850.0	0.1884	0.3448	0.6665	-	-	-	-	-	-	-	-
p2	p2_run_2	3	-4273630.1	8505336.8	0.1406	0.3062	0.2469	0.6728	-	-	-	-	-	-	-
p2	p2_run_3	4	-8831505.8	17624746.2	0.1485	0.6456	0.3303	0.3923	0.5220	-	-	-	-	-	-
p2	p2_run_4	5	-12858006.7	25679238.1	0.2255	0.6447	0.3668	0.1438	0.4364	0.3202	-	-	-	-	-
p2	p2_run_5	6	-14854142.8	29673743.9	0.1371	0.4286	0.4468	0.5407	0.0105	0.4478	0.6182	-	-	-	-
p2	p2_run_6	7	-15624625.2	31213487.3	0.1609	0.4348	0.3894	0.4203	0.3902	0.0162	0.6621	0.0097	-	-	-
p2	p2_run_7	8	-14263137.1	28487288.8	0.2644	0.6386	0.0113	0.0113	0.3322	0.0291	0.3724	0.0109	0.2900	-	-
p2	p2_run_8	9	-15428433.9	30817592.7	0.2311	0.0110	0.0091	0.3872	0.3388	0.1855	0.6432	0.0110	0.2518	0.0112	-
p2	p2_run_9	10	-15188473.7	30337035.4	0.1843	0.1574	0.0100	0.0106	0.6124	0.1757	0.3820	0.0101	0.0107	0.0104	0.2729

By looking at the data, Ln P(D) is the highest negative number for K=7, meaning structure predicts that K value is most likely. This is true for the data, as it contained 7 different populations.

For K=7:



From this graph, it appears the last two populations are fairly separated from the others and are heavily related to each other. These two populations are the humans and the spartans, respectively, which were indeed separated from other populations and closely related, they did not descend from the Covenant races and only migrated between each other. For the remaining races, we see a small section of population 1 in each population, which makes sense given Population 1 was the ancestral population that all other Covenant races descended from. We see a high correlation between Population 2 and Population 3, which is true to the script as these two populations migrated between each other. However, Populations 4 and 5, also migrated between each other, but they show very few signs of similarities or migration. Based on just the data output from structure and not on our actual knowledge of the script, we estimate high migration between Populations 2 and 3 as well as between Populations 6 and 7.

