A tail of two pandas— Whole Genome K-mer Signature Analysis of the Red Panda (Ailurus fulgens) and the Giant Panda (Ailuropoda melanoleuca)

Matyas Cserhati (csmatyi1@gmail.com)
independent scholar

Research article

Keywords: red panda, giant panda, whole genome k-mer signature, Pearson correlation, mustelid, ursid, procyonid, mephitid

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-29891/v2

Version of Record: A version of this preprint was published at BMC Genomics on April 1st, 2021. See the published version at https://doi.org/10.1186/s12864-021-07531-3.

A tail of two pandas— Whole Genome K-mer Signature Analysis of the Red Panda (*Ailurus fulgens*) and the Giant Panda (*Ailuropoda melanoleuca*)

Author: Matthew Cserhati*, independent scholar

*Corresponding author

Email: csmatyi1@gmail.com

Postal address: 5814 N Walnut Grove Ave, San Gabriel, CA 91775

Abstract

Background: The red panda (*Ailurus fulgens*) is a riddle of morphology, making it hard to tell whether it is an ursid, a procyonid, a mustelid, or a member of its own family. Previous genetic studies have given quite contradictory results as to its phylogenetic placement.

Results: A recently developed whole genome-based algorithm, the Whole Genome K-mer Signature algorithm was used to analyze the genomes of 28 species of Carnivora, including *A. fulgens* and several felid, ursid, mustelid, one mephitid species. This algorithm has the advantage of holistically using all the information in the genomes of these species. Being a genomics-based algorithm, it also reduces stochastic error to a minimum. Besides the whole genome, the mitochondrial DNA from 52 mustelids, mephitids, ursids, procyonids as well as *A. fulgens* were also aligned to draw further phylogenetic inferences.

The results from the whole genome study show that *A. fulgens* is a member of the mustelid clade ($p = 9 \cdot 10^{-97}$). *A. fulgens* also separates from the mephitid *Spilogala gracilis*. The giant panda, *Ailuropoda melanoleuca* also clusters away from *A. fulgens*, together with other ursids ($p = 1.2 \cdot 10^{-62}$). This could be due to the geographic isolation of *A. fulgens* from other mustelid species. However, results from the mitochondrial study based on the sequence identity matrix seem to place *A. fulgens* into its own group.

Conclusions: The main conclusion that we can draw from this study is that on a whole genome level *A. fulgens* belongs to the mustelid clade, and not an ursid or a mephitid. This despite the fact that previously some researchers classified *A. fulgens* and *A. melanoleuca* as relatives. Since the genotype determines the phenotype, molecular-based classification takes precedence over morphological classifications. This affirms the results of some previous studies, which studied smaller portions of the genome. The mitochondrial results could be due to differing mutational pressures compared to the nucleus. It cannot be said for sure, but it is likely that *A. fulgens* belongs to the mustelid clade.

Keywords: red panda, giant panda, whole genome k-mer signature, Pearson correlation, mustelid, ursid, procyonid, mephitid

Background

The red panda (*Ailurus fulgens*) is an enigmatic animal and is hard to classify based on its morphology. It lives in parts of India, Nepal and China, and has a distinct red-white coloration, and a striped, bushy tail. It goes by several nicknames, such as the ‘bear-cat’, the ‘cat-bear’, the ‘lesser panda’ or the ‘fire-fox’. Some researchers think *A. fulgens* is a relative of the giant panda (*Ailuropoda melanoleuca*) based on several

physical characteristics. These include an almost exclusive diet of bamboo (both species eat meat on occasion), and have an enlarged radial sesamoid bone, which they use to process bamboo [1,2].

Because of these similarities, the giant panda even received its name from the red panda. According to other opinions, *A. fulgens* has been classified as a member of the family Procyonidae (raccoons). Yet others put the red panda into its own family (Ailuridae) [3]. *A. fulgens* also has some unique characteristics: a large zygomatic arch, a powerful jaw, and complex cheek teeth, following a P2-3 pattern [1].

Previous genetic studies

According to new genetic evidence, there are two species of red panda, the Himalayan red panda (*A. fulgens*), and the Chinese red panda (*A. styani*) [4]. Due to reduced numbers, the red panda is an endangered species. Previous studies based on different combinations of nuclear and mitochondrial genes have given contradictory results as to the taxonomic relationship of *A. fulgens* with other carnivores. This may be because only several mitochondrial and/or nuclear genes were analyzed, and not the entire whole genome sequence (WGS).

The red panda's classification as a procyonid or procyonid-relative is based on immunological, DNA-DNA hybridization, and isozyme evidence [5]. A phylogenetic tree based on Bayesian analysis of cytochrome-b put *A. fulgens* next to Canidae [6].

For example, Peng et al. classify *A. fulgens* either as a mustelid, placing them next to the American marten (*Martes americana*), or as a mephitid, next to the striped skunk (*Mephitis mephitis*). This was based on the analysis of 13 concatenated mitochondrial proteins, based on neighbor-joining (NJ) and maximum likelihood (ML) phylogenetic methods, respectively [7]. In a study of three mtDNA genes (12S rRNA, 16S rRNA and cytochrome b) and intron 1 of the nuclear transthyretin gene, Flynn et al. also found that *A. fulgens* is neither an ursid, nor a procyonid, nor a mephitid, but a mustelid [1]. Another study including three mitochondrial and three nuclear genes by Fulton and Strobeck, based on 16 arctoid species, with *Canis lupus* as an outlier, placed *A. fulgens* in close relationship to *M. mephitis* [8].

