|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| GenBank ID | Species (Haplogroup) | Radiocarbon Date | Median BEAST calibrated date | Location | Genome Coverage (% 1sf) | Published paper |
| MH662441.1 | Aardwolf | Present Day | - | - | 100.0 | [1] |
| MH662442.1 | Aardwolf | Present Day | - | - | 100.0 | [1] |
| MH662443.1 | Aardwolf | Present Day | - | - | 100.0 | [1] |
| MH662444.1 | Aardwolf | Present Day | - | - | 100.0 | [1] |
| MF593938.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593939.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593940.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593941.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593942.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593943.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593944.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593945.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593946.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593947.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593948.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593949.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593950.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593951.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| MF593952.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| NC\_038159.1 | Brown Hyaena | Present Day | - | - | 100.0 | [1] |
| JF894376.1 | Striped Hyaena | Present Day | - | - | 100.0 | [1] |
| NC\_020669.1 | Striped Hyaena | Present Day | - | - | 100.0 | [1] |
| MN320467.1 | Cave and Spotted Hyaena (Haplogroup A) | Present Day | - | Somalia | 100.0 | [1] |
| MN320458.1 | Cave and Spotted Hyaena (Haplogroup A) | 40,770+-350 | 44,323 | Horosno-Vynjava site (Ukraine) | 86.8 | [1] |
| JF894377.1 | Cave and Spotted Hyaena (Haplogroup A) | Present Day | - | Kira Zoo, France | 100.0 | [3] |
| JF894378.1 (NC\_020670.1) | Cave and Spotted Hyaena (Haplogroup A) | Late Pleistocene | 25,230 | Coumere Cave, Ariege (France) | 100.0 | [3] |
| JF894379.1 | Cave and Spotted Hyaena (Haplogroup A) | Late Pleistocene | 19,241 | Coumere Cave, Ariege (France) | 100.0 | [3] |
| KU937361.1 | Cave and Spotted Hyaena (Haplogroup A) | Late Pleistocene | 94,344 | Grotte-aux-Ours, Lot (France) | 94.1 | [4] |
| MN320459.1 | Cave and Spotted Hyaena (Haplogroup A) | Late Pleistocene | 23,474 | AufhausenerHöhle (Baden Württemberg, Germany | 98.3 | [1] |
| MN320463.1 | Cave and Spotted Hyaena (Haplogroup A) | Present Day | - | Ghana | 100.0 | [1] |
| MN320464.1 | Cave and Spotted Hyaena (Haplogroup A) | Present Day | - | Tierpark, Berlin (Germany) | 100.0 | [1] |
| MN320461.1 | Cave Hyaena (Haplogroup B) | 26, 820+-100 | 30,967 | LindenthalerHöhle (Gera, Germany) | 97.6 | [1] |
| MN320462.1 | Cave Hyaena (Haplogroup B) | 18,030+- 50 | 23,798 | LindenthalerHöhle (Gera, Germany) | 97.6 | [1] |
| MN320456.1 | Cave Hyaena (Haplogroup B) | 51,200 +-4,000 | 51,200 | Certovapec (Slovakia) | 98.1 | [1] |
| MN320449.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Botswana | 100.0 | [1] |
| MN320450.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Zambia | 100.0 | [1] |
| MN320451.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Tanzania | 100.0 | [1] |
| MN320452.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Tanzania | 100.0 | [1] |
| MN320453.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Kenya | 100.0 | [1] |
| MN320454.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Kenya | 100.0 | [1] |
| MN320455.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Kenya | 100.0 | [1] |
| MN320465.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Tierpark,  Berlin,  Germany | 100.0 | [1] |
| MN320466.1 | Spotted Hyaena (Haplogroup C) | Present Day | - | Namibia | 100.0 | [1] |
| MN320457.1 | Cave Hyaena (Haplogroup D) | Late Pleistocene | 56,115 | Tonghe Bridge, Zhaodong County in Heilongjiang Province (Northeastern China) | 96.2 | [1] |
| MN320460.1 | Cave Hyaena (Haplogroup D) | 48,650 +- 4,000 | 48,650 | Geographical society cave (Eastern Russia) | 97.6 | [1] |
| MT880602.1 | Cave Hyaena (Haplogroup D) | 16,780+-40 | 20,240 | North eastern China | 89.6 | [2] |
| MT880603.1 | Cave Hyaena (Haplogroup D) | 16,790+-50 | 20,252 | North eastern China | 92.9 | [2] |

[1] - Hyena paleogenomes reveal a complex evolutionary history of cross-continental gene flow between spotted and cave hyena, (2020), Westbury et al.