Appendix

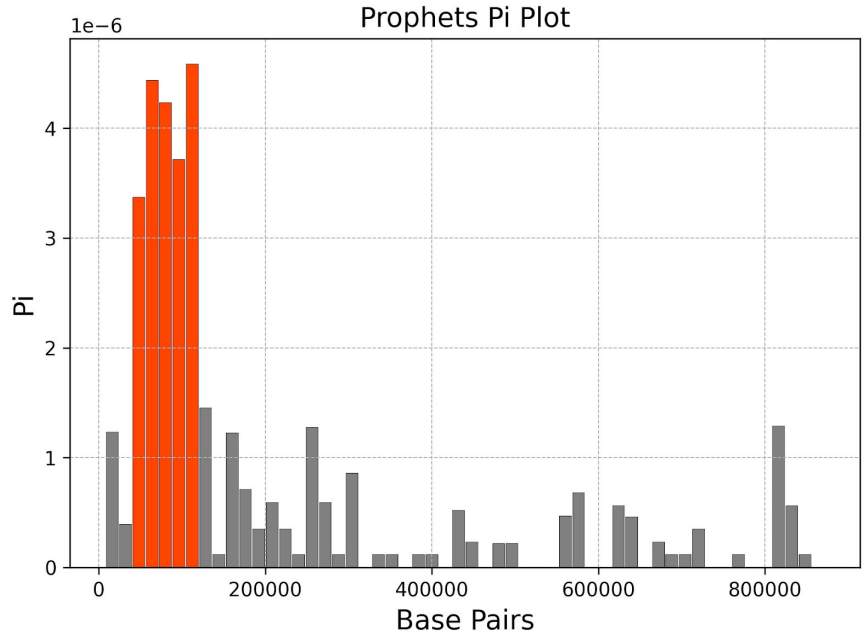


Figure 1: Pi of the Prophet Population across Base Pairs

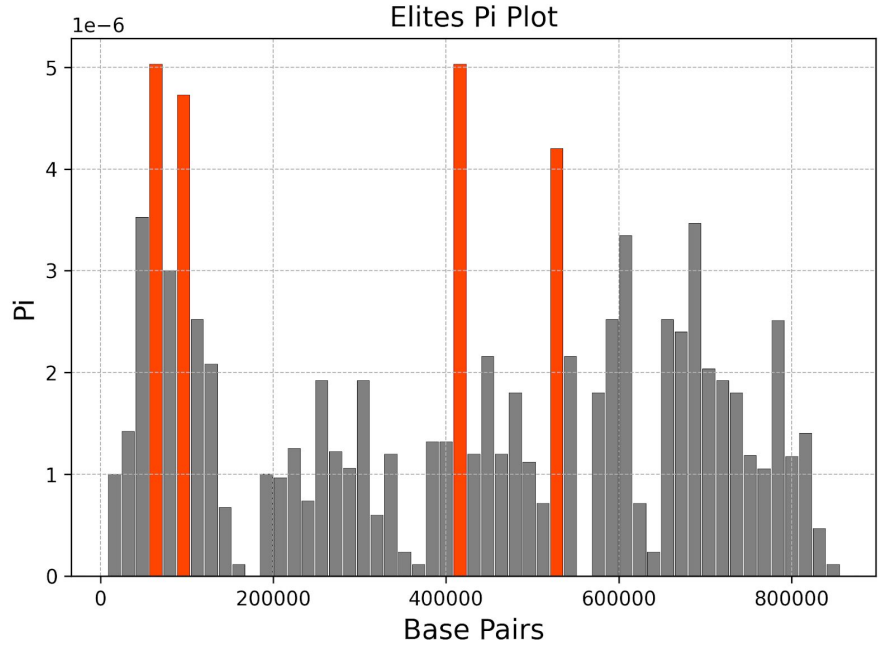


Figure 2: Pi of the Elite Population across Base Pairs

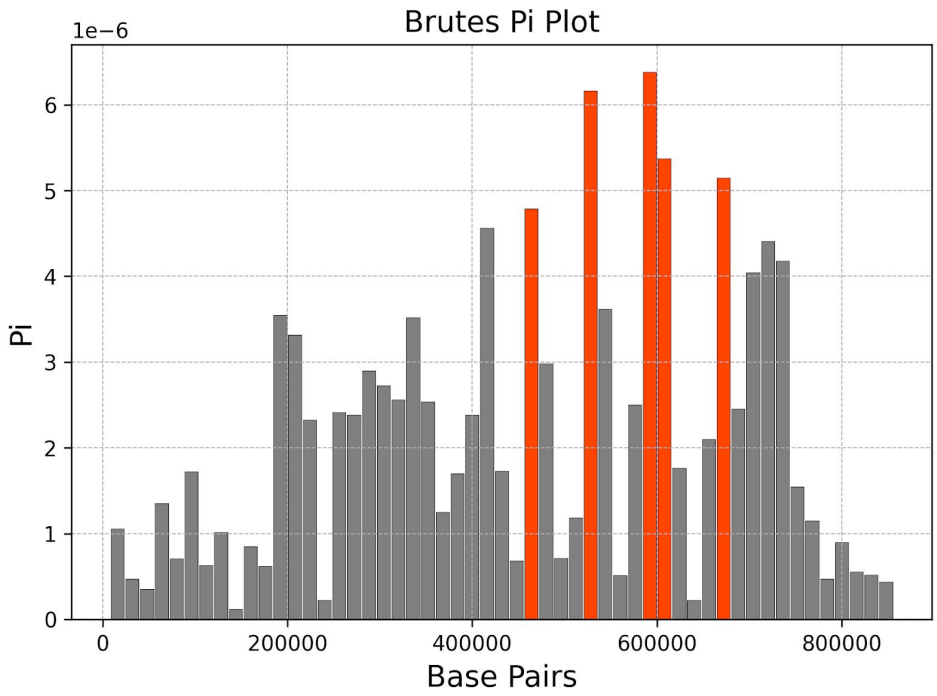


Figure 3: Pi of the Brute Population across Base Pairs

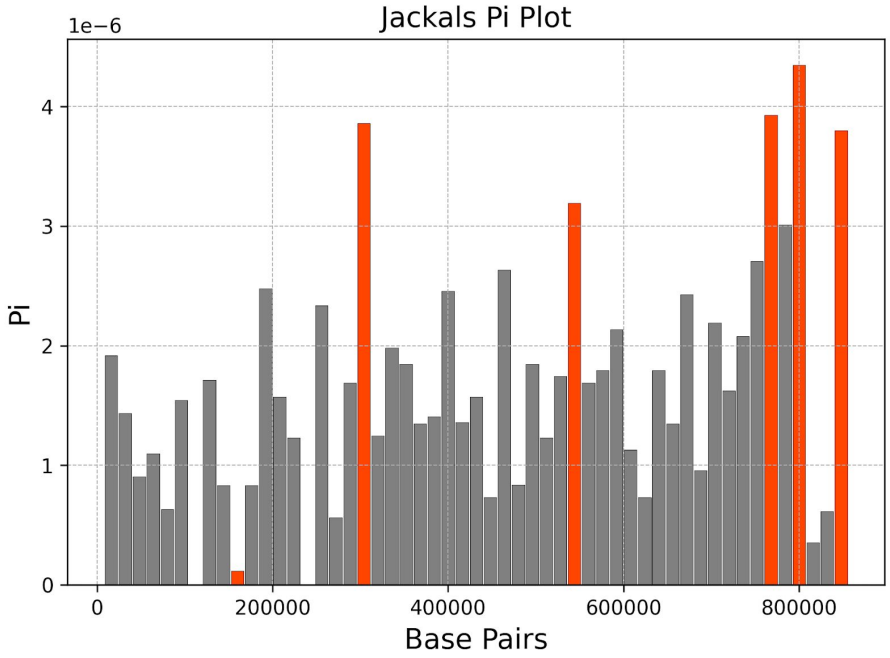


Figure 4: Pi of the Jackal Population across Base Pairs