Yu and Zhang studied introns 4 and 7 from the nuclear gene ß-fibrinogen (FGB) as well as the mitochondrial gene NADH dehydrogenase subunit 2 (ND2) in 17 species from the order Carnivora. In their results, these researchers found that *A. fulgens* is most closely related to procyonids based on analysis of intron 4 of the FGB gene. But when intron 7 was analyzed, it clustered towards ursids. Classification based on the ND2 gene *A. fulgens* clustered with mustelids, but these results had poor bootstrapping support. When the two introns were combined with analysis of the genes IRBP and TTR, *A. fulgens* was closest to mustelids [9].

Sato et al. analyzed a 5.5 Kbp segment of DNA coding for five genes, AOPB, BRCA1, RAG1, RBP3, and VWF, and found that *A. fulgens* clusters together with procyonids and mustelids, and not with mephitids (skunks and stink badgers) [10]. An earlier, similar result was attained when studying a 3.2 Kbp segment containing the genes APOB, RAG1 and IRBP [11]. Genomically, *A. fulgens* shares several apomorphic chromosome fusions with mustelids, namely F2+C1p and A1p+C1q [12]. However, *A. fulgens* differs in several other chromosomal rearrangements indicating that it diverged early from other mustelids.

Interestingly, several genes have been found in both species which show convergent development. For example, changes in the amino acid composition of the DYNC2H1 and PCNT proteins lead to polydactyly in humans and mice, but to the pseudothumb in the giant and red pandas. Three other convergent genes (PRSS1, PRSS36, and CPB1) are responsible for more efficient uptake of nutrients from bamboo, which makes up a large part of their diet as well. Four other genes, ADH1C, CYP3A5, CYP4F2, and GIF also

enable the more effective utilization in the giant and red pandas of vitamins A and B12 as well as arachidonic acid, which are absent or very low in bamboo [2].

Intron analysis is useful, since these sequences are not under mutational constraint. An analysis of 22 Kbp of nuclear intron sequences from 16 carnivore species groups *A. fulgens* with Musteloidea *sensu stricto* (Mustelidae+Procyonidae) to the exclusion of mephitids [13]. These results, however, contradict results coming from mtDNA analyses [14].

Principle of analysis

Since morphology-based classification of *A. fulgens* is ambiguous, it would be helpful to determine the precise taxonomic status of this species based on a whole genome-based algorithm. To this end, the Whole Genome K-mer Signature (WGKS) algorithm [15] is used to analyze the genomes of five bear species, eleven cat species and ten species from the family Mustelidae (weasels, otters, martens, and badgers), *Spilogala gracilis*, a mephitid species, as well as *A. fulgens*, making 28 species in total.

The advantages of using a genomics-based algorithm to analyze the WGS of these organisms is that it takes all the information present in the WGS, as opposed to just a handful of genes, utilized in gene studies. Deciding which genes are important is subjective and may vary between investigators. Whole genome-based algorithms also have the advantage that they greatly reduce stochastic error, due to the vast number of characters (DNA bases) that they analyze [16]. Using this algorithm can provide additive results as to the phylogenetic classification of *A. fulgens*.

While the WGKS algorithm may not be a *sensu stricto* phylogenetic algorithm, it can still be used to classify species, based on their WGS into different groups. There are several metagenomics methods, which use k-mer analysis to classify Next-generation read sequences, such as kraken [17], the Naïve Bayes Classifier (NBC) [18], and PhymmBL [19]. For example, kraken splits read sequences into k-mers, which it then maps to a taxonomic tree. The leaf node/species which has the most reads assigned to it is the designated as the species that the read came from. The NBC also splits a read into constituent N-mers, and then calculates the a posteriori probability of a given N-mer belonging to a specific strain, species, genus, or other taxon.

The NBC algorithm and the WGKS algorithm are similar in that they both utilize the k-mer signature of a DNA sequence in order to classify it. One could view the whole genome sequence as a very extended read sequence. Using k-mer methods on whole genome sequences (WGS) should give even more accurate results than on read sequences because the WGS represents a much larger search space. Individual k-mers occur in much larger numbers than in short reads, which are between 75-300 bp or so. In other words, the k-mer 'coverage' is much, much higher in a WGS than a single read.

Besides a whole genome approach, it would also be useful to complement the results from the whole genome analysis using a multiple alignment of several genes. To this end the mitochondrial DNA of 52 ursid, mephitid, mustelid, procyonid species along with the ailuronids, *A. fulgens* and *A. fulgens styani* were analyzed. Not only does the mtDNA contain more than a dozen conserved genes, these genes are localized to the same part of the genome and also largely follow the same order. The mtDNA also contains non-coding DNA, which is not under mutational constraint, and thus better reflects species relationships. Mitochondrial genes would be more conducive to this kind of analysis as opposed to artificially concatenating together genes from different parts of the genome. These mtDNA sequences were aligned using the online MUSCLE tool at the EBI website.

Results and Discussion

Pre-clustering analysis

The list of species and the PCC matrix can be seen in Additional File 1 online. The Hopkins statistic is 0.9, which means that the data set is of very good quality for clustering. The silhouette plot (Supplementary figures 1 and 2) gave a maximum average silhouette width of 0.82 for three clusters. This value was 0.8 for four clusters. The average silhouette width was studied for two to seven clusters. The only difference was the placement of the mephitid, *S. gracilis* into its own group (cluster 4 in Supplemental figure 2).

Whole genome analysis

In Figure 1 we can see three visible clusters, felids, ursids and mustelids, with *S. gracilis* in between the mustelids and the ursids. Based on the results in Table 1, *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. If we compare *Felis nigripes* (the black-footed cat) with other cats, it has a mean PCC value of 0.89±0.02, whereas felids having an even greater mean PCC of 0.97±0.03. Yet we know that cats are a monophyletic group. Table 2 shows the minimum, mean, maximum PCC for all three putative clades, as well as the p-value, which is statistically significant for all three groups.

