PERL PROJECT

Estimating Stop Codons Frequencies

by Iñigo Oyarzun

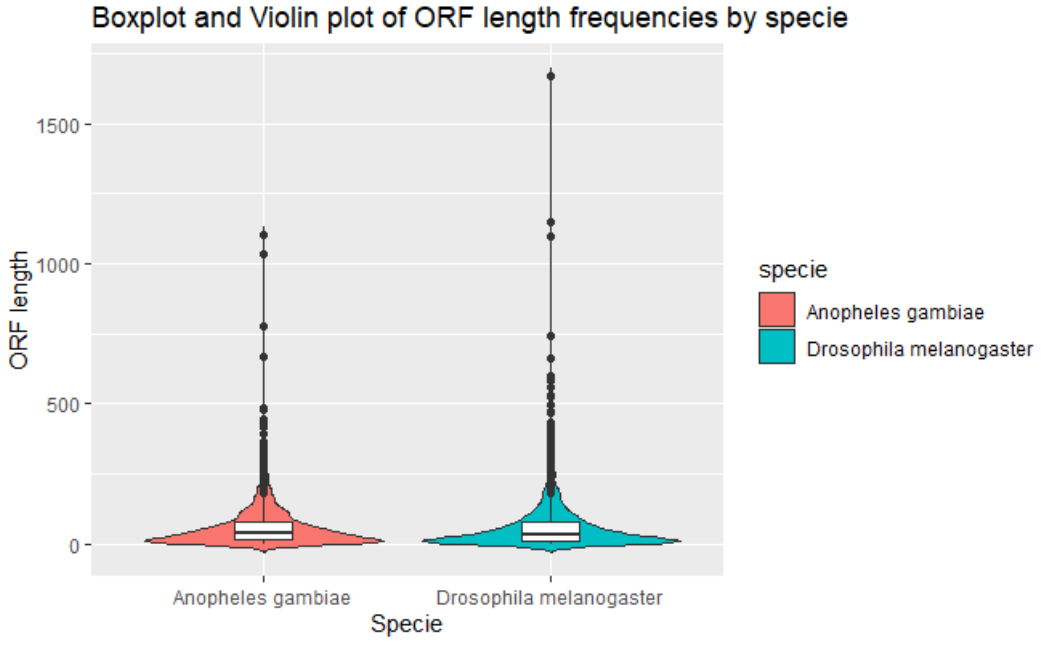
After run our program with both genomic sequences, Anopheles gambiae and Drosophila melanogaster, we obtained these two tables with the information about the frequency of the different stop codons.

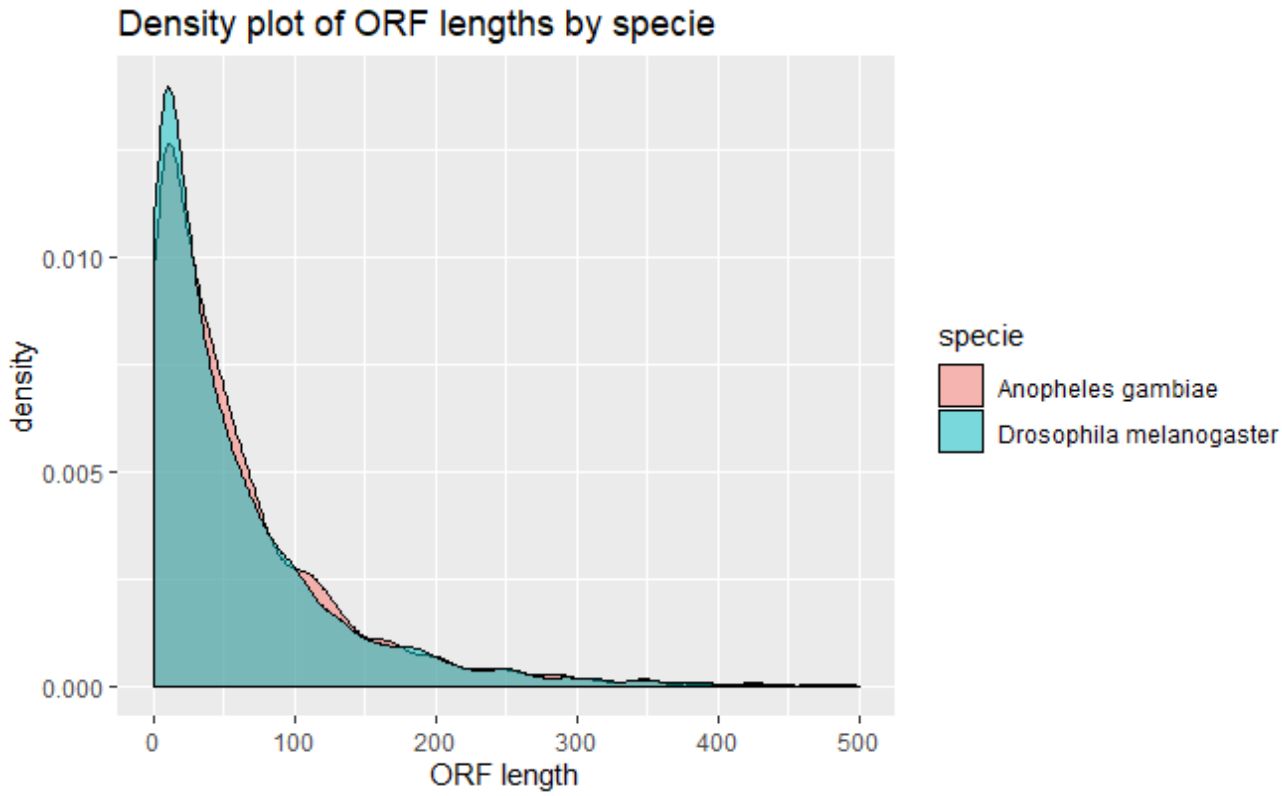
Anopheles gambiae Stop Codons	Relative Frequency All	Relative Frequency Stop Codons
TGA	1,668%	36,435%
TAG	1,036%	22,63%
TAA	1,874%	40,935%