[2] - Ancient mitochondrial genomes from Chinese cave hyenas provide insights into the evolutionary history of the genus Crocuta, (2021), Jiaming Hu et al.

[3] - Coprolites as a source of information on the genome and diet of the cave hyena, (2012), Bon C et al.

[4] - Genome data on the extinct Bison schoetensacki establish it as a sister species of the extant European bison (Bison bonasus), (2017), Palacio P et al.

**Phylogenetic Mapping Using These Files:**

See pdf file.

**Short summary of the current understanding of Hyaena Genetics:**

The Hyaenidae family is the smallest carnivora group extant today. In the late Miocene the species had a much greater genetic diversity comparatively, with a predicted peak of 80-100 co-existing species[ref]. Despite the small number of species today they inhabit a wide range of biological niche roles with the Aardwolf (Proteles cristata) inhabiting a solitary insectivore lifestyle, with the other three species belonging to the sub-family Hyaenidae which specialise in bone-crunching jaws that allow for the digestion of bone marrow from carcasses. The two species Striped hyena (Hyaena hyaena) and brown hyaena (Parahyaena hyeana) are exclusively kleptoparasites, whilst the species Spotted hyena (Crocuta crocuta) is a hunter alongside being a kleptoparasite. The current genetic studies are heavily biased towards spotted hyena due to their recent ecological and biological changes that coincided with their speciation into the extinct cave hyena found abundantly throughout Eurasia at excavated cave sites, due to their long term occupation and teeth marks left behind on bones found there. In 2005 [ref] Rohland et al. looked at the mitochondrial gene cyt b among extinct cave hyena found throughout Eurasia and sequenced extant spotted hyena in Africa. They identified 4 Haplogroups (A, B, C, and D) that represented different geographical areas (North Africa/European, Pure European, South African, East Asian respectively). Interestingly they found that the cave and spotted hyena samples where genetically mixed together not forming reciprocally monophyletic clades. From this study they hypothesized an out of Africa model that showed 3 migration events (figure 1). Later a full mitochondrial analysis in 2020 identified the same groups however found that the dispersion order was slightly different (figure 2) with the pure European diverging from Africa prior to the B, which contains the North African and European mix. They also found that the haplogroup B showed a reciprocally monophyletic clade with the north African and European branching off ~400 ka. Two Chinese hyena samples were later added on to the cave hyena mitochondria samples and showed the genetic diversity of the east Asian samples to be alike to that of the haplogroup B which included the two North African and European clades, showing a possibility of a second east Asian group. This could possibly be interesting to see when more East Asian samples get sequenced to see how many haplotypes they are able to find. The same study [ref] was also able to sequence full nuclear genomes of ancient bone and teeth fossils and constructed PCA graphs that showed clear genetic variation patterns between the cave and spotted hyaena, forming two reciprocally monophyletic clades. To support this they created phylogenetic trees from 2 million base pair non-overlapping sequences that also showed the separate grouping of cave and spotted hyena. They showed that the cave hyena came from one initial migration into east Asia which then eventually migrated into Eurasia where it had contact with Spotted hyena due to the genetic evidence showing introgression events between the two groups (figure). They found two events with an initial bidirectional gene transfer and then followed by a subsequent unidirectional gene transfer into spotted hyena. It has been hypothesized that the reason for the indistinct mitochondrial results between cave and spotted hyaena is from this introgression. A large number of the genes from these introgression events in the extant Spotted hyaena are involved with central nervous system (reference) this may have been crucial in there survival with their social groupings being the most complex of any species known ranging from 80-120(reference). In 2021 a paper looking into the cave at Denisova identified diagnostic positions in the mitochondria to identify unknown sedimentary sequences to a species and showed a temporal change in the hyaena species prevalent at the cave in the late Pleistocene to early neolithic (figure). This is what I am writing about in greater documentation. In their tests they used datasets available in 2019, therefore I will include the extra two haplogroup D mitogenomes that have been sequenced since. Finally I will then repeat this looking to see if there are big changes for the quicksand pipeline that is less conservative with its filtering and see if there are any clear signs of false positive results (figure 4).

