

<https://creationismeweersproken.blogspot.com/2023/01/de-rode-panda-en-cserhati-9-clusters-op.html>

THE RED PANDA AND CSERHATI (9): CLUSTERS ON WGKS DATA

Cserhati uses two techniques for red panda and giant panda placement based on the Whole Genome K-mer Signatures: a phylogenetic tree and clustering.

Cserhati does not place much emphasis on the phylogenetic tree based on his WGKS data, but gives ample attention to clustering of the 28 species. Based on clustering, Cserhati says in 'Results and Discussion':

Based on this evidence, A. fulgens would belong to mustelids as a monophyletic group.

In 'Conclusion' Cserhati says:

In conclusion, A. fulgens possibly belongs to Mustelidae, based on the analysis of the WGKS.

The UPGMA phylogenetic tree on WGKS data indicates that the red panda is the sister group of the marten family Mustelidae, but does not belong to the marten family. These statements by Cserhati come from his clustering of the WGKS data.

What does Cserhati do with the WGKS data?

In his analysis he makes a correlation matrix of pairwise correlations of the WGKS data of the species. He displays this matrix in a 'heat map' in which the size of the correlation is shown on a light-dark scale.

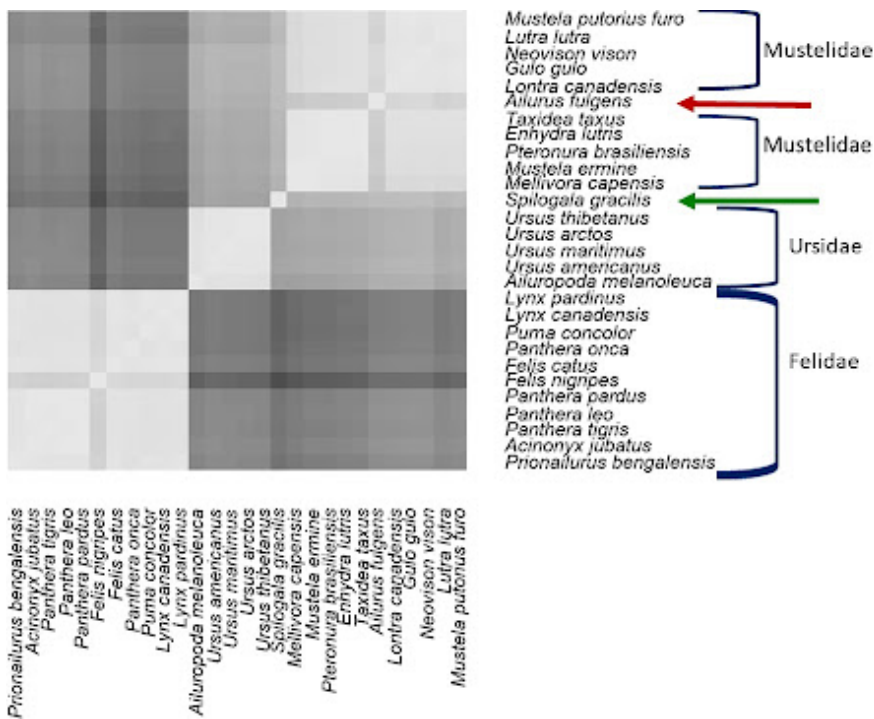


Figure 1 Heat map: pairwise correlations on WGKS data. The order on the x-axis from left to right is the same as the order on the y-axis from bottom to top. The south-west to north-east diagonal gives identity. Red arrow: red panda *Ailurus fulgens*; green arrow skunk *Spilogale gracilis*. Blue: family names. Lighter is higher correlation between species. This is Cserhati fig 1 BMC Genomics with group names added.

How the species order on both axes came about is not mentioned in Cserhati's BMC Genomics article. It is clear that the cats are grouped together, and also the bears and Musteloidea are grouped together. The heatmap thus provides a rough but clear first access to the hierarchical classification of the animals.

It is clear that the red panda differs from the species of the family Mustelidae: we see a 'Finnish flag' pattern: the correlations of the red panda with the species of the marten family are lower than the correlations between the species of the family Mustelidae themselves. That's why we see those dark stripes. But why is the red panda in the middle of the mustelids? No explanation from Cserhati.

Let's take another look at the heatmap. Fortunately, Cserhati has put the numbers of the correlation matrix in one of the supplementary files of the article, so that a colored version can be made in Excel:

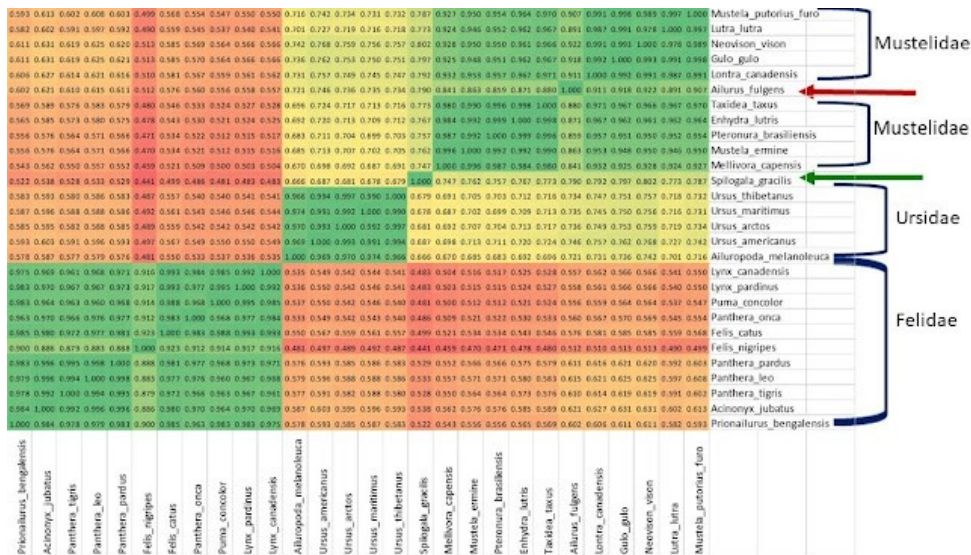


Figure 2 Heatmap according to Cserhati on pairwise correlations on WGKS data. The order on the x-axis from left to right is the same as the order on the y-axis from bottom to top. The south-west to north-east diagonal gives identity. Greener is higher correlation between species, redder is lower correlation between species Red arrow: red panda Ailurus fulgens; green arrow skunk *Spilogale gracilis*. Blue: family names. This is Cserhati fig 1 BMC Genomics with group names added, and colored

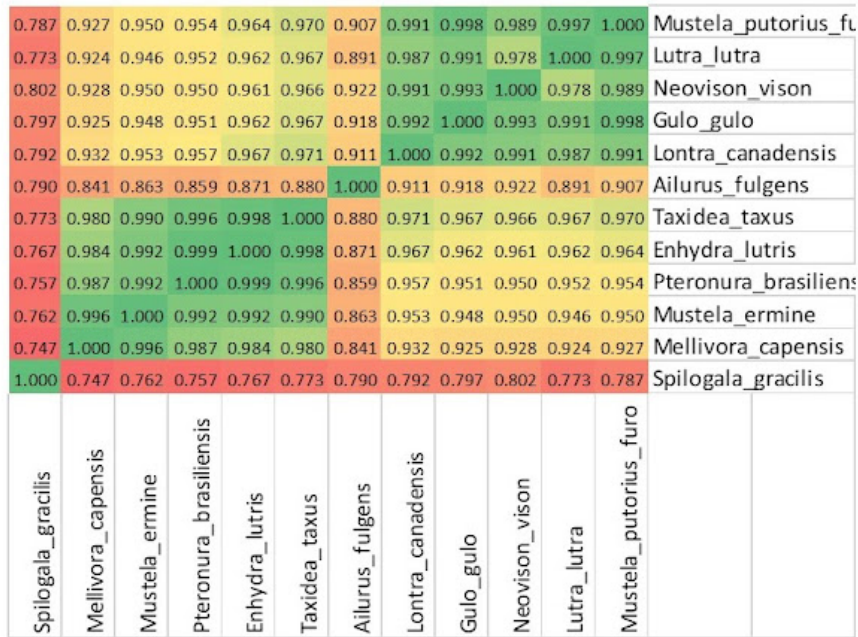


Figure 3 Heatmap according to Cserhati on pairwise correlations on WGKS data of the superfamily Musteloidea. The order on the x-axis from left to right is the same as the order on the y-axis from bottom to top. The south-west to north-east diagonal gives identity. Greener is higher correlation between species, redder is lower correlation between species. This is Cserhati fig 1 BMC Genomics with group names added, and colored

It is clearly visible that the red panda has lower (redder) correlations with the species of the marten family than the species of the marten family have among each other. It is not explained why the red panda *Ailurus fulgens* is placed amongst the species the marten family Mustelidae. Cserhati says:

A. fulgens clearly clusters together with the mustelids, although on average, it has a lower mean PCC value compared to all the other species, 0.89 ± 0.03 , whereas mustelids have a mean PCC value of 0.95 ± 0.04 .

This difference is not too significant.

How significant is "not too significant"? Considering the correlations as independent numbers, we get the mean of all correlations of the red panda with the species of the marten family as 0.89 ± 0.03 , and the mean of all correlations of the species of the marten family among themselves (without red panda) 0.97 ± 0.2 . This is a significant difference. In a two-tailed t-test with unequal variance, this difference is found with a probability of $P = 1.44 * 10^{-6}$. A one-tailed t-test with equal variance gives $P = 3.41 * 10^{-14}$. The WGKS correlations of the red panda with mustelid species differ significantly from the WGKS correlations between mustelids. It remains unclear why Cserhati finds a significance of $P = 1.44 * 10^{-6}$ or $P = 3.41 * 10^{-14}$ "not too significant".

Cserhati uses a clustering program on the data in the correlation matrix, and finds three or four clusters. The cats and bears give clear clusters. The Musteloidea with skunk, red panda and 10 species of mustelids can be considered as a cluster, but Cserhati prefers the skunk on its own and a cluster of 11 species, the 10 species of mustelids with the red panda.