Based on this evidence, *A. fulgens* would belong to mustelids as a monophyletic group. Since it has such a low mean PCC is because it may have diverged early from other mustelids, possibly due to its isolated mountainous habitat in parts of Myanmar, Burma and China. This can also be seen well in Figure 2, which shows the UPGMA-based phylogenetic tree for the 28 species in the whole genome analysis.

Also important is that the skunk species *S. gracilis* does not cluster with mustelids. When compared with mustelids, *S. gracilis* has a mean PCC value of 0.78±0.02. *A. fulgens* has a PCC value of 0.79 with this species as opposed to a mean PCC value of 0.89 with mustelids, reported previously. This also indicates that mustelids and mephitids form separate clades.

The giant panda, *Ailuropoda melanoleuca* is a clearly a member of a clade which includes the ursids, as shown in Figure 2. It has a mean PCC value of 0.97±0.003 with the other ursids. Other genetic evidence classifies the giant panda as a member of Ursidae. This includes mtDNA, chromosome banding patterns, and serological and immunological evidence [20, 21].

Analysis of mitochondrial genomes

The result of the analysis of the mitochondrial genomes can be seen in Figure 3. The Hopkins clustering statistic is 0.841, which means that the sequence identity matrix is of good quality for clustering. Three larger clusters and two smaller clusters are visible in the heat map. The clusters and statistics for these five groups are available in the ‘clusters’ tab of Additional File 2, and Table 3 respectively. The list of species, accession numbers, and the results from this analysis are also available online at github in Supplementary File 2.

Figure 4 depicts a hierarchical tree, showing the position of the different clades. Ursids and Musteloidea form two large clades, with 15 and 37 species, respectively. Within Musteloidea we have three smaller groups besides Mustelidae. The first one consists of both species of *A. fulgens.* The second is made up of three mephitids, *S. gracilis*, *M. mephitis*, and *Conepatus chinga*. Lastly, two procyonids, *Procyon lotor* (raccoon), and *Nasua nasua* (ring-tailed coati) make up the third group. Supplementary figure 3 shows the average silhouette width according to the number of clusters, with an average silhouette width of 0.51 for two clusters.

Ledje et al. [3] also found that *A. fulgens* was distinct from all other caniforms, and placed it in its own monotypic family. However, this analysis was based on the analysis of only the mitochondrial 12S rRNA gene. Flynn et al. also reached a similar conclusion based on the analysis of three mitochondrial genes [1]. On the other hand, Peng et al. [7] classified *A. fulgens* as a mustelid, based on the analysis of thirteen concatenated mitochondrial proteins.

These results may seem to contradict the results of the WGKS analysis, by placing *A. fulgens* into its own group, apart from mustelids. Let us bear in mind, that even though the mitochondrial genome is a good way to study multi-gene alignments, it is still only a fraction of the entire genome. This may be due to something which is known as mito-nuclear discordance, the difference between selective and mutational pressures between nuclear and mitochondrial DNA, due to migration patterns. Mito-nuclear discordance is a common phenomenon [22, 23, 24]. We must also remember that *A. fulgens* is a geographically isolated species, which may lead to its genetic isolation from other mustelids as well. Fulton and Strobeck analyzed four nuclear sequence-tagged sites and one exon of the gene IRBP within 79 carnivore species, and also found discordant results between the results of the mitochondrial and nuclear analyses. In their study, the mtDNA results supported the monophyly of Ailuridae and Mephitidae, whereas the nuclear results suggested otherwise [25].

Conclusion

In conclusion, *A. fulgens* probably belongs to the mustelids, based on the analysis of the WGKS. This species also clusters away from *S. gracilis*, indicating that mustelids and mephitids belong to separate clades, which is reinforced by the mtDNA results as well. This is based on whole genome data as opposed to the contradictory results in previous studies involving just a handful of genes, one even in two different exons of the same gene. This demonstrates the utility of the WGKS algorithm, which takes a holistic approach of analyzing the WGS. The mtDNA results appear to place *A. fulgens* into its own group, but this could be due to mito-nuclear discordance, which is a common phenomenon. *A. melanoleuca*, on the other hand, belongs to the ursids, as shown consistently in both the WGS results as well as the mtDNA results.

Methods

Data and programs used

The Python script motif_analysis_k-1.py at github.com/csmatyi/motif_analysis was used to generate WGKS profiles. Version 3.6.0. of R was used. The heatmap was generated using the R command ‘heatmap’, using the ‘ward.D’ clustering algorithm for the WGKS analysis, and the ‘single’ algorithm for the mitochondrial data. Clusters were generated using the ‘cutree’ command and were depicted in hierarchical trees using the UPGMA method [26]. To determine the optimal number of clusters, the ‘cluster’ and ’factoextra’ libraries and the fviz_nbclust command were used, setting the method parameter to ‘wss’. The

'fviz_silhouette' plot was used to construct the Silhouette plot. The 52 complete mitochondrial genome Refseq sequences were downloaded from the nucleotide database at NCBI. Additional Excel files and figures as well as the mitochondrial genome fasta file can be found online at github.com/csmatyi/ailurus.

Description of algorithm

The WGKS algorithm that was used in the analysis is an alignment-free k-mer sequence comparison method [27]. These methods involve the statistical comparison of k-mers between species. A k-mer is a segment of DNA k bp long, which can correspond to the core segment of a transcription factor binding site, a repeat element or other regulatory element. These elements take part in protein binding and gene regulation and are conserved across different species. The advantages of using a k-mer based alignment-free algorithms over alignment-based ones is that they process input much faster and are unbiased by guide trees imposed upon the data [28, 29].