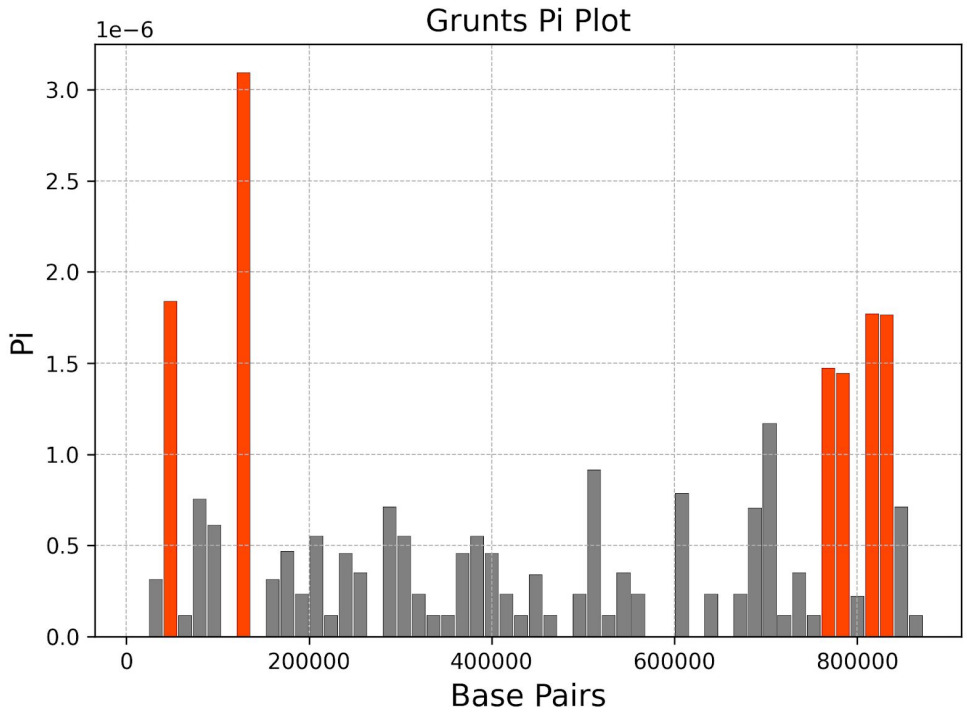


Figure 5: Pi of the Grunt Population across Base Pairs

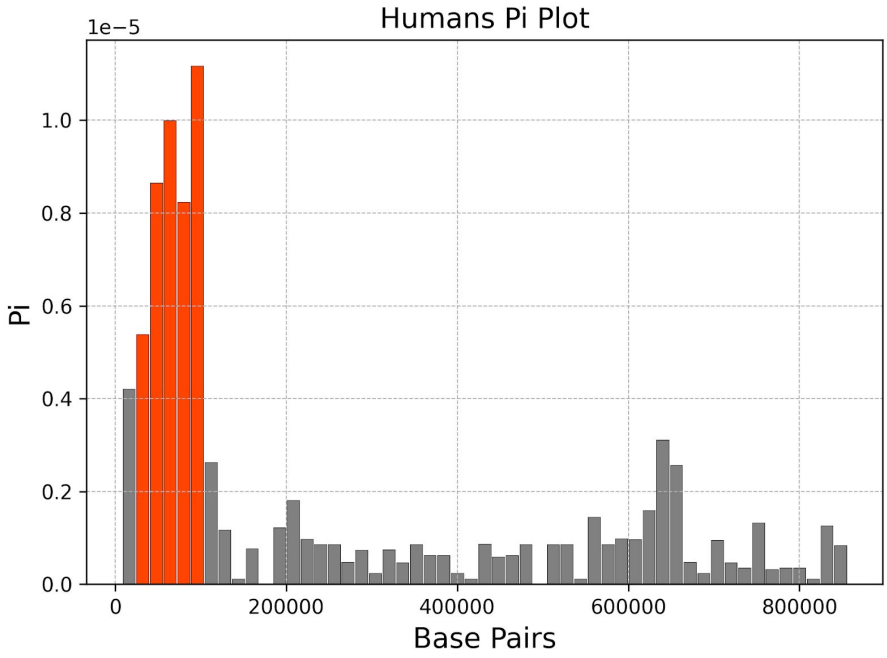


Figure 6: Pi of the Human Population across Base Pairs

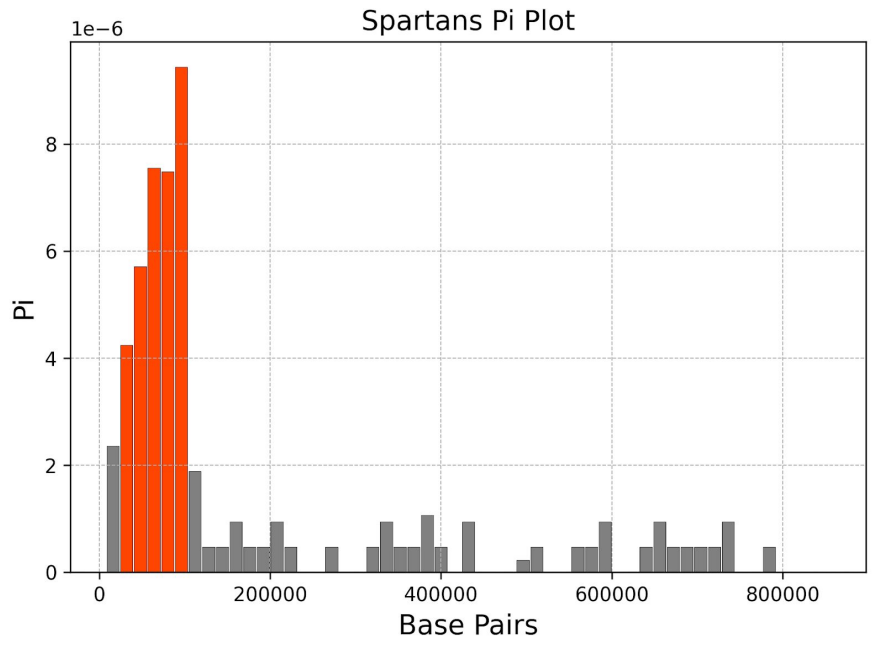


Figure 7: Pi of the Spartan Population across Base Pairs

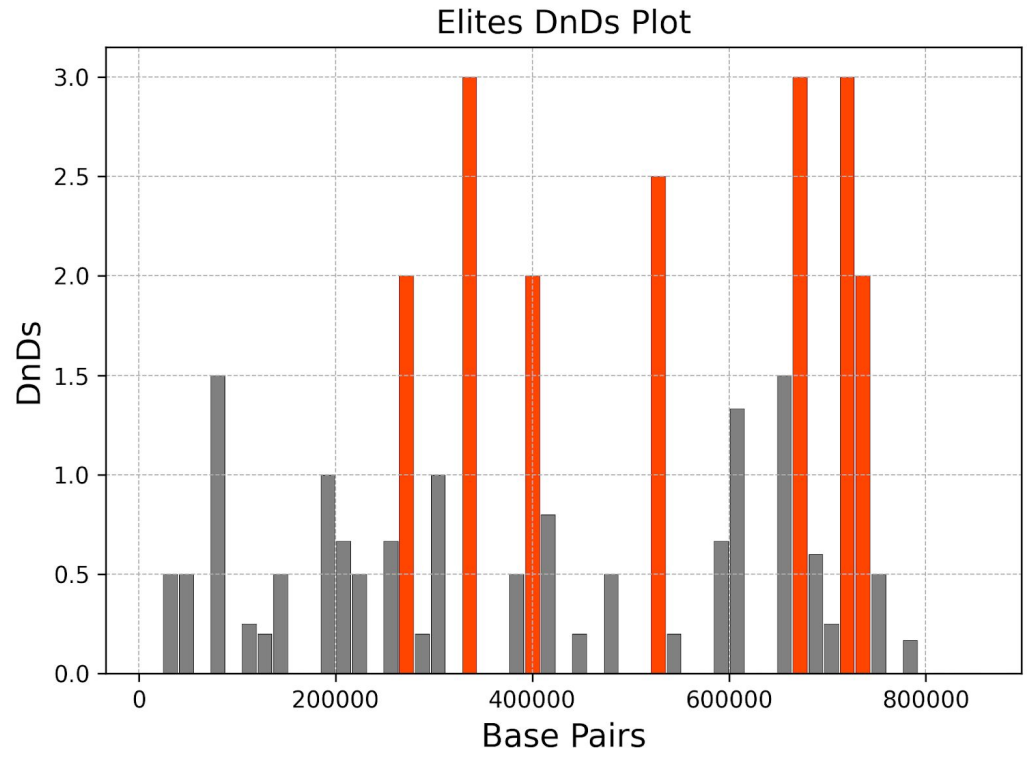


Figure 8: DnDs of the Prophet Population across Base Pairs

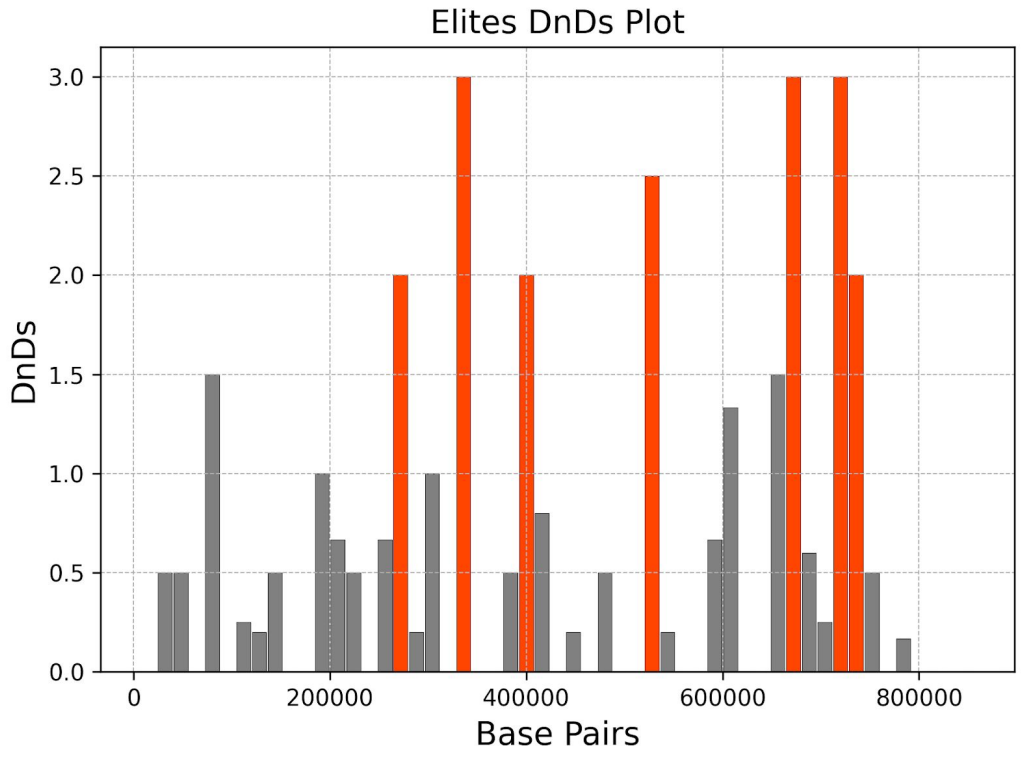


