https://creationismeweersproken.blogspot.com/2023/01/de-rode-panda-en-cserhati-11-clustering.html

THE RED PANDA AND CSERHATI (11): CLUSTERING MTDNA

In addition to WGKS, Cserhati uses mitochondrial DNA to classify the red panda. He uses the full mtDNA as reported in GenBank for 52 species: 15 species and subspecies of the bear family, the two subspecies of the red panda, three species of skunks, 30 species of the mustelid family, and now also two species of the family of the raccoon family, the raccoon itself and the coati *Nasua nasua*.

Cserhati again starts by making a heat map (BMC Genomics Figure 3) to show the correlation matrix between the species. Cserhati writes:

> T*hree larger clusters and two smaller clusters are visible in the heat map.*

That can hardly be right.

It looks more like two really big clusters; or eight or even nine clusters.

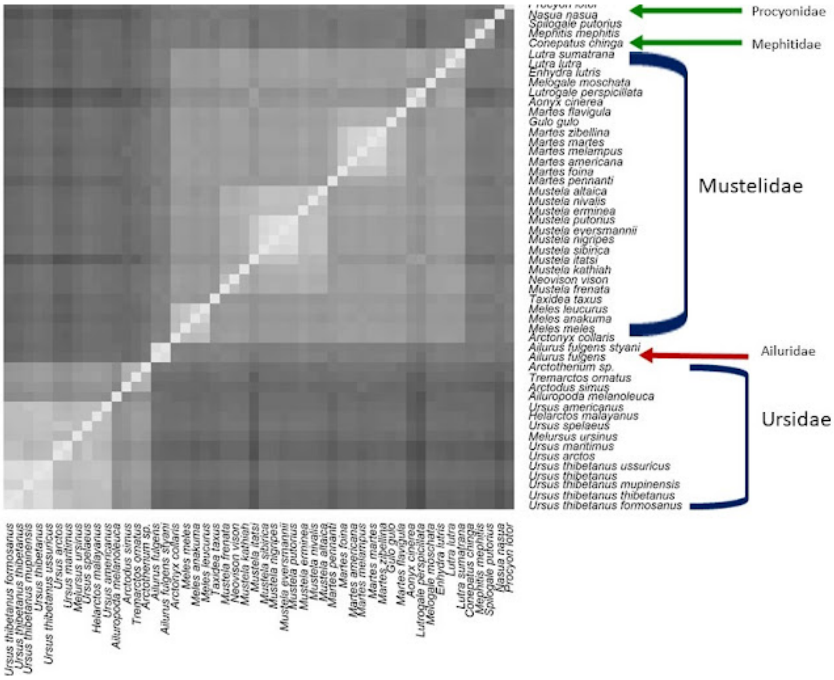


Figure 1. Figure 3 BMC Genomics. Heatmap showing the size of the correlations in mtDNA between the species. The red pandas, family Ailuridae, are plotted here far from the skunks Mephitidae and the raccoon and coatis Procyonidae. Lighter is higher correlation, darker is lower correlation.

Cserhati gives the matrix with correlations in mtDNA between the 52 species in Additional File 2. I copied that correlation matrix in Excel, and sorted it so that the red pandas, the raccoon, the coatis and the skunks are next to each other, and then used the high-low color option in Excel. The following version of Cserhati's matrix emerges:

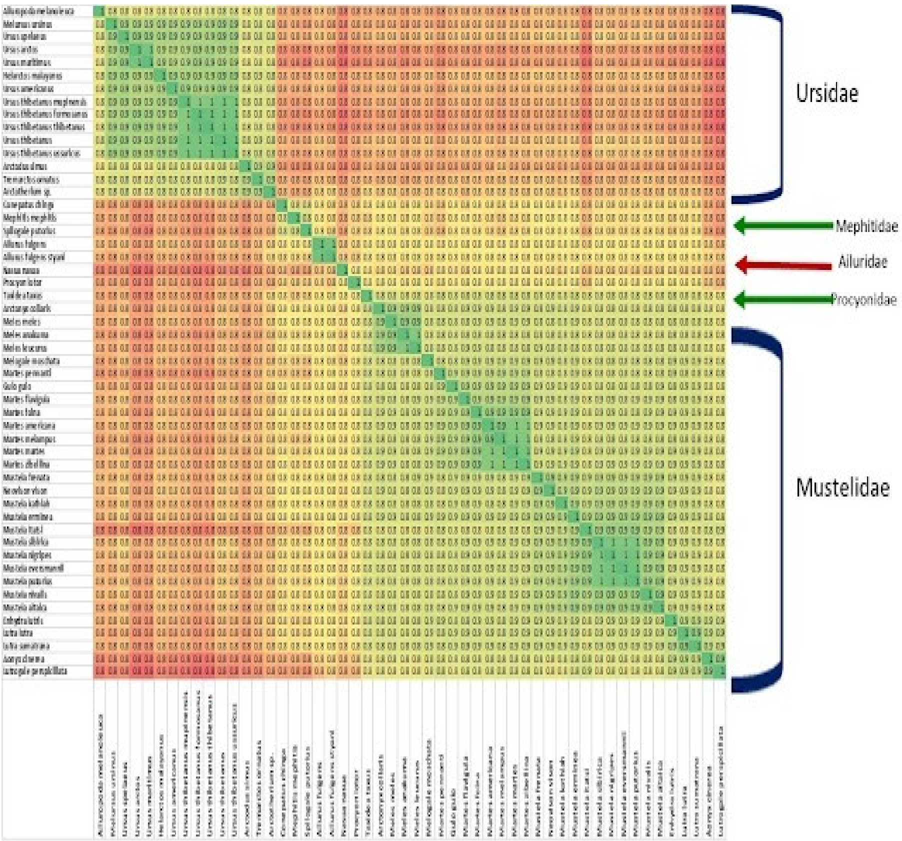
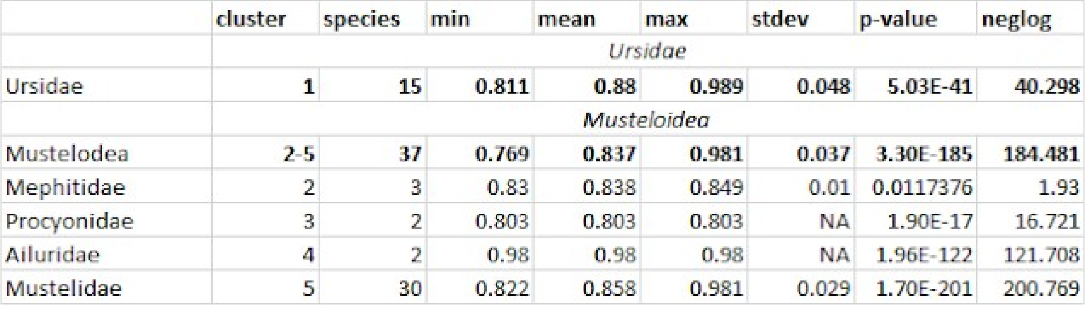


Figure 2. Heatmap showing the size of the correlations in mtDNA between the species. Greener is high correlation, redder is low correlation, yellow in between. Figure 3 of Cserhati BMC genomics with the species in different order, and different color.

The first thing to notice is the division into two large groups: the bears Ursidae and the superfamily Musteloidea with the four families Mephitidae, Ailuridae, Procyonidae and Mustelidae. The contrast yellow / red makes the Musteloidea visible in the heatmap.

The Musteloidea as an important group is reflected in the clustering data given by Cserhati in his Table 3. The first split into clusters is between the Ursidae and the Musteloidea. Two clusters is what the data say.



	cluster	species	min	mean	max	stdev	p-value	neglog
	Ursidae							
Ursidae	**1**	**15**	**0.811**	**0.88**	**0.989**	**0.048**	**5.03E-41**	**40.298**
	Musteloidea							
Mustelodea	**2-5**	**37**	**0.769**	**0.837**	**0.981**	**0.037**	**3.30E-185**	**184.481**
Mephitidae	2	3	0.83	0.838	0.849	0.01	0.0117376	1.93
Procyonidae	3	2	0.803	0.803	0.803	NA	1.90E-17	16.721
Ailuridae	4	2	0.98	0.98	0.98	NA	1.96E-122	121.708
Mustelidae	5	30	0.822	0.858	0.981	0.029	1.70E-201	200.769

Table 3 of Cserhati with minor modification from Additional file 2. Min, Max refer to lowest and highest correlation of a group. Mean and StDev to mean and standard deviation of a group's correlations. P-value must refer to the clusters.

Cserhati moves from two to five clusters, each representing a family. Why does he stop at five clusters? Why doesn't Cserhati come up with clusters within the family Mustelidae?

