Review of Gene Location Project Results and Methodology

Kathi Munoz, Andrew Meng, Joseph Zou

Methodology

Our test for gene location searched for coding mutations without any intergenic mutations between them. We created a map using the mutation data, which stored the type, position, and number of individuals with that mutation. Those data were then mapped to their local mutation IDs. We then created a sorted list of mutations of each population based on their mutation locations. Next, we used that sorted list to find all possible gene regions with non-coding mutation locations as boundaries. After that, we calculated pn/ps and dn/ds data for each possible region to determine their selection type using a threshold determined by the user.

Results

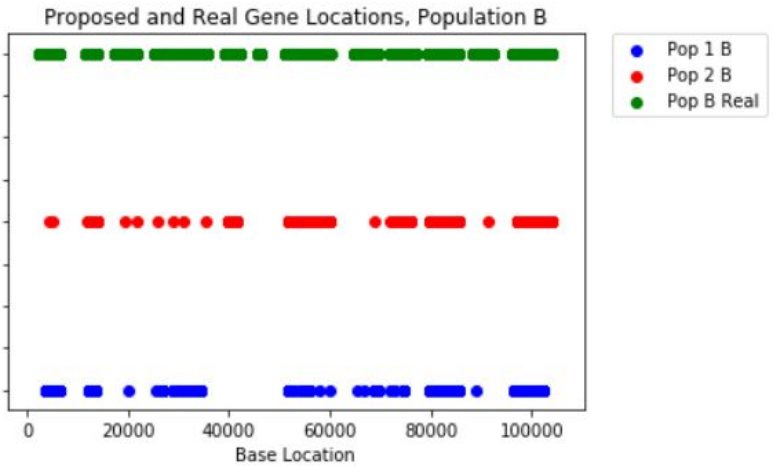

Figure 1: Visual comparison of proposed locations of genes using our methodology to true gene locations.

Discussion

Compared to the correct regions provided, our program provided many smaller regions that when combined intuitively would yield gene locations close to the correct positions. One of the problems we have is that we didn't review the possible gene regions we found, resulting in many regions of length one. Those regions should be combined with their neighboring regions to form larger genes. Another issue in our program is that we select gene boundaries based purely on their mutation types, while genes often extend beyond their highest and lowest position mutations. Consequently, our estimates of gene size were always on the small side.

Our interpretation of the two populations was as two entirely independent populations as well; as a result, we provided two separate analyses. It is interesting to us that one population was found to experience no positive selection and that the other experienced no negative selection. Gene locations identified in the two populations were found to overlap in some spots, as well, but the kinds of selection never overlapped unless both locations were found to experience neutral selection. Our overestimation of the number of genes in the populations makes analysis of the kinds of selection inherently flawed, it appears that we were able to recognize the kind of selection acting on a gene when the genes themselves were recognized appropriately. For example, the gene from sites 71853 through 76917 truly experienced positive selection, whereas we recognized two positively selected genes within this region, both of nontrivial (but smaller) size than that of the true gene. Similarly, we recognized a gene spanning sites 98707 to 101880 experiencing positive selection, and this region exists entirely within a larger region identified as a true gene.