Figure 9: DnDs of the Elite Population across Base Pairs

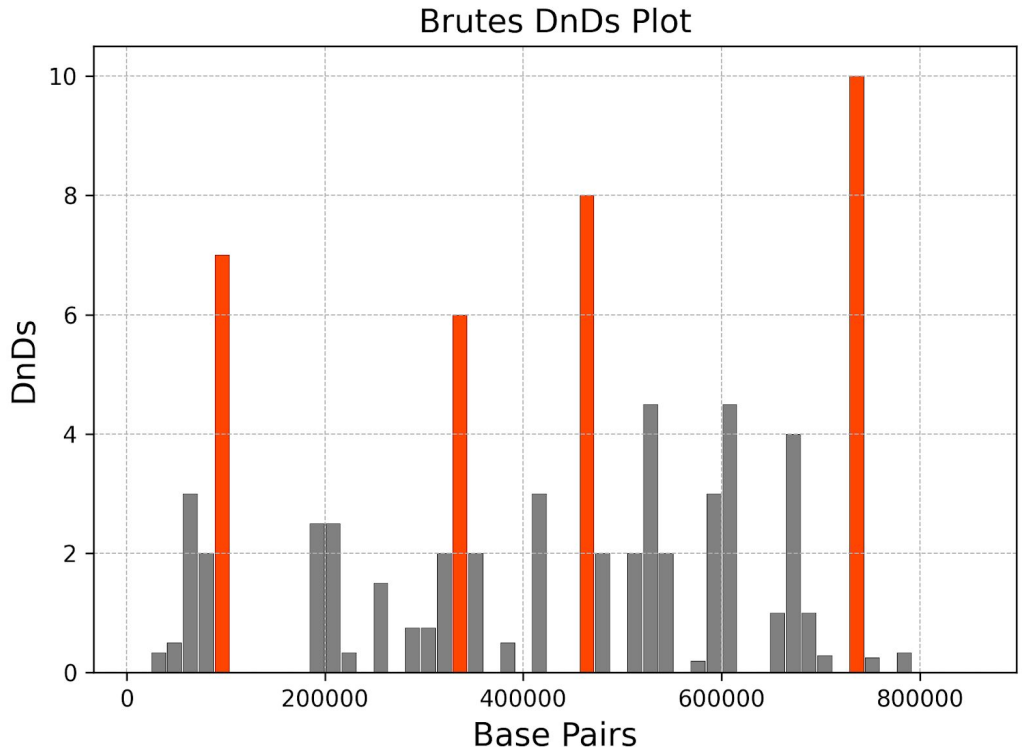


Figure 10: DnDs of the Brute Population across Base Pairs

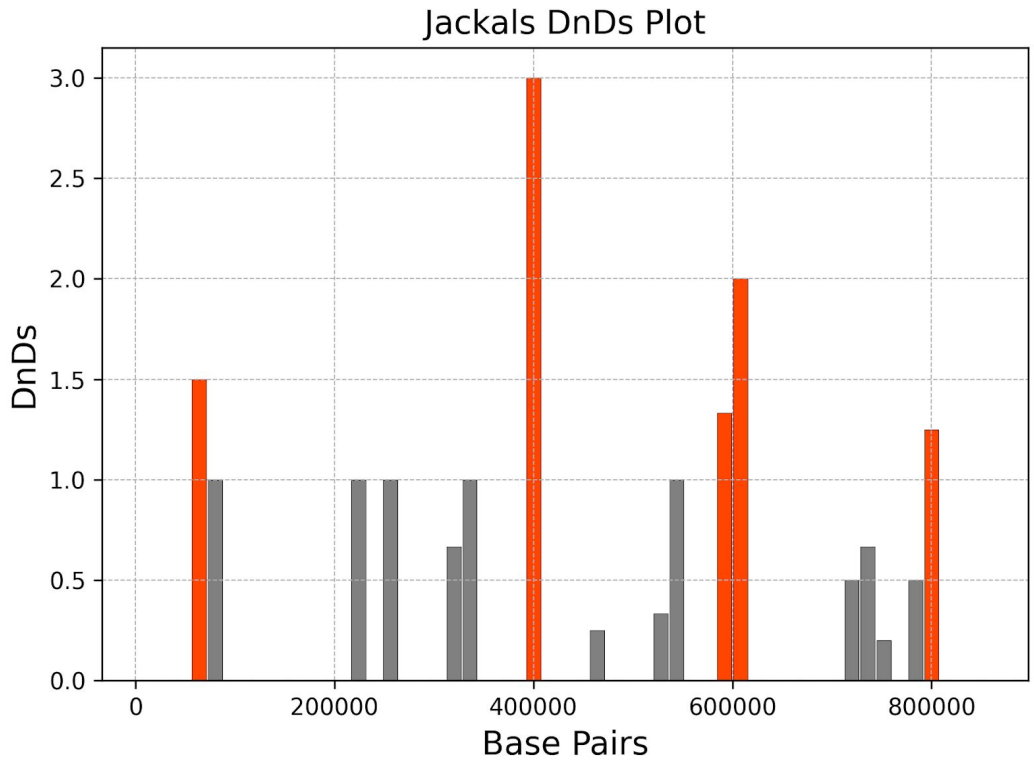


Figure 11: DnDs of the Jackal Population across Base Pairs

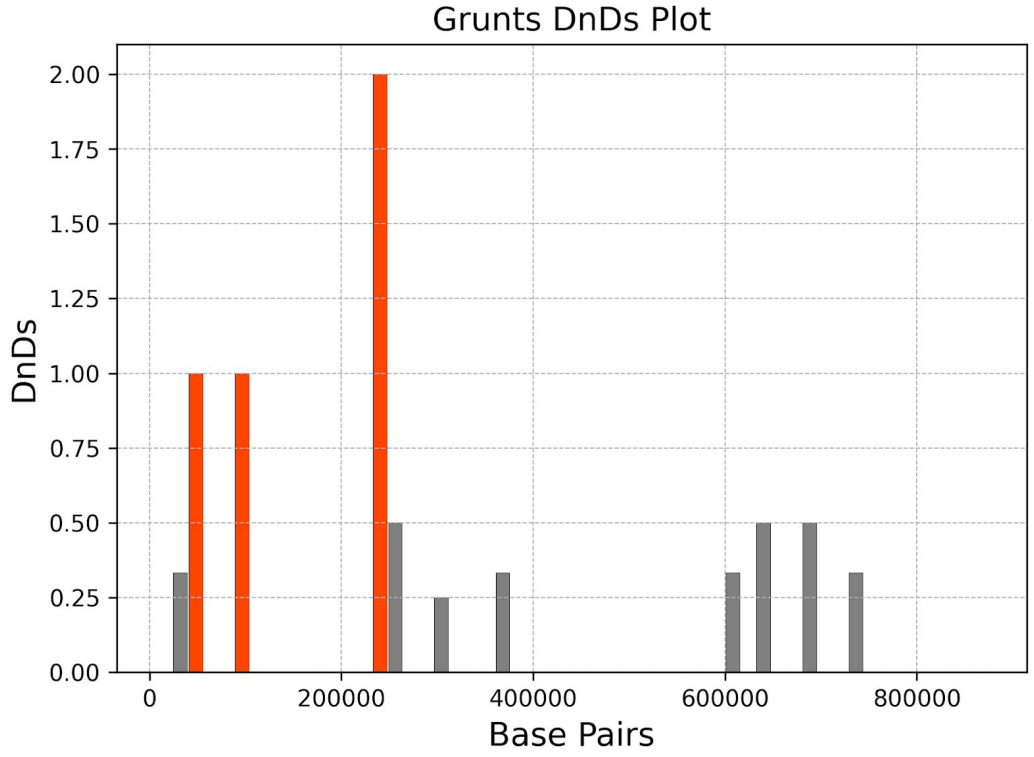


Figure 12: DnDs of the Grunt Population across Base Pairs

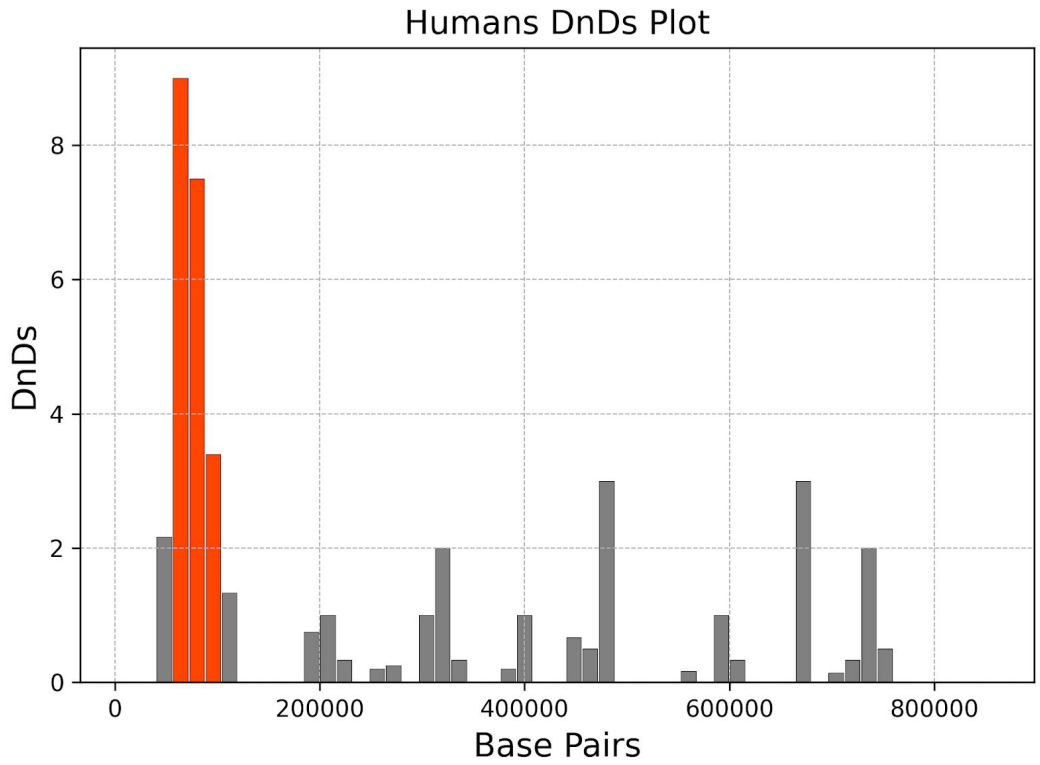


