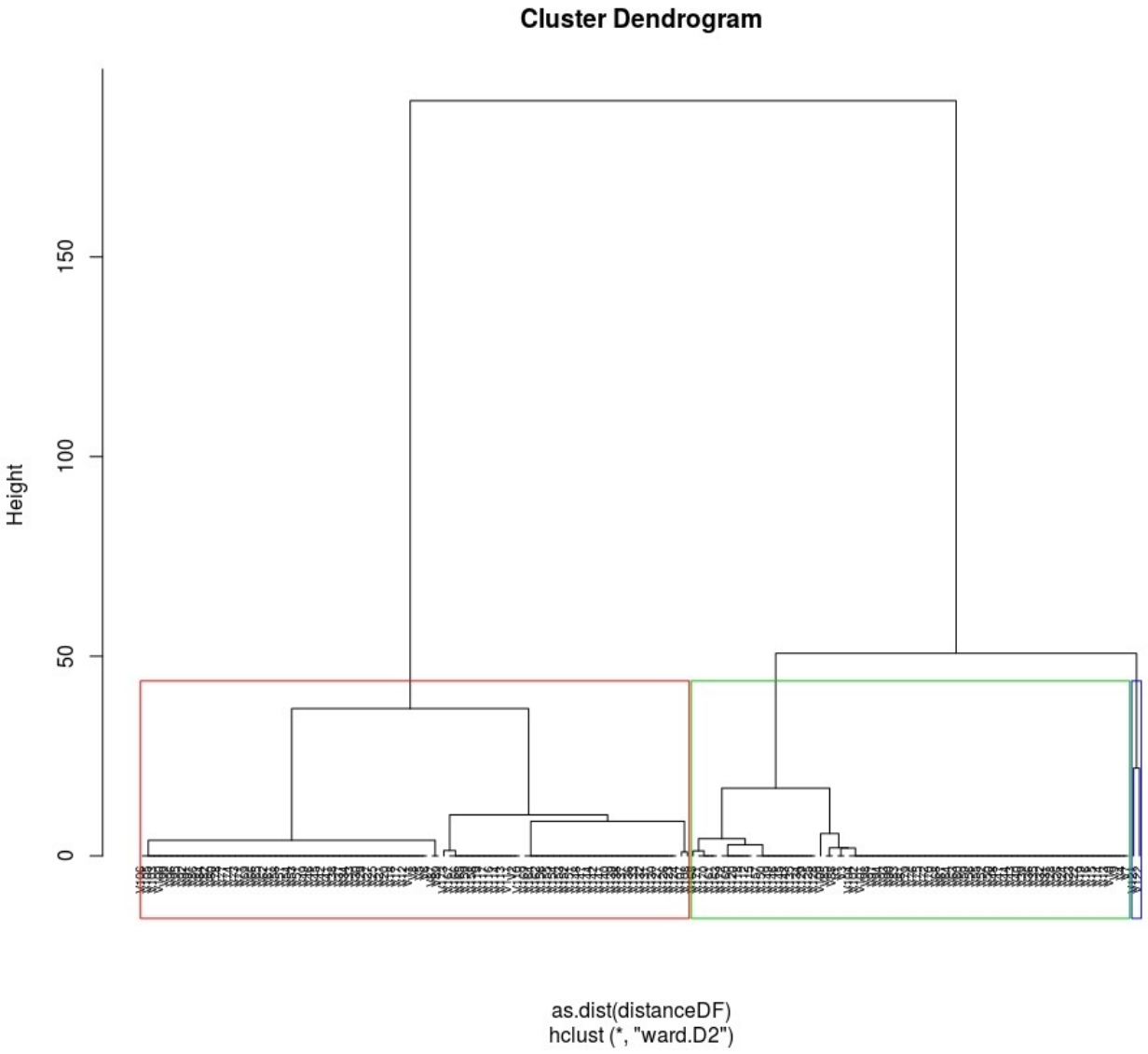
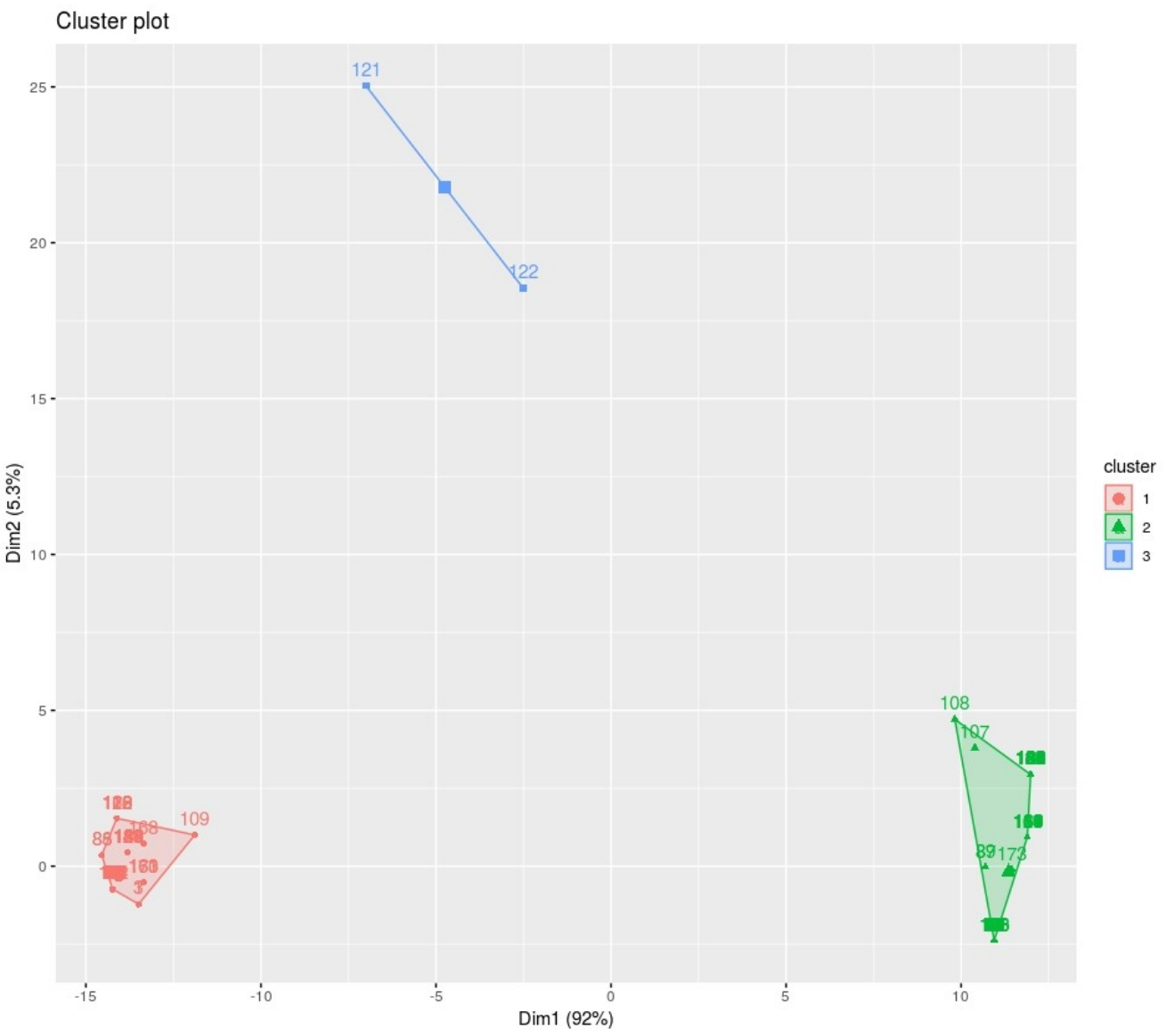
Clustering CAT Genes Using the Levenshtein (edit) distance between gene sequences. using Agglomerative clustering which works in a bottom-up manner. I used Ward.D2 method to measure the dissimilarity between two clusters of observations which minimizes the total within-cluster variance.