|  |
| --- |
| A diagram of a graph  Description automatically generated with medium confidenceA map of the world  Description automatically generated  **D**  **A**  **B**  **C** |
| Figure 1) |

|  |
| --- |
| **C**  **A**  **B**  **D** |
| Figure 2) |

|  |
| --- |
| **D**  **A/B**  **C/A** |
| Figure 3) |

**Method:**

A diagram of a diagram of a shelf

Description automatically generated with medium confidence

Multiple Sequence Alignment:

After research, identified the best full mitochondrial genomes available in genbank for the Hyaenidae family, for this study I decided to use the Proteles cristata as the basal clade rather than including a feline as I think this would remove a lot of the possible data. Most of the data was from [1] but there was recently a paper showing 2 new Chinese full mitogenomes that were included to supplement the haplogroup D inferred fixed genotype. Each species/group had atleast 4 DNA sequences used in the multiple alignment.

After collecting I aligned them all to the Crocuta crocuta reference genome (NC\_020670.1) and these will be used as coordinates for the diagnostic positions later. Then I ensured that the mitochondrial genomes all started at the same location by cutting and pasting overhangs from regions that hadn’t aligned quite properly. Then using ClipKit, that removes regions in multiple sequence alignments that have poor conservation (lower than 50%). Using these clipped genomes in MegaX, I performed a bootstrap 500 phylogenetic reconstruction to find similar results to what was expected (see pdf file).

Next I identified the diagnostic positions for the four crocuta haplogroups, which worked by identifying diagnostic positions based on ancestral and derived states figure 1. For this it was important to try and reduce the number of false positives by removing tri-allelic mutations as well as regions that showed poor DNA preservation , eg regions of DNA repeats that could be poorly aligned and often just artefacts from the computational binding. The coordinates of these corresponded to where they aligned to the reference Crocuta crocuta sample.

Then using .csv file with the records for the 2021 Denisova paper on previous work done on the sedimentary DNA. I re-ran the samples on the new diagnostic positions which now included the 2 extra haplogroup D samples see workflow for method, and then from these got a count of the number of sequences for each specific haplogroup of spotted hyena were counted. Then I got the proportion values and made a graph see figure. Removing these positions that had above 2 standard deviations of the mean count. Leaving with my filtered diagnostic positions for which I re-ran on the same data set to get my final output files, which I concatenated together for each haplogroup.  
Finally, I created a script which read in all these concat files simultaneously along with data files that held: all ID values, the appropriate marker for each sampleID, and another file with the appropriate time gaps for each layer. (see in /r1/people/Thomas\_harris\_snell/hyena\_mtDNA/crocuta/python\_scripts/) excluding SampleID whose derived count : ancestral count ratio was significantly lower than 10% as well as samples that didn’t have at least 3 diagnostic positions covered. Then discarding any duplicate samples, I then counted each of the samples and looked at how the coverage of each changed over time. See results (figure). After doing this I then looked at how the MEGAN pipeline compared with that of quicksand (figure 2).

Results:

|  |  |
| --- | --- |
| A graph with red and blue lines  Description automatically generated | A graph with red and blue lines  Description automatically generated |
| A graph with red lines  Description automatically generated | A graph with numbers and lines  Description automatically generated |
| Figure) A,B,C,D showing the diagnostic positions that had more than two times the sequence coverage that were removed to reduce sampling bias from the mitochondrial capture method. | |

A graph with different colored lines

Description automatically generated with medium confidence

7798 = number of A sequences 360 = B 0 = C 8717 = D

This was my initial attempt, in this I counted each of the individual reads from each sample. This introduces a strong bias towards the most evolutionary distinct haplogroup, which in this case is D. As they have more possible positions therefore more likely of getting hits. Then in future I counted the number of samples depending on if they had positive results for a specific haplogroup, this helped reduce the bias.