Figure 13: DnDs of the Human Population across Base Pairs

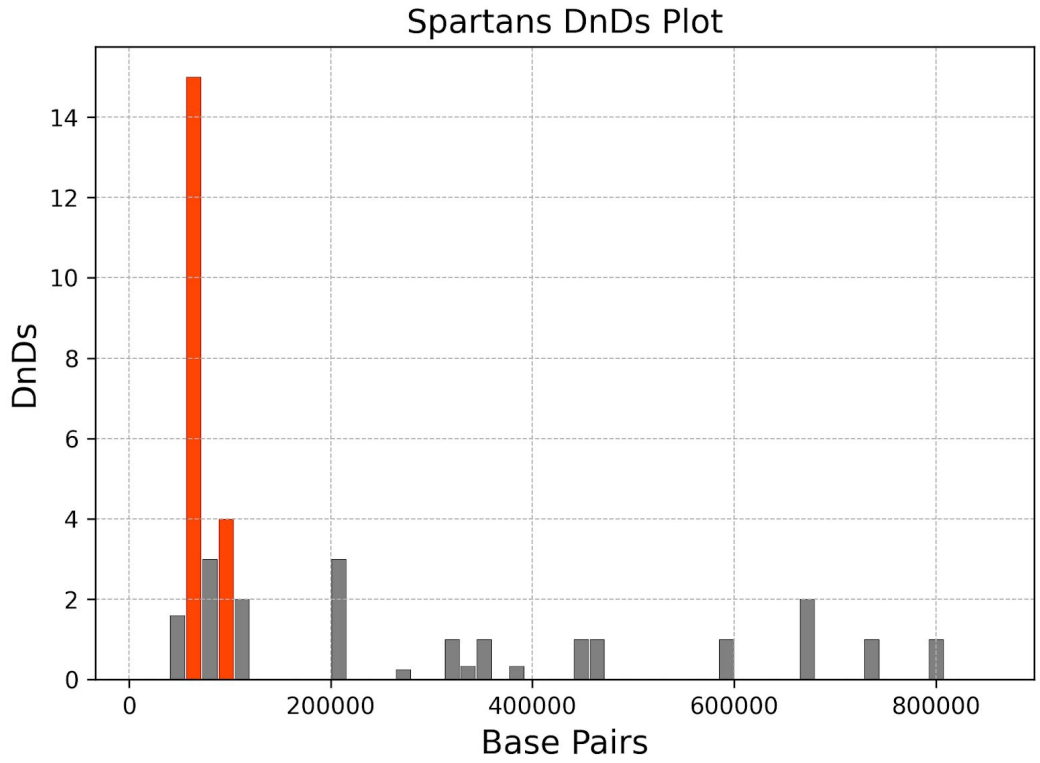


Figure 14: DnDs of the Spartan Population across Base Pairs

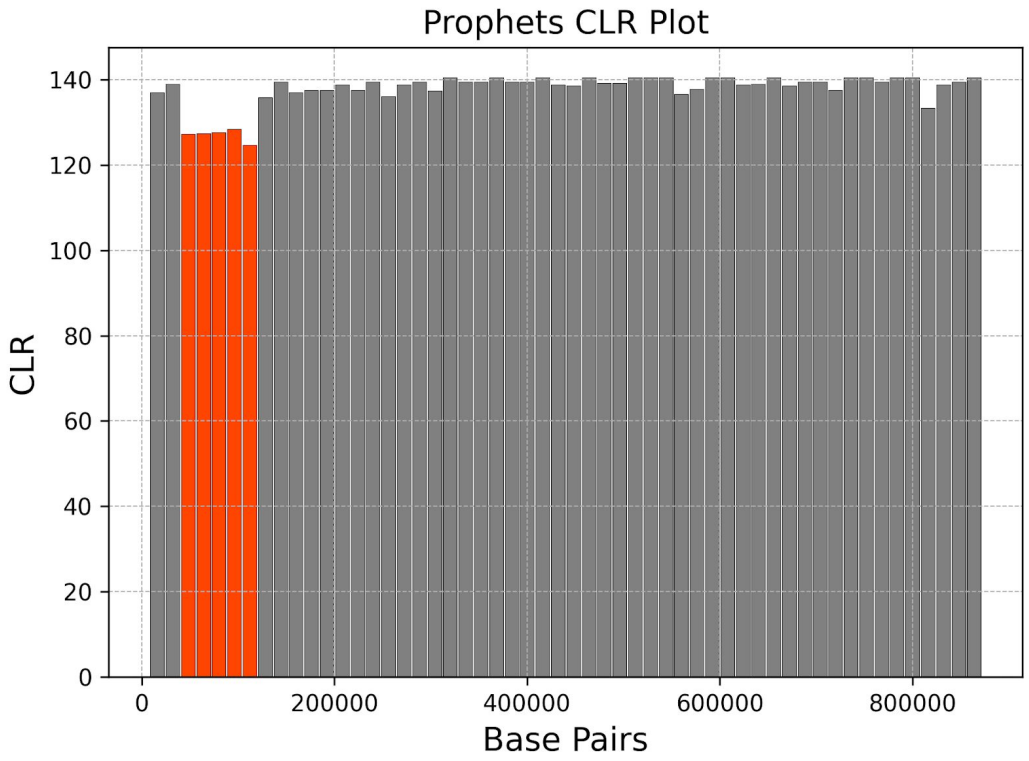


Figure 15: CLR of the Prophet Population across Base Pairs

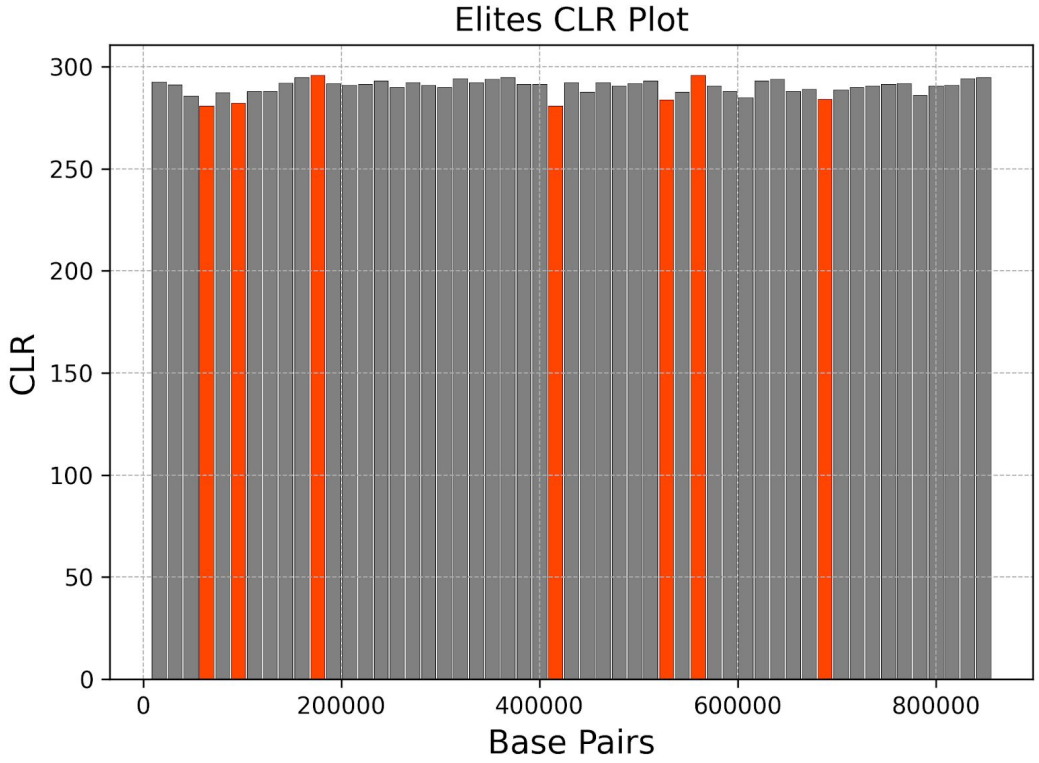


Figure 16: CLR of the Elite Population across Base Pairs

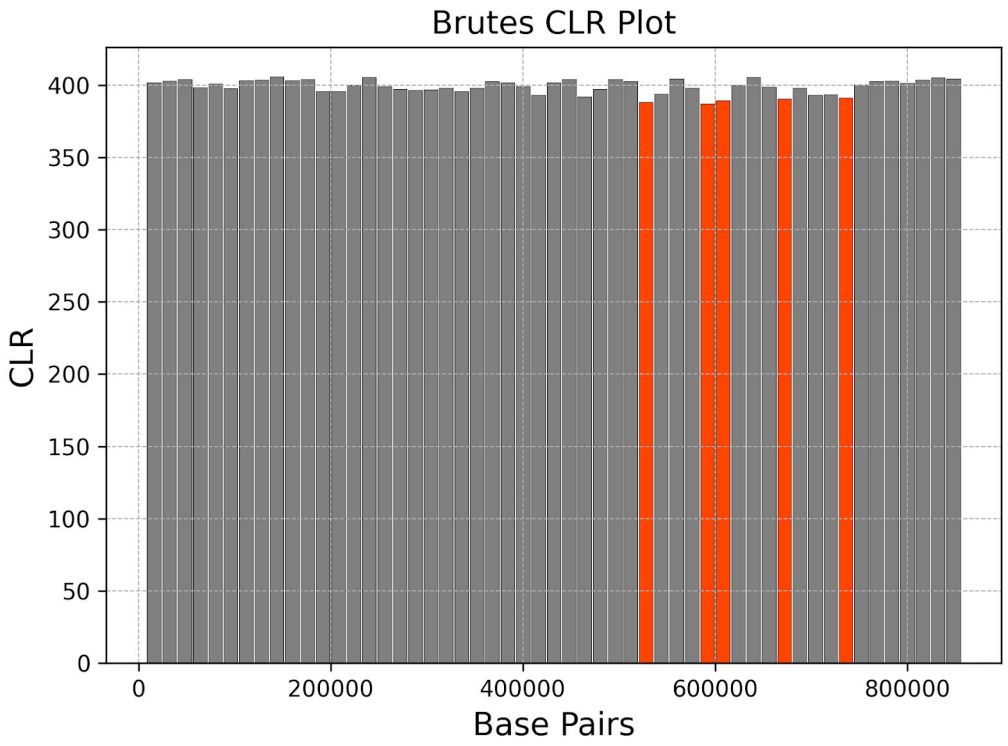