Why is visible in the heatmaps for all species used, as soon as the correlations are sorted by their difference with the red panda:

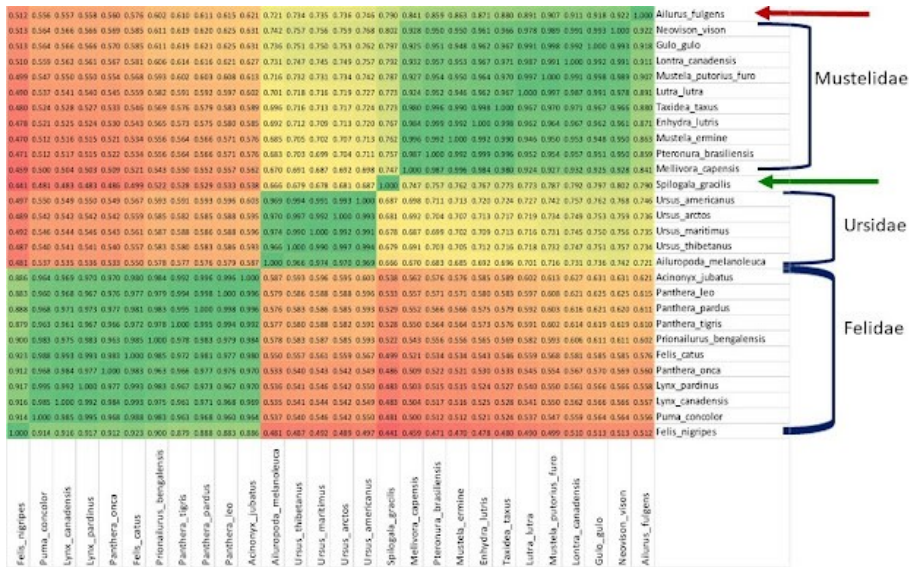


Figure 4 Heatmap on pairwise correlations on WGKS data from Cserhati. The correlations are sorted by similarity to the red panda, which is now far right on the x-axis and highest on the y-axis. The order on the x-axis from left to right is the same as the order on the y-axis from bottom to top. The south-west to north-east diagonal gives identity. Greener is higher correlation between species, redder is lower correlation between species Red arrow: red panda *Ailurus fulgens*; green arrow skunk *Spilogale gracilis*. Blue: family names

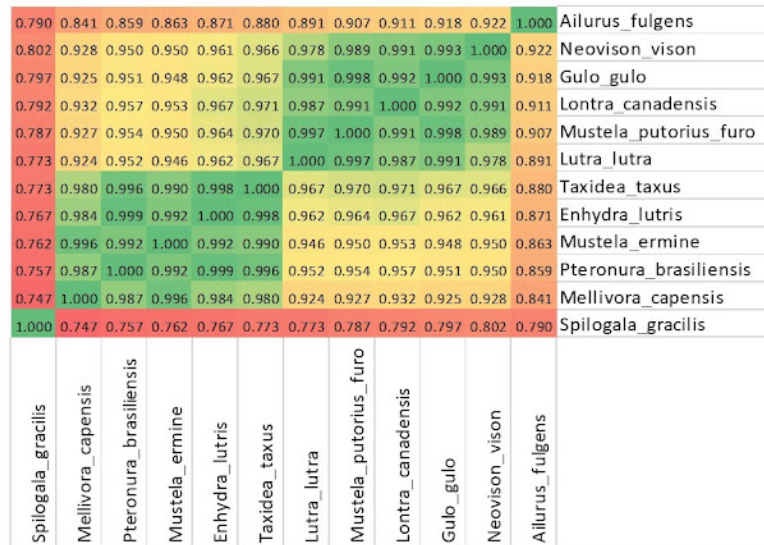


Figure 5 Detail of figure 4, only the 12 species of the Musteloidea. Heatmap on pairwise correlations on WGKS data from Cserhati. The correlations are sorted by similarity to the red panda, which is now far right on the x-axis and furthest on the y-axis. The order on the x-axis from left to right is the same as the order on the y-axis from bottom to top. The south-west to north-east diagonal gives identity. Greener is higher correlation between species, redder is lower correlation between species

When we look at the superfamily Musteloidea in more detail in a heatmap, it is clear that setting the skunk apart is obvious. Not much remains of a cluster of 10 species of mustelids + the red panda: the red panda clearly differs from the mustelids (figure 5). The two groups within the mustelid family are the two groups found within the mustelids in the phylogenetic tree with UPGMA on the WGKS data. (That mess in which the otters and the weasels end up as strange bed fellows.)

All in all:

Although the red panda clusters with the species of the mustelid family when a remote group as the cats are present, the red panda differs from the mustelids; exactly as in the phylogeny on the same data. It is clear from the phylogeny that the red panda and the marten family are monophyletic, but as sister groups: not because the red panda belongs to the marten family Mustelidae. When Cserhati says:

Based on this evidence, A. fulgens would belong to mustelids as a monophyletic group.

the "*monophyletic*" is correct, but "*would belong to mustelids*" not. After all, the family Ailuridae and the family Mustelidae are sistergroups, and that means monophyletic.

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