**Problem:** Finding outliers of aligned sequences for each gene model as potential sequences with anti-codon shifts. Normalizing the distribution of scores for each gene model. Scoring outliers against each model to find possible anti-codon shifts.

**Input**: Aligned tRNA gene sequences from alignment pipeline.

**Step1**: Finding outliers in each cluster of tRNA gene models.

I used package OD-seq, which is based on finding sequences whose average distance to the rest of the sequences in a dataset, is anomalous. I found a total of 39 outliers which 9 of them were the same genes with possible anti-codon shift, I found before.

Table on the right side shows the number of outliers found in each model. (script1)

| model | outlier#/total |
|---|---|
| A | 0/316 |
| R | 1/522 |
| N | 0/187 |
| D | 1/157 |
| C | 1/96 |
| Q | 2/243 |
| E | 0/240 |
| G | 1/342 |
| H | 1/120 |
| I | 2/256 |
| L | 0/504 |
| K | 19/274 |
| M | 1/145 |
| X | 0/114 |
| F | 1/156 |
| P | 3/300 |
| S | 2/342 |
| T | 3/327 |
| W | 0/78 |
| Y | 0/117 |
| V | 1/361 |
| X | 0/114 |

total number of outliers = 39

**Step2:** Excluding the outliers from the gene file.

**Step3:** Creating the profiles for each model.

**Step4:** Scoring each sequence against all the models.

**Step5:** Visualizing the distribution of the scores for each model. (script2)

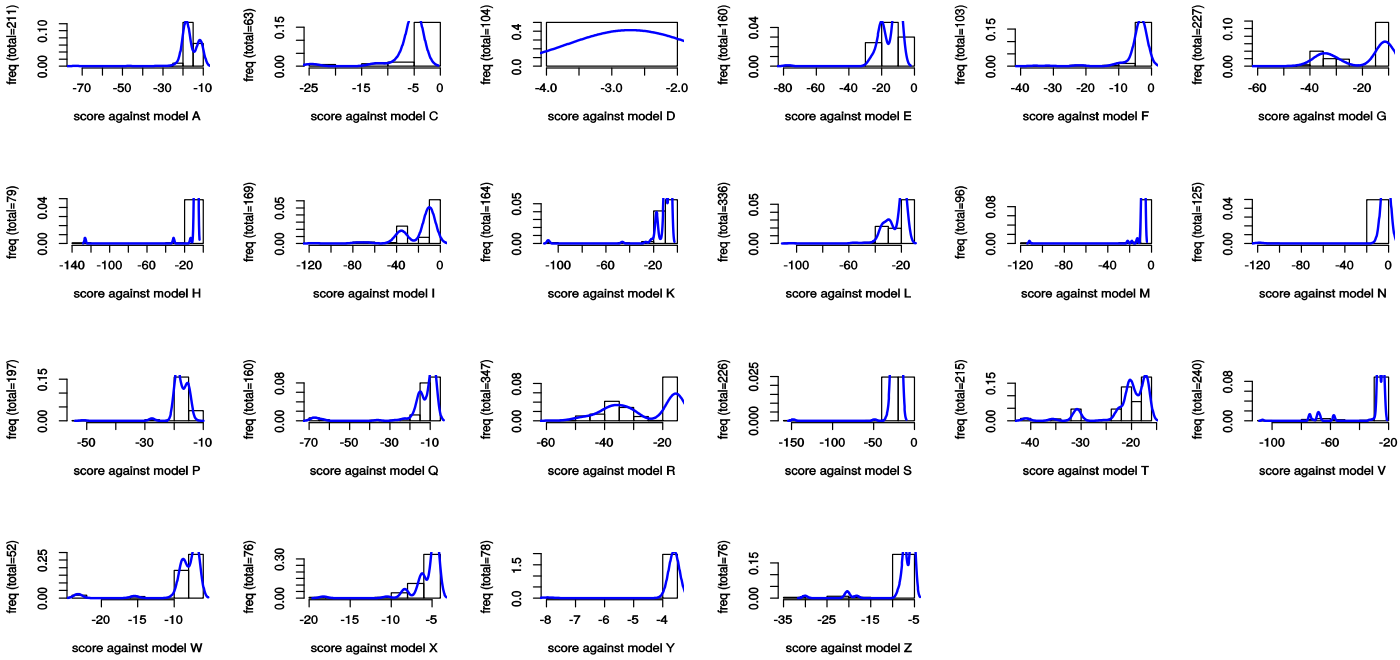


*fig 1: distribution of scores for each model for all the TriTryp tRNA genes from the intersection set.*

As you can see in figure 1, the distribution of some of the models is not normal.