Figure 17: CLR of the Brute Population across Base Pairs

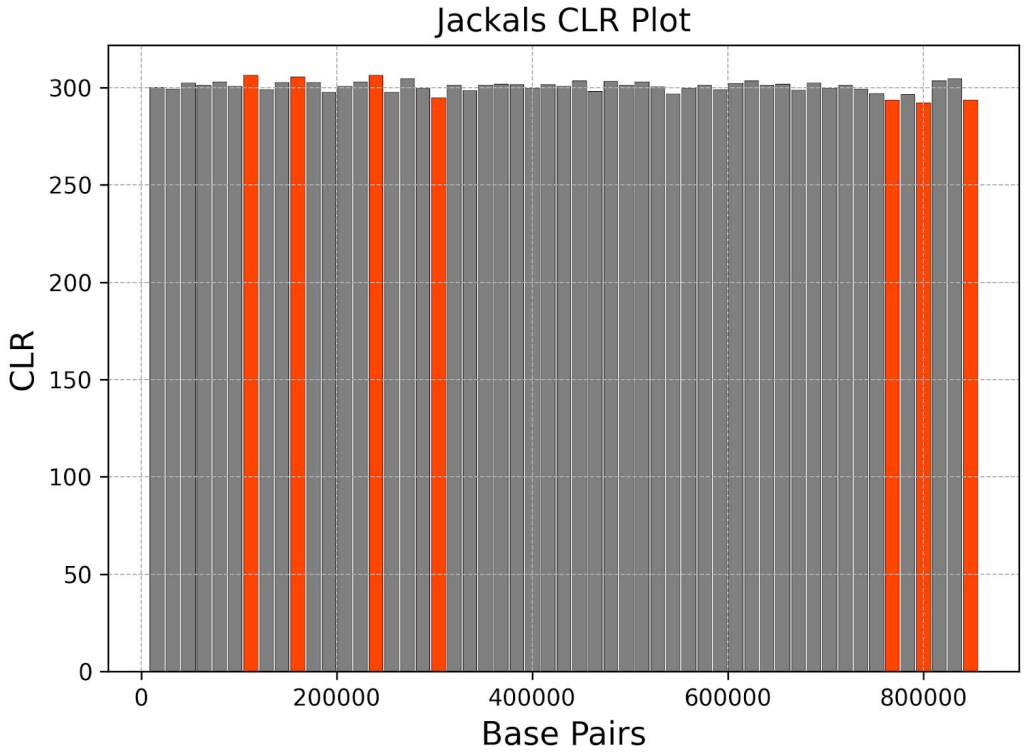


Figure 18: CLR of the Jackal Population across Base Pairs

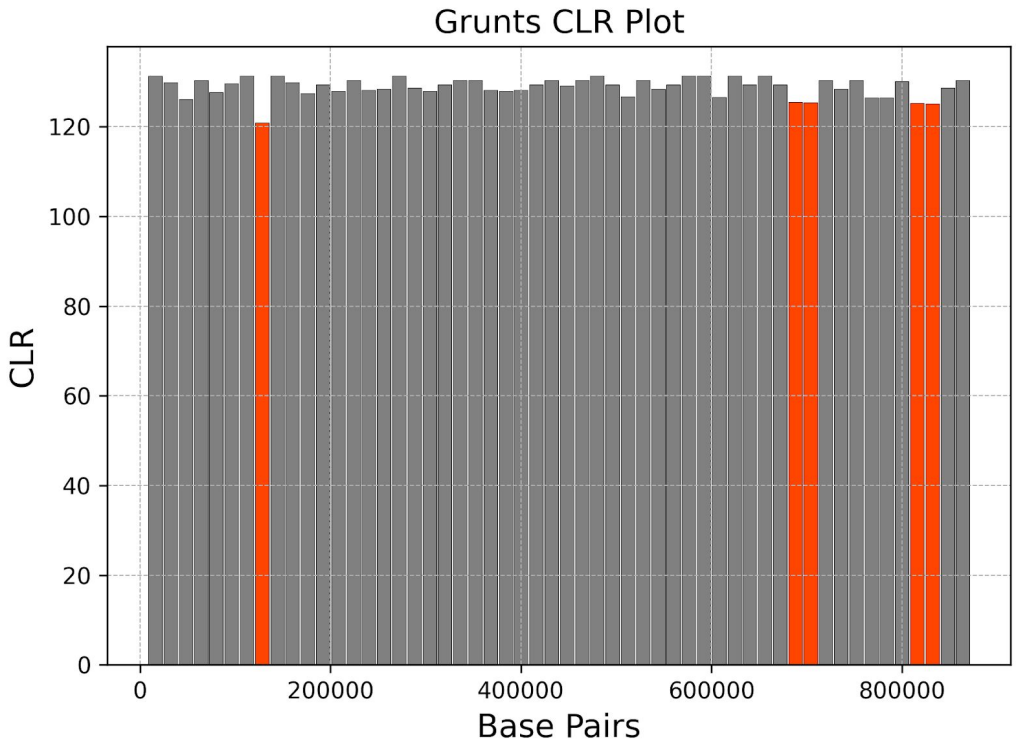


Figure 19: CLR of the Grunt Population across Base Pairs

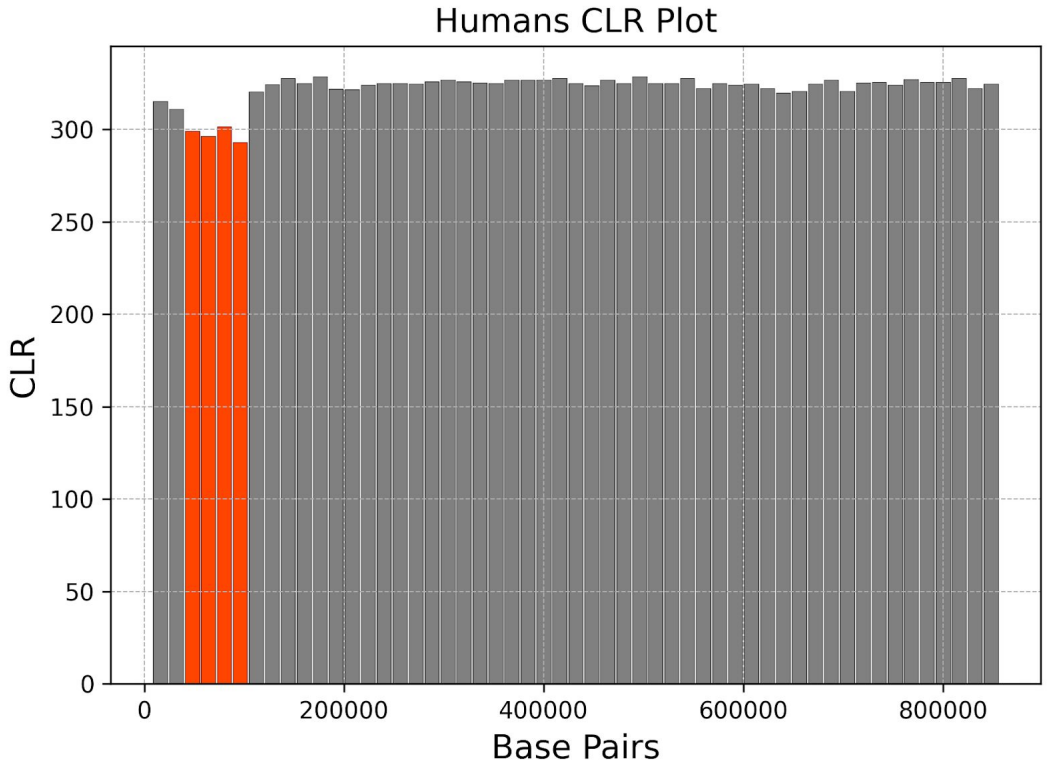


Figure 20: CLR of the Human Population across Base Pairs

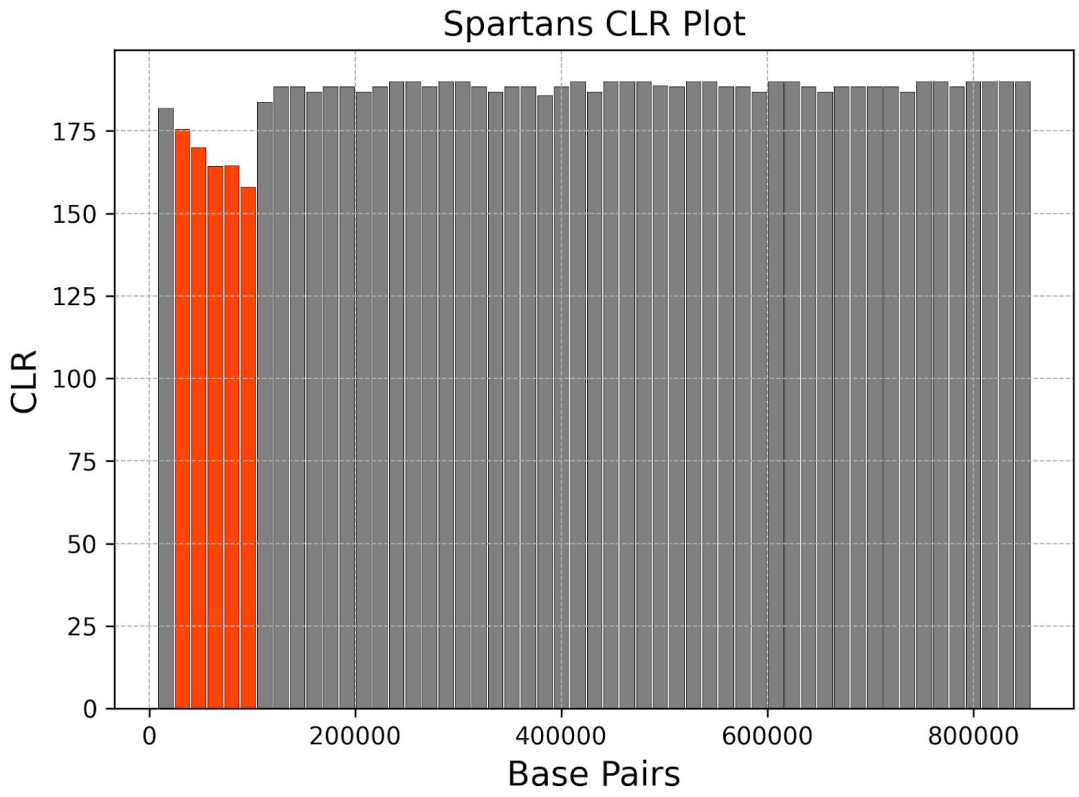