For the CAT genes detected by both ARA and TSE, we have the fallowing Dendrogram:



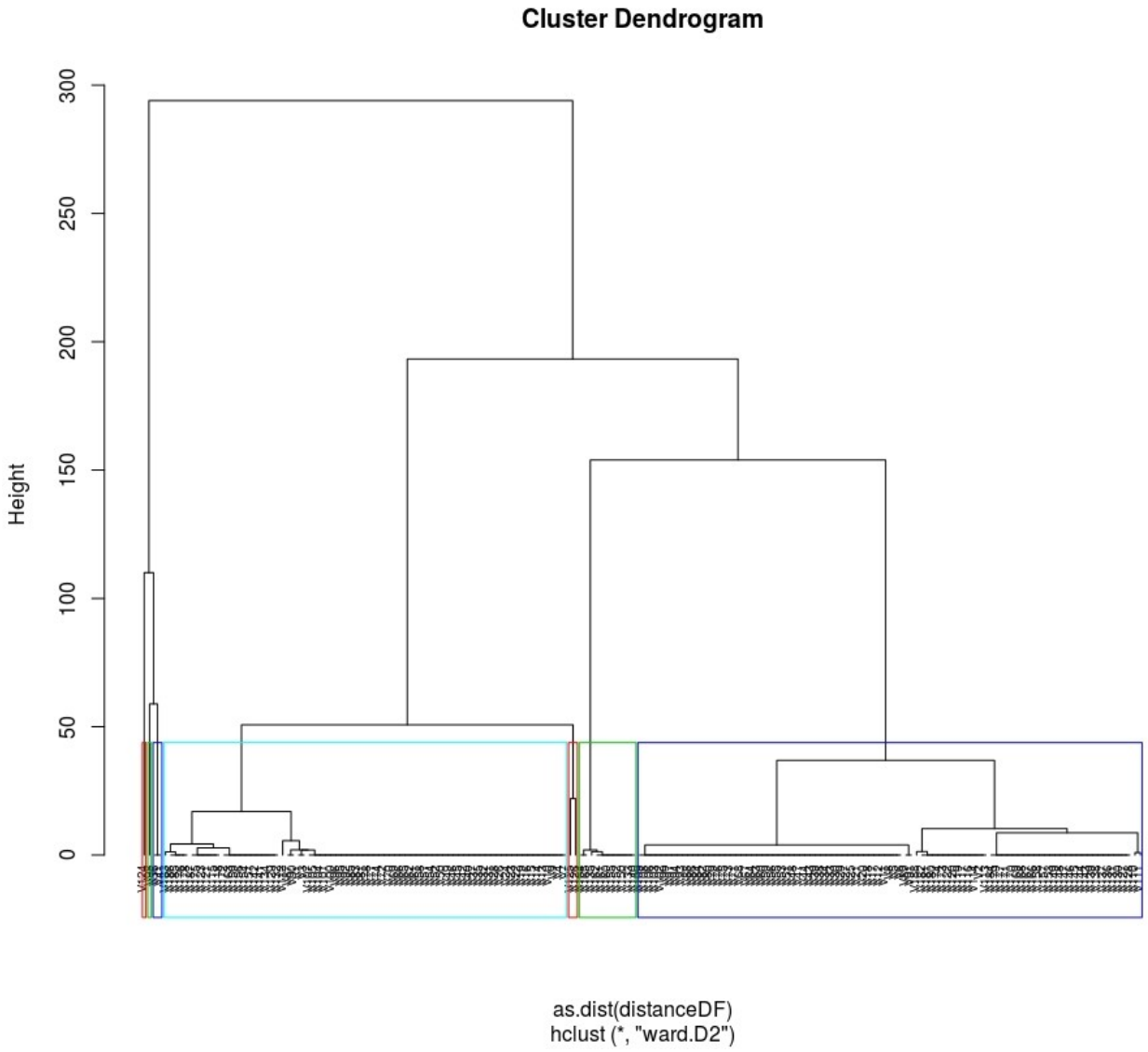
The height of the cut to the dendrogram controls the number of clusters obtained. (It plays the same role as the k in k-means clustering). So, in order to make tight groups of clusters, I ended up cutting the tree into 3 clusters. The fallowing plot displays a "summarising" scatterplot to visualize the clusters:



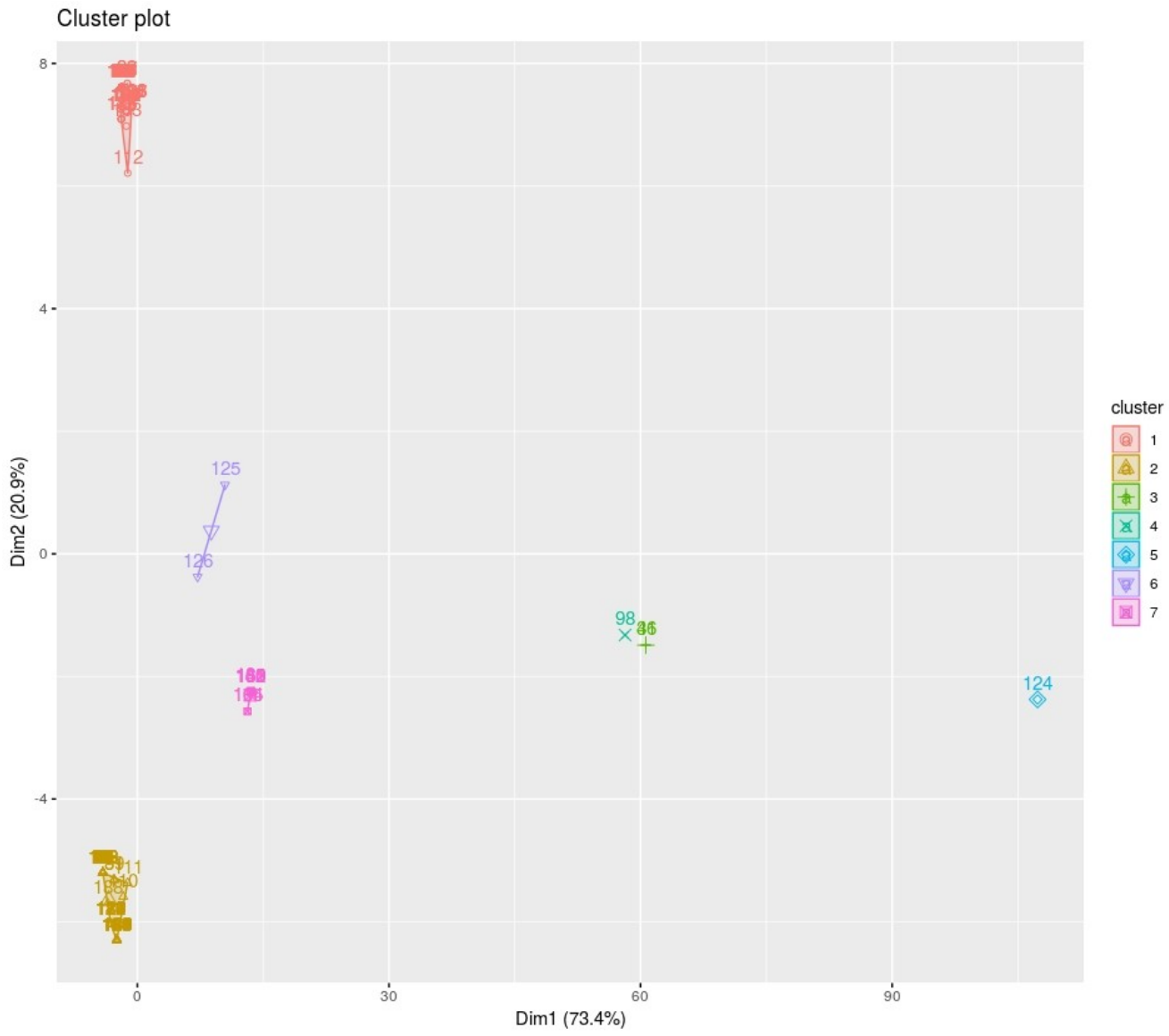
And in the fallowing table you can see number of sequences in each cluster and their range of similarity (edit distance)

	sub_groups	Freq	editdistanceRange
1	1	76	0-6
2	2	95	0-8
3	3	2	0-22

Next, I ran the same clustering method for gene sequences detected as CAT by either of ARA or TSE.



For this set of sequences, I ended up cutting the tree into 7 clusters to make clusters almost as tight as the previous set.



Here is the table of clusters.

	sub_groups	Freq	editdistanceRange
1	1	76	0-6
2	2	95	0-8
3	3	2	0-0
4	4	1	0-0
5	5	1	0-0
6	6	2	0-22
7	7	11	0-2

So, cluster 1,2 and 6 look like the same clusters of genes we had for the previous set of genes, And cluster 3,4,5,7 are new added clusters for the CAT sequences that are detected as CAT by only one of the genefinders.

Table1 in the paper by CHRISTIAN MARCK and HENRI GROSJEAN, shows the base distribution (number of A, C, G, or T in each position) for Eukarya initiators for 41 sequences. From this table, we see that 32 positions in these tRNAs are conserved (all 41 sequences have same nucleotide). So, I assumed that the edit distance for initiator tRNAs should not exceed 44, assuming that the length of tRNAs is 76. I used this assumption as an upper bound to decide about number of cuts in dendogram to make the maximum edit distance in each cluster less than 44!