Using the alignment viewer Seaview I aligned the sequences of model R (which has a bimodal distribution), and saw that the alignment changes when the genomes/clusters changes. As you can see part of the alignment in this figure.

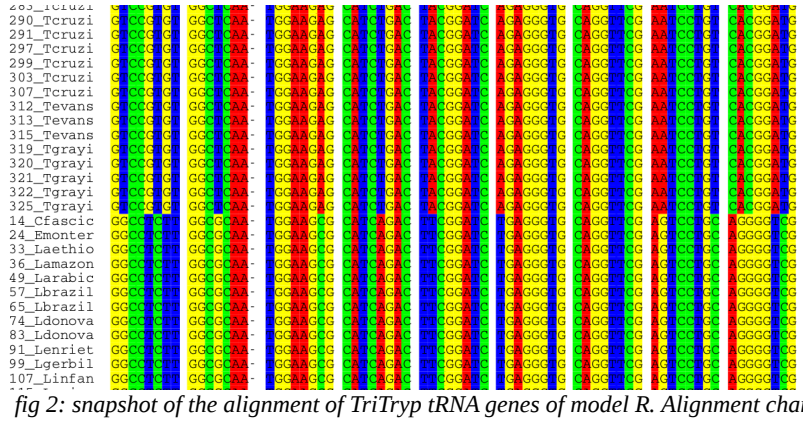


*fig 2: snapshot of the alignment of TriTryp tRNA genes of model R. Alignment changes from Tripanosoma to Leishmania*

So, as an attempt to make the distributions normal, I made two clusters of Leishmania and Trypanosoma and created the profiles for each of the clusters.
Here is a visualization of the distribution of scores for each cluster:

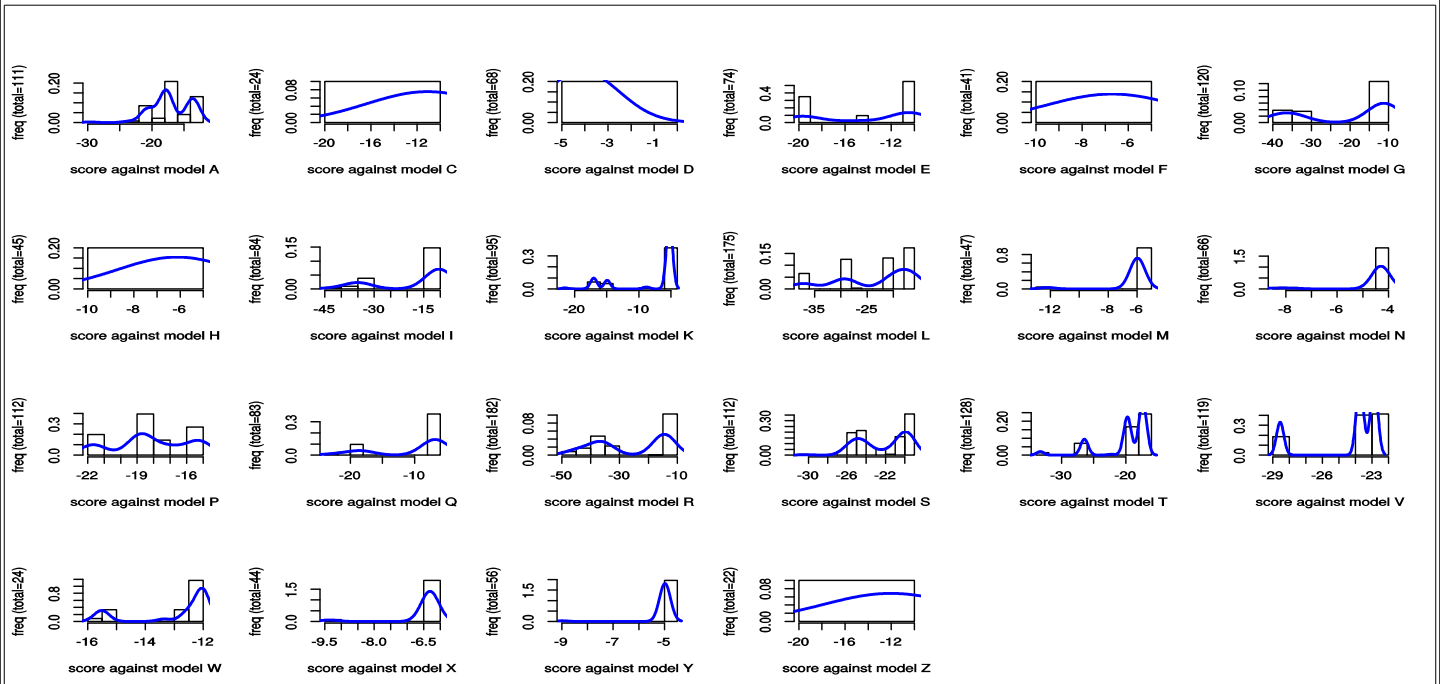


*fig 3: distribution of scores for each model of Leishmania tRNA genes*

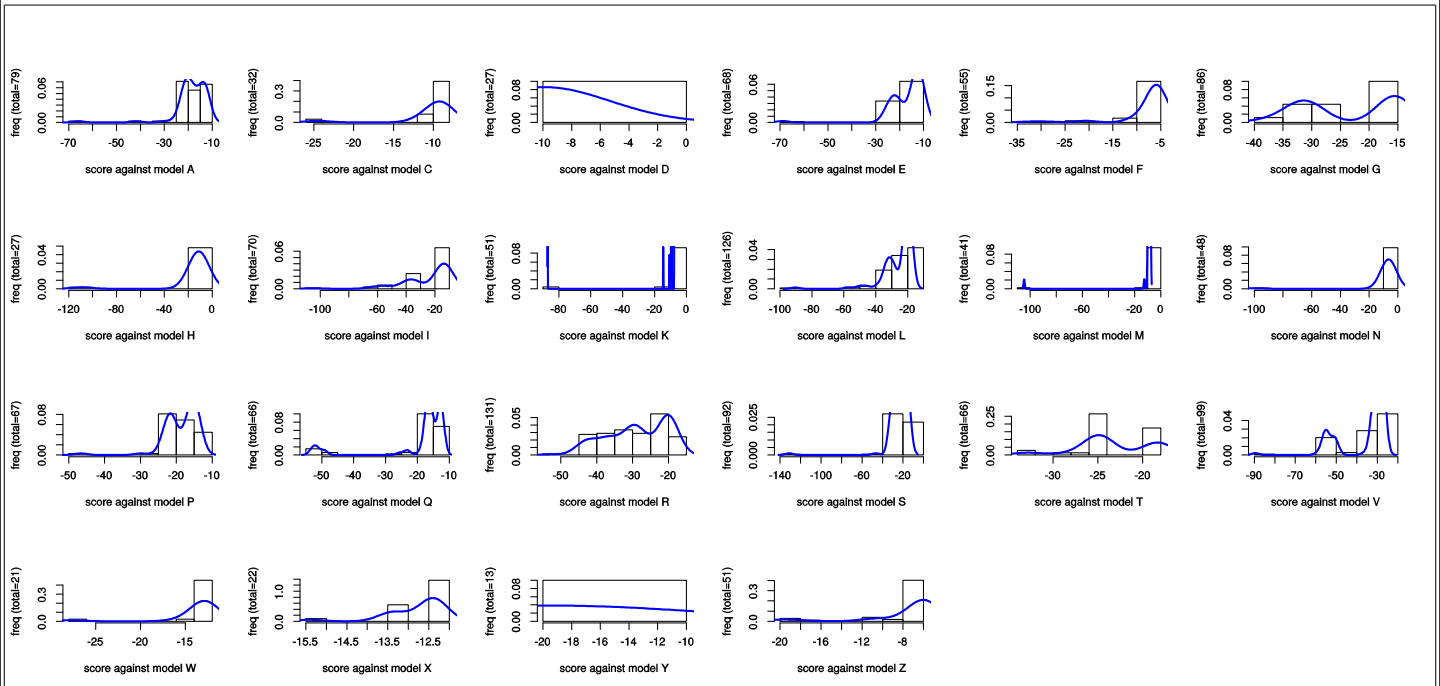


*fig 4: fig 3: distribution of scores for each model of Trypanosoma tRNA genes*

As you can see the distribution of the models R and T in Trypanosoma has become more like the normal distribution. Although the distribution of Leishmania scores, has not improved!
I assume that by making the clusters smaller, distributions will become more like normal distribution. however, by making smaller clusters of genomes the number of sequences for each model will become very small, and making a profile based on them is not reliable.
The next step is to find the score of the outliers against each model. Should I make the models from smaller clusters?