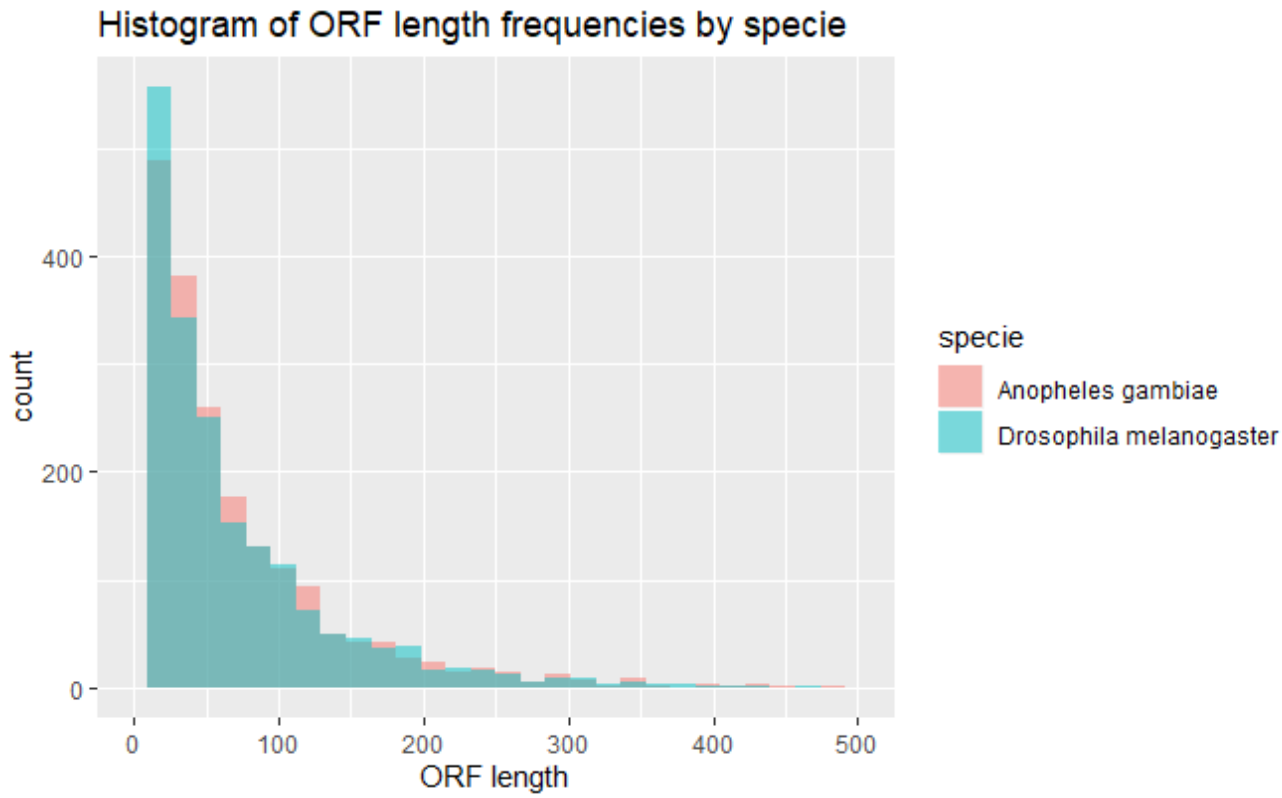
Drosophila melanogaster Stop Codons	Relative Frequency All	Relative Frequency Stop Codons
TGA	1,554%	33,333%
TAG	0,932%	19,991%
TAA	2,176%	46,675%

As we can see, in both species, the most frequent stop codon is TAA followed by TAG and the least frequent stop codon is TGA. We can also say that both species have a similar relative frequency of these stop codons, both when relativizing with the whole quantity of possible codons and when relativizing just with the quantity of stop codons. The stop codon that most vary in these measurements from one specie to the other is TAA.

Also, we have obtained information about the ORF lengths of these sequences and we have tested their distributions. In both species the differences in the distribution of the ORF lengths between the forward strain and the reverse strain were not significant, using Wilcoxon test. However, the differences in the distribution of the ORF lengths between the species were significant using the Wilcoxon test with a p-value = 0.01769. So we decided to create some plots for visually check these significant differences.







In order to visualize the distribution differences better, we have cut-off the x-axis at 500 at the density plot and at the histogram. Using this procedure, we have skipped 14 rows, or ORFs, from the initial data that their length is over 500. But we have to keep in mind that, actually, these 14 outliers could be the most interesting to study from the sequences cause as longer an ORF is as bigger is the probability of it coding for a real protein.