Look again at the Mustelidae: their correlations are colored just a little greener, not as yellowish as the correlations with the other species of the superfamily Musteloidea. Within the Mustelidae we see beautiful green areas. Perhaps clusters within the Mustelidae? The Mustelidae are a large family with subfamilies. The weasel subfamily gives a cluster as good as the skunk family, for example.

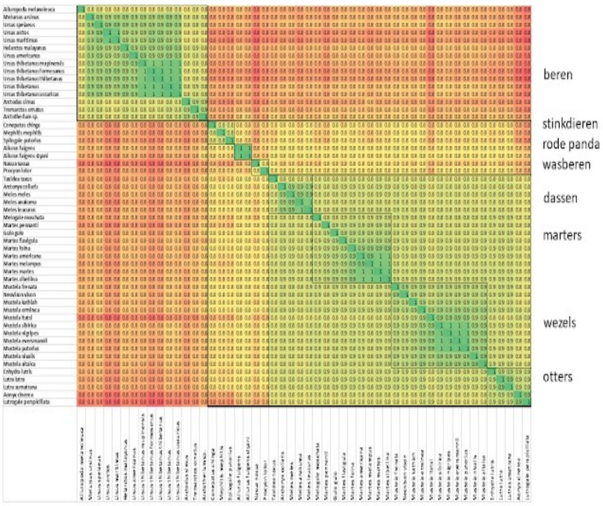


Figure 3. Heatmap showing the size of the correlations in mtDNA between the species. Greener is high correlation, redder is low correlation, yellow in between. Families and subfamilies have now been given a box. Figure 3 from Cserhati BMC Genomics with the species in different order. (beren = bears; stinkdieren = skunks; wasberen = raccoon family; dassen = badgers; marters = martens; wezels= weasels; otters = otters)

How does Cserhati arrive at five clusters?

Not from statistics. Look at Additional file 5: Figure S3 with legend by Cserhati:

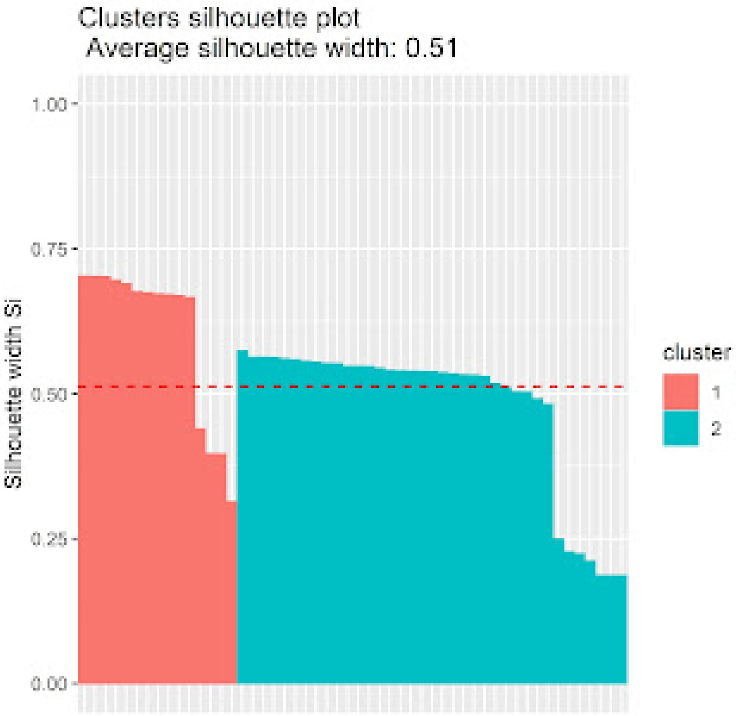


Figure 4. Additional figure 5 BMC Genomics. Plot showing the mean silhouette width according to the number of clusters for the mitochondrial data, based on the ‘silhouette’ method. The maximum average silhouette width is 0.51 for two clusters.

Additional file 5 Figure S3 in the BMC genomics article gives two clustersthe bear family and the superfamily Musteloidea.

Compare this figure with Additional Figure S3 from the CRSQ article, below. Cserhati’s other article, but about the same data. The optimal number of clusters is two. Five clusters is not better than six, seven, eight or ten clusters: all are perfectly good possibilities. I've indicated eight clusters in Figure 3 here, emphasizing the subfamilies within the Mustelidae; five clusters is not a better solution than eight clusters.