For a lengthy description of the algorithm, the reader is referred to Cserhati et al., 2019 [15]. However, a short description is provided here for better understanding. The WGKS algorithm is divided into three steps.

First, all possible k-2, k-1, and k-mers in the genome of a given species are enumerated to give the observed occurrence O. Then, based on these observed occurrences, the expected occurrence E can also be calculated with the following equation:

$$(1)\ E_k = O_{1..k-1} \bullet O_{2..k}/O_{2..k}$$

where $O_{1..k}$ is the expected occurrence of the k-mer, $O_{1..k-1}$ is the expected occurrence of the k-1-mer from positions 1 to k-1, $O_{2..k}$ is the expected occurrence of the k-1-mer from positions 2 to k, and $O_{2..k-1}$ is the expected occurrence of the k-2-mer from positions 2 to k-1.

The score value S can be calculated in the following way:

$$(2)\ S_{k-mer} = \frac{O-E}{O+E}$$

Score values can be interpreted in three ways:

$$(3)\ O \gg E : S_{k-mer} \to 1 \text{ (overrepresented k-mer)}$$

$$(4)\ O \ll E : S_{k-mer} \to -1 \quad \text{(underrepresented k-mer)}$$

$$(5)\ O = E : S_{k-mer} \approx 0 \quad \text{(randomly occurring k-mer)}$$

Even if the genome is partially or completely duplicated, then the score value will not change. This is because both the Observed and Expected values will increase by the proportion that the duplicated genome is compared to the pre-duplication genome.

The next step involves comparing the k-mer signature between two species. The k-mer signature is simply a list of all k-mers ordered in lexicographical order from AA…A to TT…T, together with their score values. For a given value k, there are 4^k possible k-mers. Thus, the k-mer signature also corresponds to a vector of 4^k numbers. Since octamers were analyzed, this corresponds to 65,536 possible octamers. Two of these vectors can be compared to one another for two different species using the Pearson Correlation Coefficient

(PCC). PCC values closer to 1 represent a pair of closely related species, within the same clade. Lower PCC values denote two unrelated species. This step is performed between all possible pairs of species to derive a square PCC matrix. P-values for clusters were calculated by comparing the PCC values for all species pairs within the cluster with all PCC values for all species pairs where one species came from the cluster, and the other species was outside the cluster.

The last step involves visualizing the PCC in a heatmap and using clustering algorithms to detect monophyletic groups. Clustering can be done for example using the k-means clustering algorithm, or the Partitioning Among Medoids (PAM) algorithm.

Mitochondrial DNA analysis

The 52 complete mitochondrial genome sequences for the ursid, mephitid, mustelid, procyonid species and the two *A. fulgens* species were aligned using the online MUSCLE tool [30], version 3.8 at ebi.ac.uk/Tools/msa/muscle using default parameters. The sequence identity matrix was derived from the alignment using BioEdit, version 7.2.5 [31].

Abbreviations

Ailuronid: a member of the family Ailuronidae, the red panda.

Mephitid: a member of the family Mephitidae, or skunks.

mtDNA: mitochondrial DNA.

Mustelidae: a member of the family Mustelidae, or a group of animals including weasels, otters, ferrets, minks, martens and wolverines.

PCC: Pearson Correlation Coefficient.

Procyonid: a member of the family Procyonidae, or raccoons.

UPGMA: unweighted pair group method with arithmetic mean, am agglomerative hierarchical clustering method.

Ursid: a member of the family Ursidae, or bears.

WGS: whole genome sequence.

WGKS: whole genome k-mer signature.

Declarations:

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The Python script motif_analysis_k-1.py at github.com/csmatyi/motif_analysis was used to generate WGKS profiles. Additional Excel files and figures can be found online at github.com/csmatyi/ailurus.

Competing interests

The author declares that he has no competing interests.

Funding

No funding was used for this study.

Author contributions

M.C. designed the whole study, ran all calculations and scripts, and wrote the article.

Acknowledgements

No specific acknowledgements.

Tables

Table 1. Classification of the 28 species used in the WGKS analysis.

Species	**group**
Ailurus fulgens	1
Enhydra lutris	1
Gulo gulo	1
Lontra canadensis	1
Lutra lutra	1
Mellivora capensis	1
Mustela ermine	1
Mustela putorius furo	1
Neovison vison	1
Pteronura brasiliensis	1
Taxidea taxus	1
Ailuropoda melanoleuca	2
Ursus americanus	2
Ursus arctos	2
Ursus maritimus	2
Ursus thibetanus	2
Acinonyx jubatus	3
Felis catus	3
Felis nigripes	3
Lynx canadensis	3
Lynx pardinus	3
Panthera leo	3
Panthera onca	3
Panthera pardus	3

Panthera tigris	3
Prionailurus bengalensis	3
Puma concolor	3
Spilogala gracilis	4

Table 2. Statistical measures for each of the three clusters in the WGKS analysis.

group	**name**	**no. species**	**min**	**mean**	**max**	**stdev**	**p-value**
1	mustelids	11	0.841	0.954	0.999	0.04	8.97E-97
2	ursids	5	0.966	0.983	0.997	0.012	1.23E-62
3	felids	11	0.879	0.965	0.998	0.032	6.17E-95

Table 3. Statistical measures for each of the five clusters in the mitochondrial analysis.

group	name	no. species	min	mean	max	stdev	p-value
Ursidae							
1	ursids	15	0.811	0.880	0.989	0.048	5.03E-41
Musteloidea							
2-5	Musteloidea	37	0.837	0.769	0.837	0.981	3.3E-185
2	mustelids	30	0.822	0.858	0.981	0.029	1.7E-201
3	ailuronids	2	0.980	0.980	0.980	NA	1.96E-122
4	procyonids	2	0.803	0.803	0.803	NA	1.9E-17
5	mephitids	3	0.830	0.838	0.849	0.01	0.012

Figure legends

Figure 1. Heatmap depicting group relationships for 28 species based on results from the WGKS algorithm. Brighter colors represent species pairs which are in the same group, with a PCC value closer to 1. Darker colors represent species pairs which are in different group, with a PCC less than 1.