A graph of different colored bars

Description automatically generated with medium confidence

Using Elena dp and file input

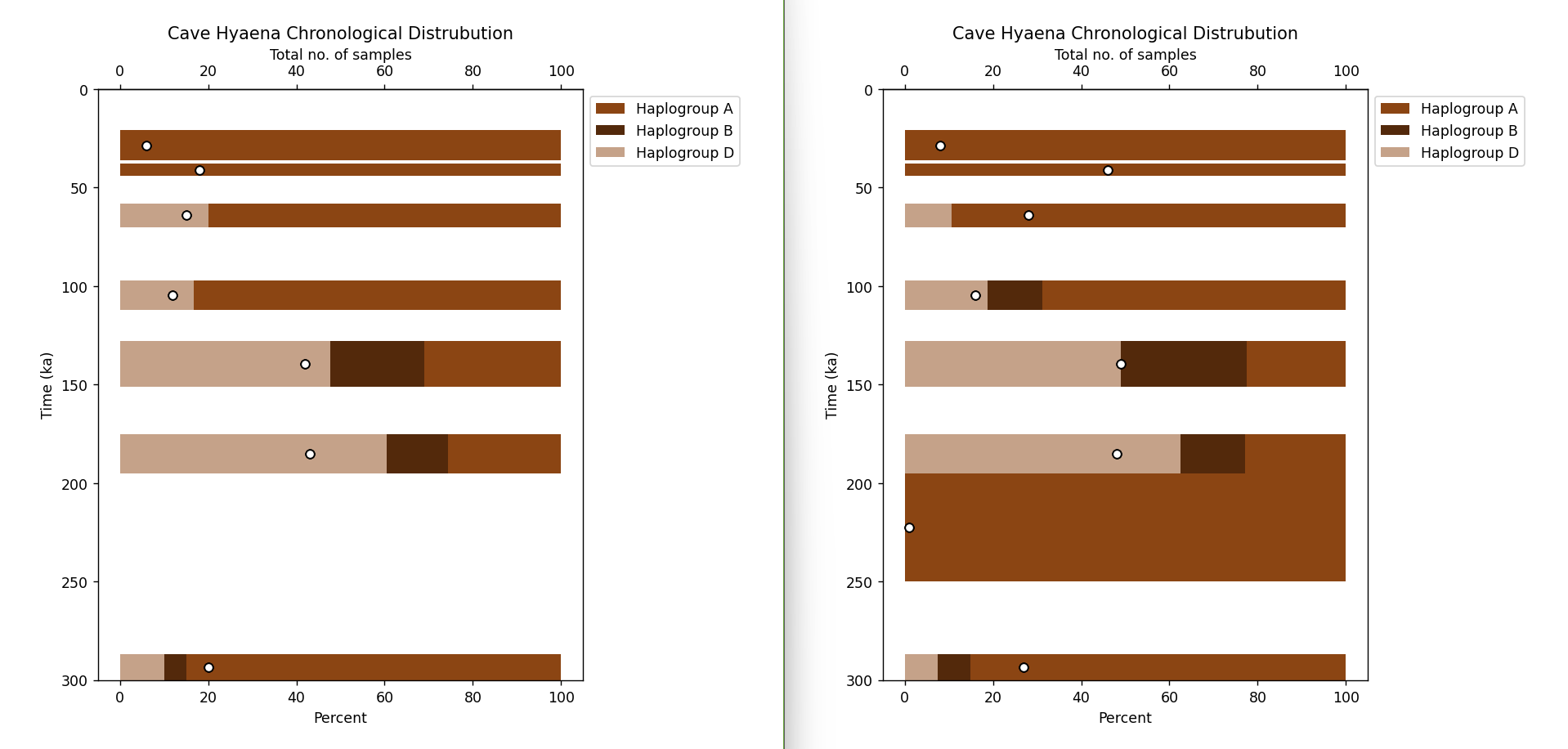
A screenshot of a computer

Description automatically generated

Using mine and Elena dp, but my file input

Clear that the datafile used in her output isn’t == read\_me.rtf file that I found. Would be nice to know exactly what files she used. Seems as if more data used for the earlier files. As for the difference between my diagnostic positions and Elena’s I see a higher haplogroup B percentage, but otherwise seems very similar.

(left MEGAN, right quicksand)



This is the output with my diagnostic positions.

Notes:

Higher number of samples, in pretty much every layer.

Higher number of haplogroup B representation (found in in 110ka)

Same trend A, A, D, D, A, A, A, A

Consistent results with MEGAN.