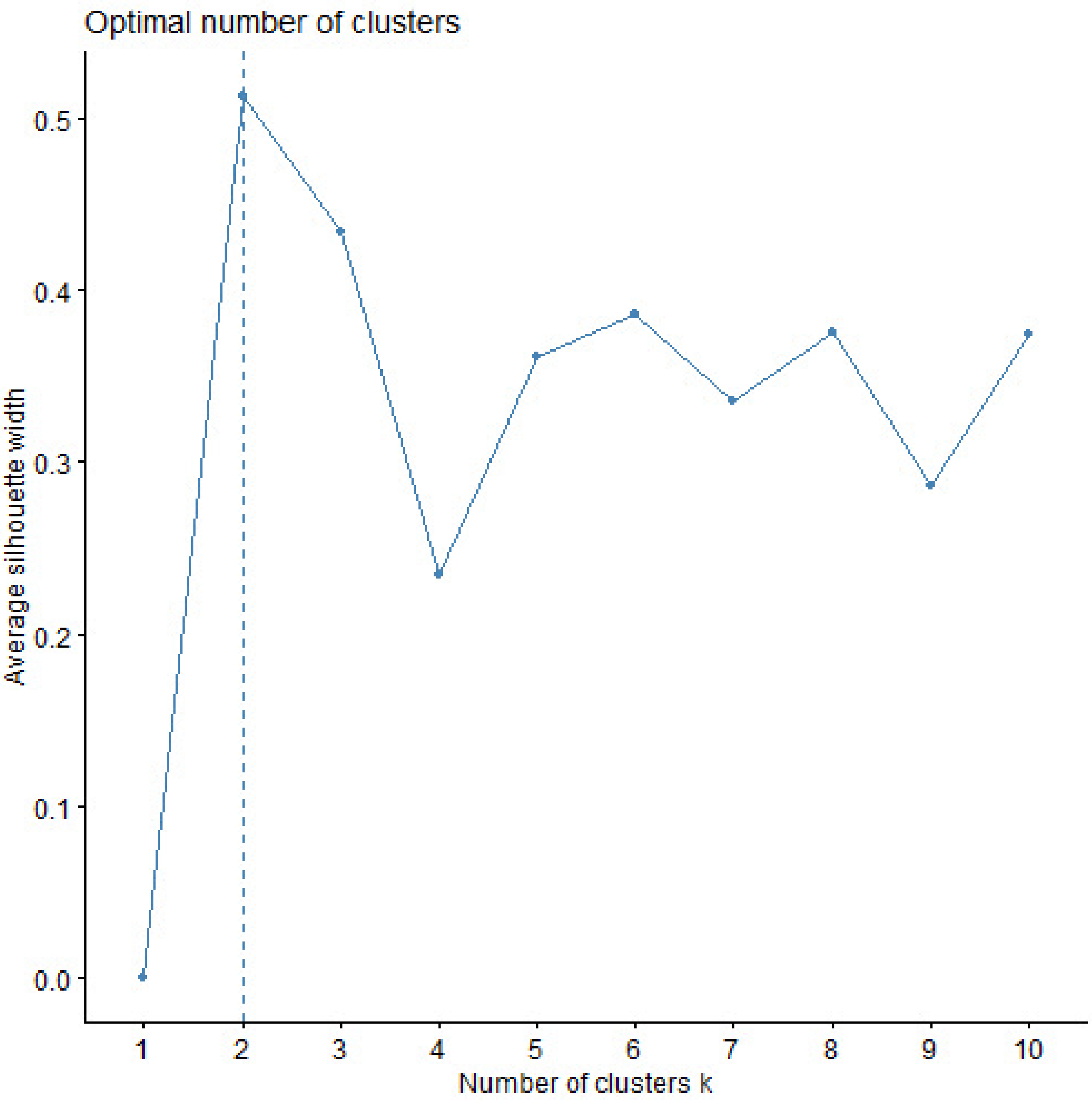


Figure 5 Additional Figure S3 from the CRSQ. Cluster width at different numbers of clusters, and optimal number of clusters, for the mitochondrial data.

Statistically, Cserhati should have stayed with two clusters, family Ursidae and superfamily Musteloidea. Based on clustering, eight clusters is just as good a solution as five clusters.

It might be that Cserhati prefers five clusters because there are species from five families in the data. If so, the clustering was superfluous - the result was known.

***

Cserhati, M., 2021, A tail of two pandas – whole genome k-mer signature analysis of the red panda (Ailurus fulgens) and the Giant panda (Ailuropoda melanoleuca), BMC Genomics 22: 228