Figure 2. UPGMA-based hierarchical tree for the 28 species based on PCC values. Ursids, felids, and mustelids form separate clades, with *S. gracilis* in its own group.

Figure 3. Heatmap depicting group relationships for 52 carnivore species based on alignment of the mitochondrial genome using the online MUSCLE software. Brighter colors represent species pairs which are in the same group, with a sequence identity closer to 1. Darker colors represent species pairs which are in different group, with a sequence identity closer to 0.

Figure 4. UPGMA-based hierarchical tree for the 52 species analyzed in the mtDNA study, based on sequence identity metrics. Mustelids and ursids form two large clades, and mephitids, procyonids forming two small groups. *A. fulgens* and *A. fulgens styani* appear either to form their own clade, or loosely associate with mustelids.

Supplementary Figure 1. Silhouette plot for three clusters from the WGKS analysis. The average silhouette width is 0.82.

Supplementary Figure 2. Silhouette plot for four clusters. The average silhouette width is 0.8.

Supplementary Figure 3. 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 Files

Additional File 1: Results of whole genome analysis of 28 species. The files includes a list of species, and the genome sequence files downloaded from NCBI, the PCC matrix which is a result of the WGKS algorithm, as well as the species clusters and the cluster statistics.

Additional File 2: Results of the alignment of mitochondrial DNA from 52 carnivore species. This file includes a species list, the sequence identity matrix, species clustering information and cluster statistics.

References

1. Flynn JJ, Nedbal MA, Dragoo JW, Honeycutt RL. Whence the red panda? Mol Phylogenet Evol. 2000;17:190–99.
2. Hu Y, Wu Q, Ma S, Ma T, Shan L, Wang X, et al. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. Proc Natl Acad Sci U S A. 2017;114:1081–1086.
3. Ledje C and Arnason U. Phylogenetic relationships within caniform carnivores based on analyses of the mitochondrial 12S rRNA gene. Journal of molecular evolution 1996;43(6):641–649.
4. Hu Y, Thapa A, Fan H, Ma T, Wu Q, Ma S, et al. Genomic evidence for two phylogenetic species and long-term population bottlenecks in red pandas. Sci. Adv. 2020:6,eaax5751.
5. Wei F, Hu Y, Zhu L, Bruford MW, Zhan X, Zhang L. Black and white and read all over: the past, present and future of giant panda genetics. Mol Ecol. 2012;21:5660–74.
6. Agnarsson I, Kuntner M, May-Collado LJ. Dogs, cats, and kin: a molecular species-level phylogeny of Carnivora. Mol Phylogenet Evol. 2010;54:726–45.
7. Peng R, Zeng B, Meng X, Yue B, Zhang Z, Zou F. The complete mitochondrial genome and phylogenetic analysis of the giant panda (*Ailuropoda melanoleuca*). Gene 2017;397:76–83.
8. Fulton TL and Strobeck C. Novel phylogeny of the raccoon family (Procyonidae: Carnivora) based on nuclear and mitochondrial DNA evidence. Mol Phylogenet Evol. 2017;43:1171–77.
9. Yu L and Zhang YP. Phylogeny of the caniform carnivora: evidence from multiple genes. Genetica 2006;127:65-79.
10. Sato JJ, Wolsan M, Minami S, Hosoda T, Sinaga MH, Hiyama K. Deciphering and dating the red panda's ancestry and early adaptive radiation of Musteloidea. Mol Phylogenet Evol. 2009;53:907–22.
11. Sato JJ, Wolsan M, Suzuki H, Hosoda T, Yamaguchi Y, Hiyama K, et al. Evidence from nuclear DNA sequences sheds light on the phylogenetic relationships of Pinnipedia: single origin with affinity to Musteloidea. Zoolog Sci. 2006;23:125–46.
12. Nie W, Wang J, O'Brien PC, Fu B, Ying T, Ferguson-Smith MA, et al. The genome phylogeny of domestic cat, red panda and five mustelid species revealed by comparative chromosome painting and G-banding. Chromosome Res. 2002;10:209–22.
13. Yu L, Luan PT, Jin W, Ryder OA, Chemnick LG, Davis HA, et al. Phylogenetic utility of nuclear introns in interfamilial relationships of Caniformia (order Carnivora). Syst Biol. 2011;60:175–87.
14. Delisle I. and Strobeck C. A phylogeny of the Caniformia (order Carnivora) based on 12 complete protein-coding mitochondrial genes. Mol. Phylogenet. Evol. 2005;37:192–201.
15. Cserhati M, Xiao P and Guda C. K-mer based motif analysis in insect species across *Anopheles*, *Drosophila* and *Glossina* genera and its application to species classification. Computational and Mathematical Methods in Medicine 2019;4259479.

16. Heath TA, Zwickl DJ, Kim J, Hillis DM Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst Biol. 2008;57:160–66.
17. Wood DE and Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology 2014;15(3)

18. Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B. Metagenome fragment classification using N-mer frequency profiles. Advances in bioinformatics, 2008;205969.

