Jake Waro
September 16, 2020
CSci 5481
Homework 1

Question 7
(10 points) Where is the largest discrepancy in amino acid counts between the coding sequences (correct frame shift) and the whole genome sequence (random frame shift), and why?

The largest discrepancy in amino acid counts between the coding sequences and the whole genome sequence was for the **stop (stp) amino acid**(s) (i.e. **Ochre, Amber, & Opal** → codons TAA, TAG, & TGA). The whole genome had 774 counts of these, while the coding sequence only recognized a mere 12.

This is intuitive given the nature of separating out coding sequences. A stop codon should tell us where the coding sequence ends. The separated genome file lists 12 different coding sequence lines, while resulting with 12 counts of stop amino acids. This would be the expected outcome, as we should expect each coding sequence line to end with a stop codon.

The whole genome (random frame shift) doesn't add any intelligence to interpreting the genome. Based on starting from the beginning of the whole genome, we resulted with counting hundreds of stop codons, likely due to the randomness of how the codon sections worked out by starting from the very first position, clearly showing an overestimation of stop codons compared to the separated coding sequences.