Figure 21: CLR of the Spartan Population across Base Pairs

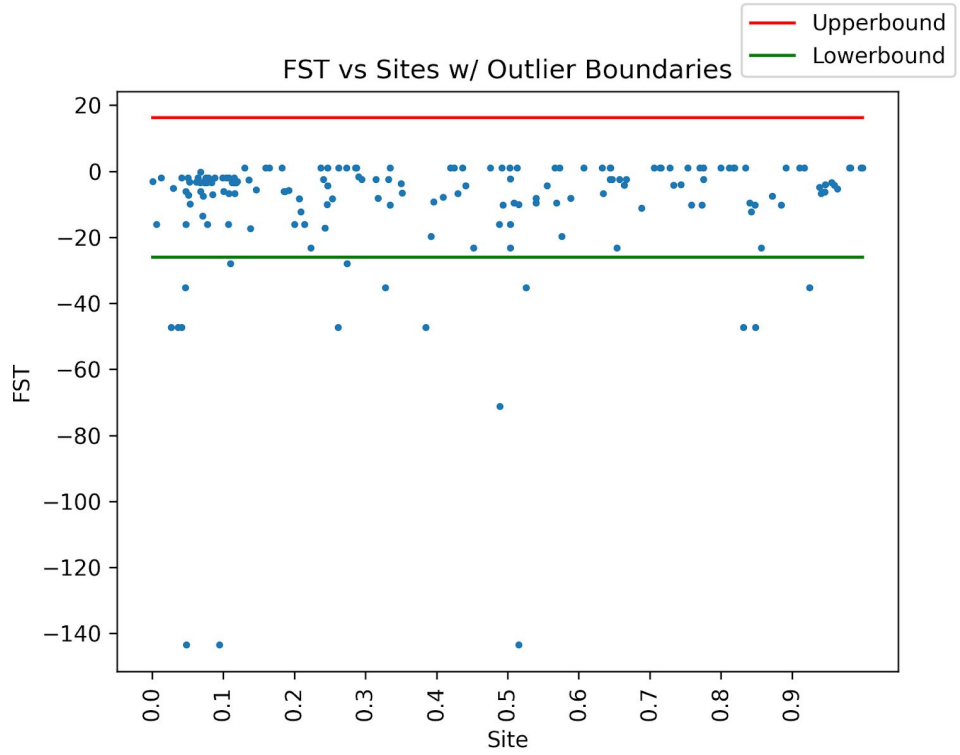


Figure 22: FST of the Prophet Population across Base Pairs

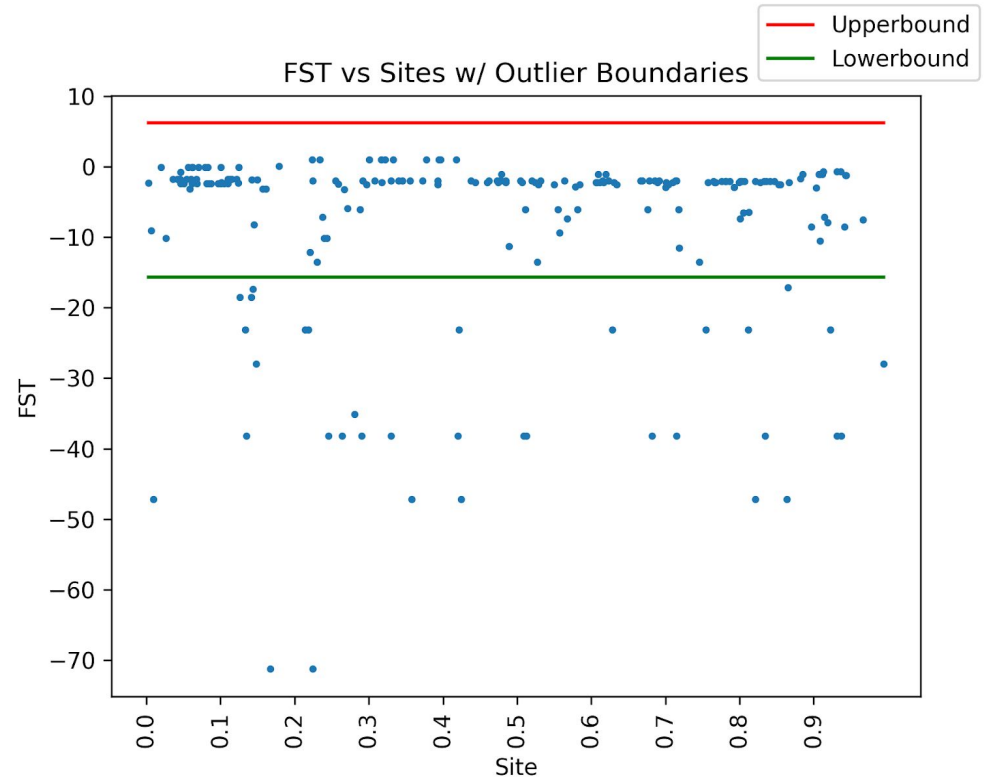


Figure 23: FST of the Elite Population across Base Pairs

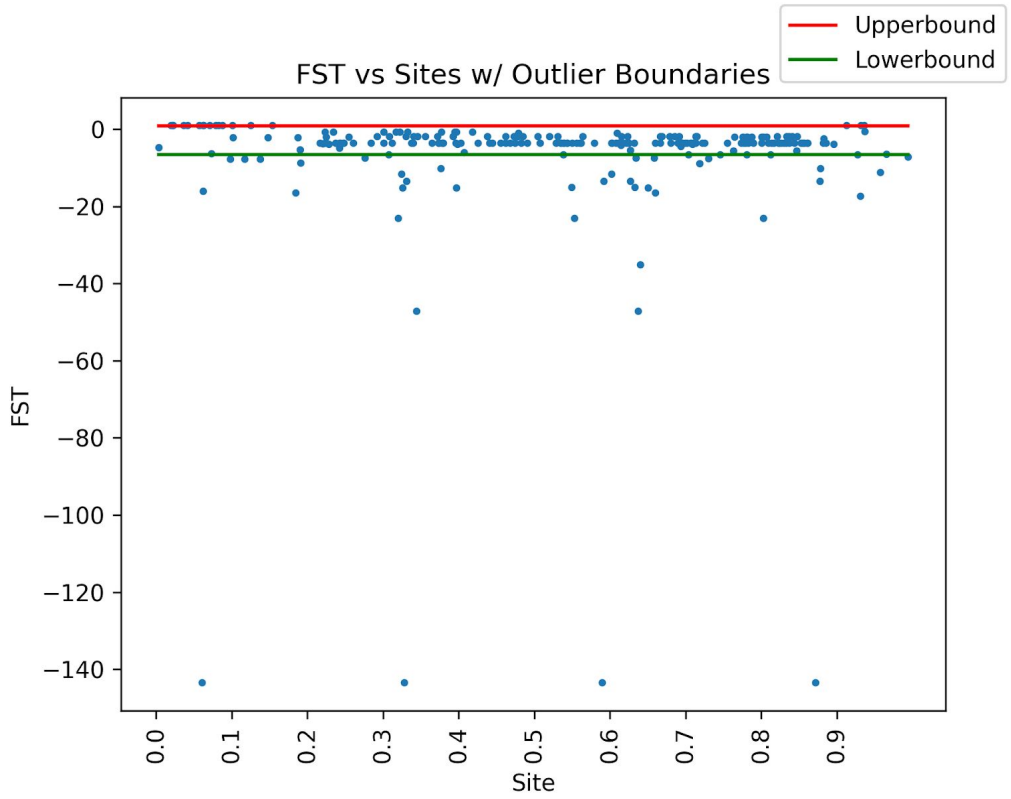


Figure 24: FST of the Brute Population across Base Pairs

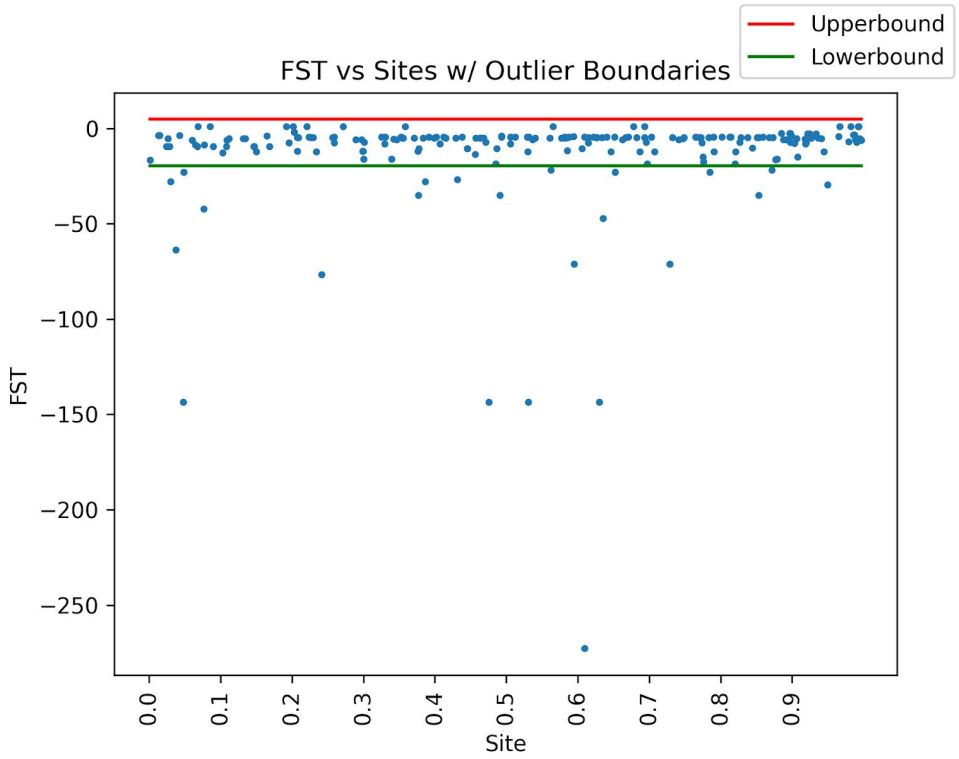


Figure 25: FST of the Jackal Population across Base Pairs