19. Brady A and Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature methods, 2009;6(9):673–676.

20. Krause J, Unger T, Noçon A, Malaspinas AS, Kolokotronis SO, Stiller M., et al. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol Biol. 2008;8:220.
21. Yu L, Li YW, Ryder OA, Zhang YP. Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. BMC Evol Biol. 2007;7:198.
22. Toews DP and Brelsford A. The biogeography of mitochondrial and nuclear discordance in animals. Molecular ecology 2012;21(16):3907–3930.

23. Morales HE, Pavlova A, Joseph L and Sunnucks P. Positive and purifying selection in mitochondrial genomes of a bird with mitonuclear discordance. Molecular ecology 2015;24(11):2820–2837.

24. Bernardo PH, Sánchez-Ramírez S, Sánchez-Pacheco S.J, Álvarez-Castañeda ST, Aguilera-Miller EF, Mendez-de la Cruz FR, Murphy RW. Extreme mito-nuclear discordance in a peninsular lizard: the role of drift, selection, and climate. Heredity 2019;123(3):359–370.

25. Fulton TL and Strobeck C. Molecular phylogeny of the Arctoidea (Carnivora): effect of missing data on supertree and supermatrix analyses of multiple gene data sets. Molecular phylogenetics and evolution 2016;41(1), 165–181.
26. Mitchener CD and Sokal RR. A quantitative approach to a problem in classification. Evolution 1956; 11:130–162.
27. Vinga S and Almeida J. Alignment-free sequence comparison-a review, Bioinformatics 2003;19:513–23.
28. Pollard DA, Iyer VN, Moses AM, Eisen MB. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2006;27:e173.
29. Yang K and Zhang L. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Res. 2008;36:e33.
30. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 2004;32(5):1792–1797.
31. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. 1999;Symp. Ser. 41:95-98.

Figures

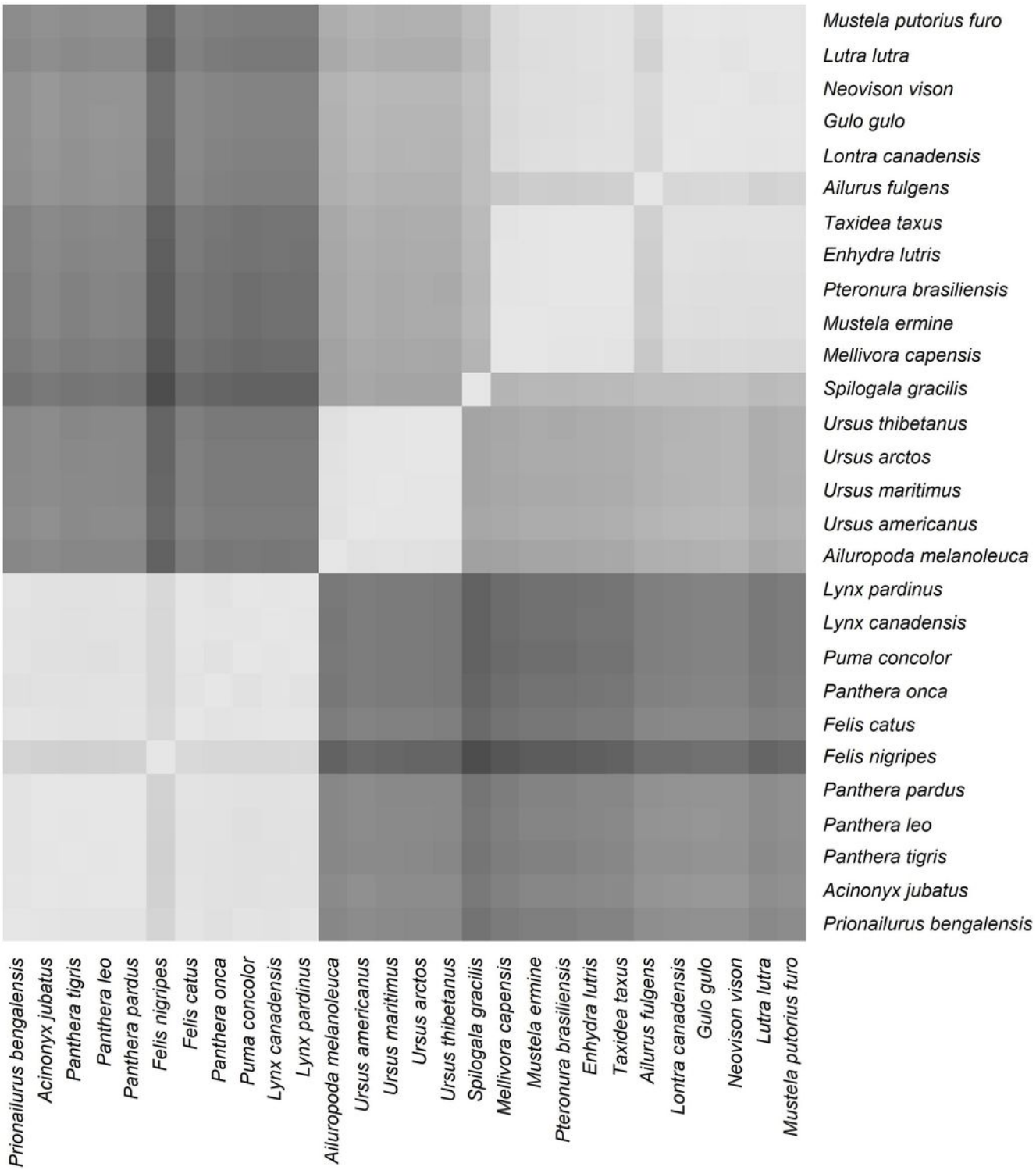

Figure 1

Heatmap depicting group relationships for 28 species based on results from the WGKS algorithm. Brighter colors represent species pairs which are in the same group, with a PCC value closer to 1. Darker colors represent species pairs which are in different group, with a PCC less than 1.

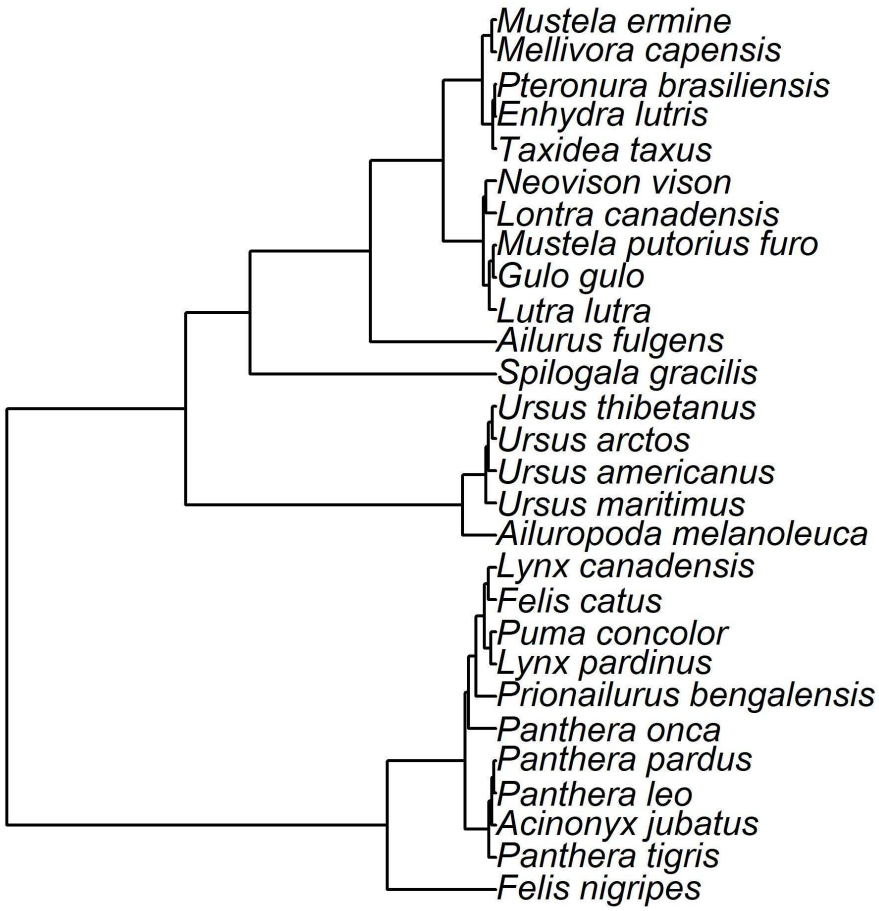

Figure 2

UPGMA-based hierarchical tree for the 28 species based on PCC values. Ursids, felids, and mustelids form separate clades, with S. gracilis in its own group.