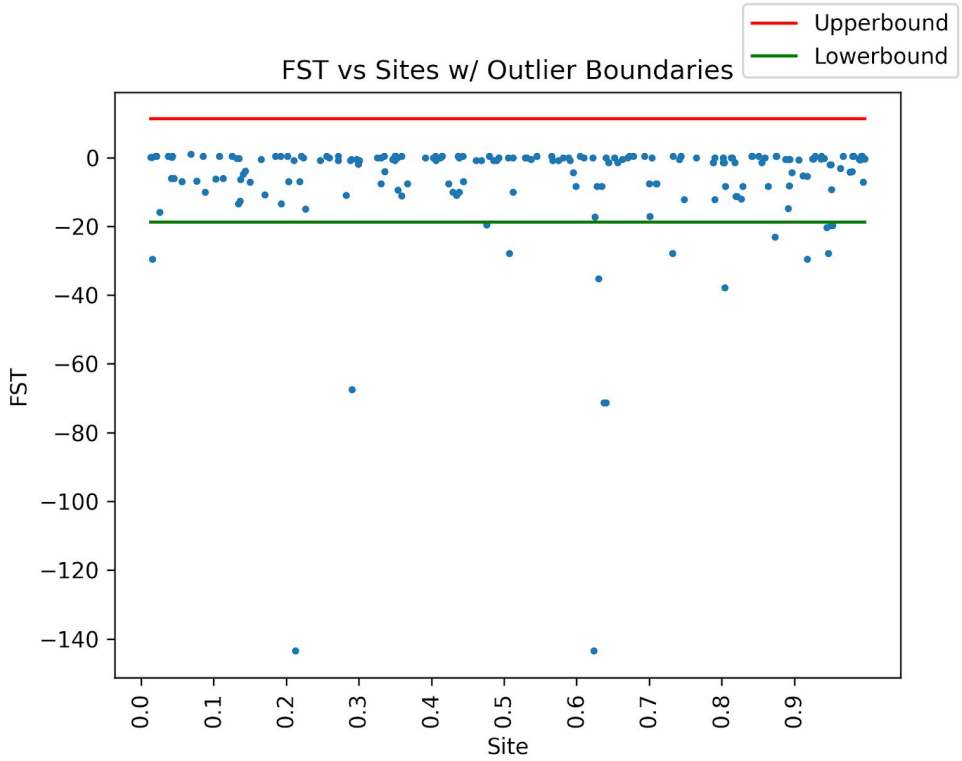


Figure 26: FST of the Grunt Population across Base Pairs

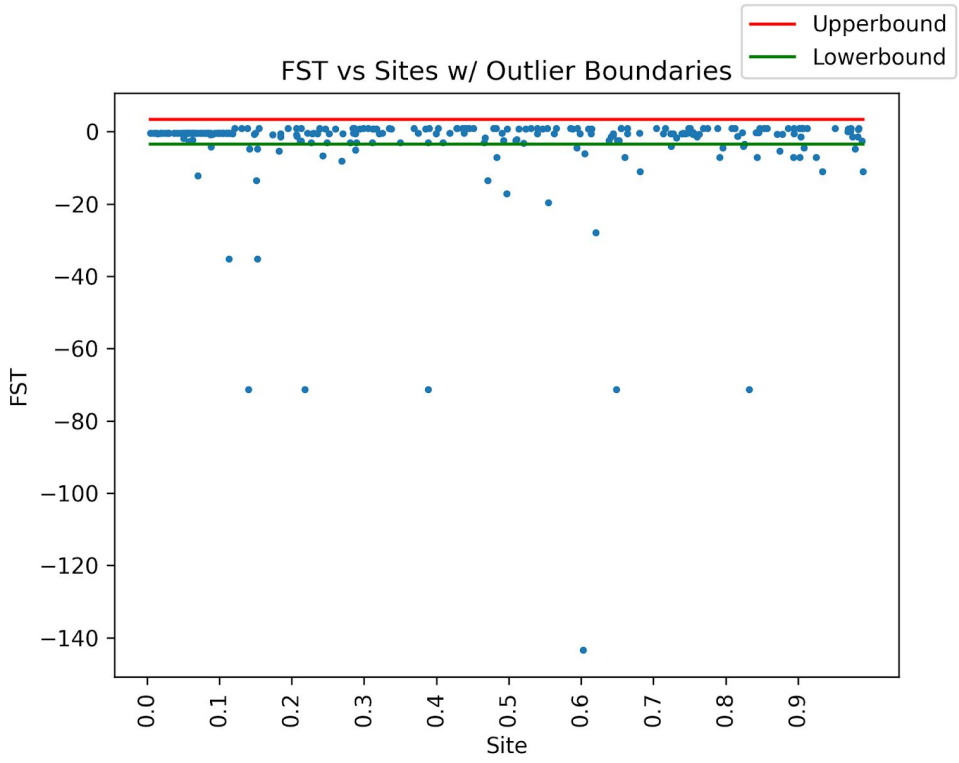


Figure 27: FST of the Human Population across Base Pairs

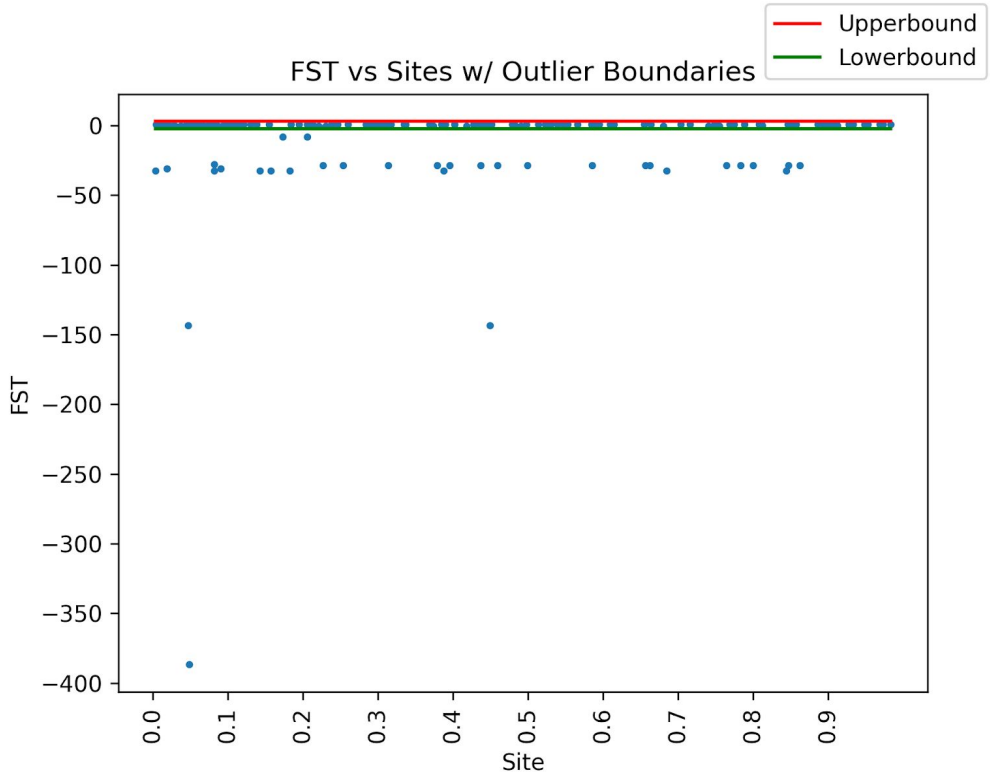


Figure 28: FST of the Spartan Population across Base Pairs

Data Changes Required:

Needed to output MS of all individuals in each population to run FST tool