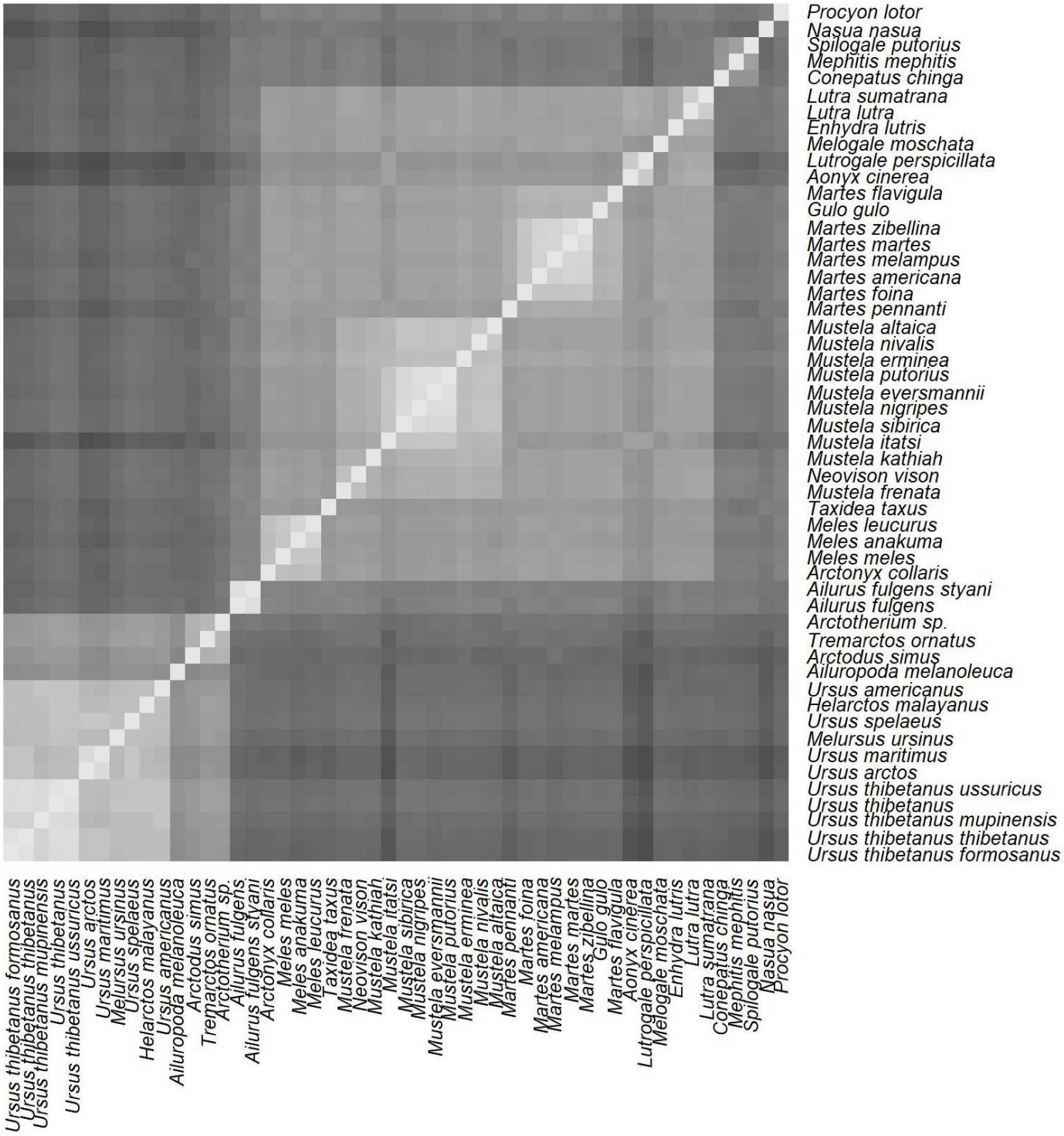

Figure 3

Heatmap depicting group relationships for 52 carnivore species based on alignment of the mitochondrial genome using the online MUSCLE software. Brighter colors represent species pairs which are in the same group, with a sequence identity closer to 1. Darker colors represent species pairs which are in different group, with a sequence identity closer to 0.

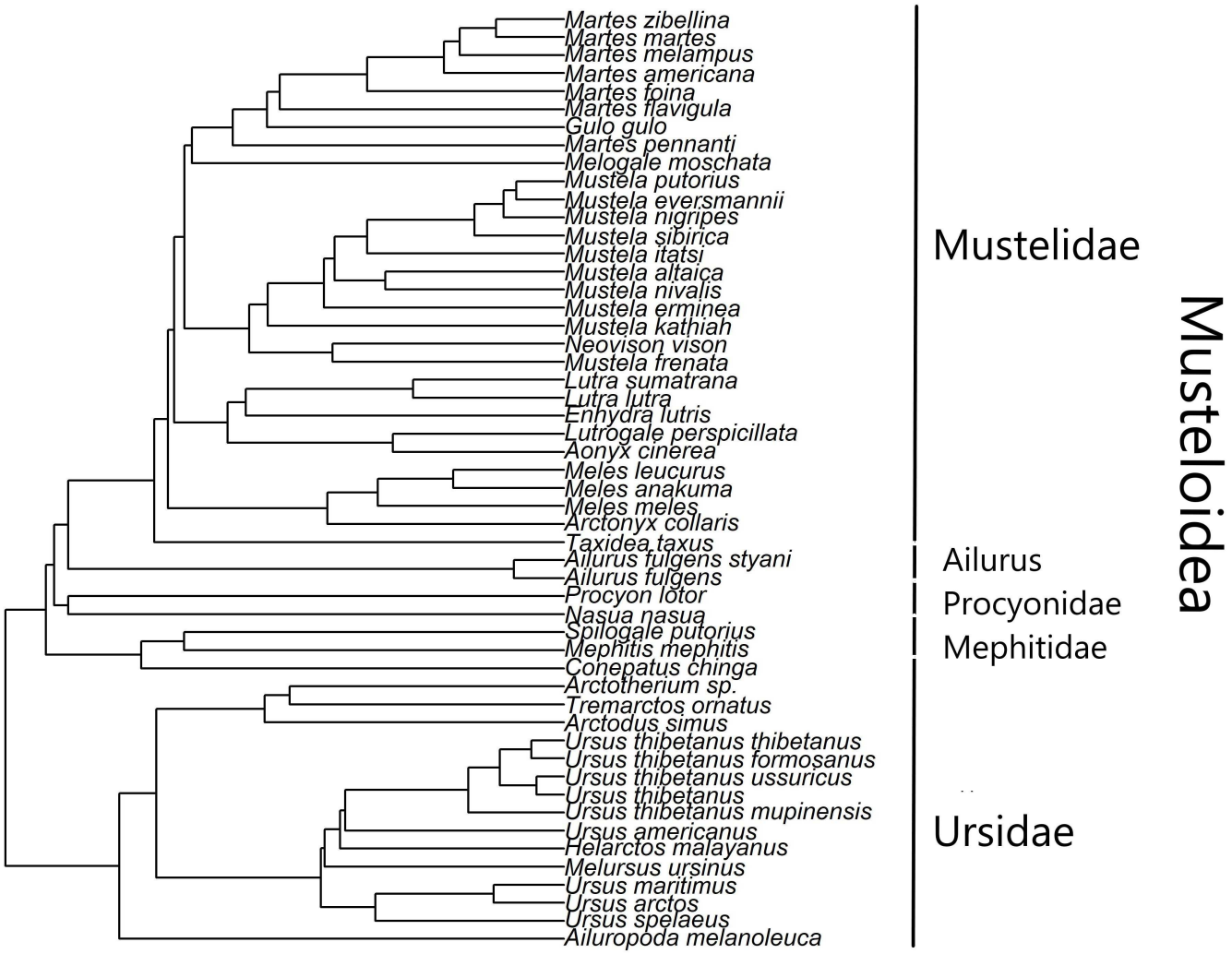

Figure 4

UPGMA-based hierarchical tree for the 52 species analyzed in the mtDNA study, based on sequence identity metrics. Mustelids and ursids form two large clades, and mephitids, procyonids forming two small groups. A. fulgens and A. fulgens styani appear either to form their own clade, or loosely associate with mustelids.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SFig1.tiff
- SFig2.tiff
- SFig3.tiff
- AdditionalFile1.xlsx
- AdditionalFile2.xlsx