Resolution of basal wolf-like canid divergence using ancient DNA and fossil data

EEOB 563 Project

Bruno do Rosario Petrucci

March 30, 2021

1 Background

The dire wolf (*Canis dirus*) has long been a source of wonder for researchers and the general public alike. While much work has been put into the study of its fossil remains, many have put into question whether its morphological similarities with the grey wolf (*Canis lupus*) are a result of close evolutionary relationships, or convergence. This is a difficult question to address, given extinct species rarely leave remains that allow for the sequencing of genetic data, which would be necessary for the construction of phylogenetic trees that could help untangle the dire wolf's relationship with the rest of the wolf-like canids (i.e. the genera *Canis*, *Cuon*, and *Lycaon*). Recent work has sequences the dire wolf DNA for a number of well-preserved specimens, however, and put us one step closer to understanding these complex relationships [1].

In this study, it became clear the dire wolf was not phylogenetically close to the other representatives of the *Canis* genus, with the dhole (*Cuon alpinus*) and the African wild dog (*Lycaon pictus*) sharing a more recent common ancestor with the wolf than the dire wolf and the jackals clade (*Canis mesomelas* and *Canis adustus*). The question of whether the latter clade should be in the *Canis* genus is also a frequently debated one, with the genus *Lupulella* being increasingly adopted for that group instead. These results supported that conclusion, and furthermore provided support to the hypothesis that the dire wolf should not be on the *Canis* genus either. The molecular data in the study failed, however, to satisfactorily resolve the relationship between *dirus*, the (*mesomelas, adustus*) group and the rest of the wolf-like canids, and it is not clear whether the dire wolf or the jackals diverged earlier in the wolf-like canids group.

Previous work has shown integrating morphology and fossil sampling data into phylogenetic inference can improve divergence time estimations (e.g. [2]), so that total-evidence analysis is a possible path to untangle these relationships. In this project, I intend to make use of the Fossilized-Birth-Death process [3] in RevBayes [4] to attempt to resolve the divergence order of the basal wolf-like canids. I will utilize nuclear DNA sequence, morphology data, and fossil samples from extant and extinct canids from the *Canis*, *Cuon* and *Lycaon* genera.

2 Data

I intend to use the nuclear genetic data from the original dire wolf paper [1]. It required correspondence with the authors, since the data is not published, but I have acquired the final alignment used on the paper, with 5 specimens of dire wolf, 3 specimens of grey wolf, 2 of African wild dog, and 1 specimen of each other of the 12 species total used in the nuclear DNA analysis. This data contains transversions only, which must be taken into account when selecting a substitution model. This data contains around 12 million SNPs, and the dire wolf DNA is highly damaged, so that some treatment will be required. First, I delete the rows for three of the dire wolf samples, since they are damaged enough that if we attempt to take sites with no missing data, we'd have less than 10 thousand bases, and for the Andean fox (*Lycalopex culpaeus*), since it is an outgroup (we have two), and we are missing its morphological data. Then, I delete every site that contains a missing ("N") value. Next, I keep only sites where the repeated species' sequences (two dire wolves, two African Wild dogs and three wolves) agree with each other. I do the same with the African golden wolf (*Canis lupaster*) and the golden jackal (*Canis aureus*) since I only have morphological data for one of them (*aureus*), and given their previous classification as conspecific this should not lead to too many deletions. I then keep only one of each of the repeated groups (the choice does not matter, since they are all the same due to the previous procedure). This leaves me with 569583 bases (keeping only samples that agree between repeated specimens deleting only 54046, or less than 10%). Finally, I cut around 70 thousand more by cutting the sites with all equal sequences. This leaves me with **499353** bases, which seems like a good number to continue the analysis with a clean data set. I might also delete singleton sites, given their pipeline (see SI for the study), but not sure yet since I am still in correspondence with the authors to understand the data better.

For fossil samples, I will used the paleobioDB package to find fossil occurrences from the paleobiology database. I downloaded all samples with genus *Urocyon*, *Canis*, *Cuon*, and *Lycaon*. Then, I kept the samples for which the species name is accepted, to make sure we are not included misidentified genera in the analysis. Then, I maintain only one fossil sample with a unique combination of taxon name, minimum age, and maximum age. This is due to the FBD model, where we cannot include more than one fossil sample with the same species and age range. Finally, I deleted species for which we had neither morphological nor dna data, since these would not contribute to the analysis. This left me with 119 fossil samples. These include both extant (+ dire wolf) and extinct species, so that some species in the data set will have morphological data only. The FBD model estimates, jointly with the topology, branch lengths and divergence times, the posterior probability that a fossil is a sampled ancestor (i.e. a fossil from a species that descended into one of the extant species) or an extinct tip.

For the morphological data, I intend to use the data matrix compiled by Graham Slater for a select group of extinct and extant canids [5]. This is a scoring of 123 characters for a group of canids that includes 10 species in our extant (+ dire wolf) data set, the two left out being the andean fox and African golden wolf. It also includes 13 species from our fossil samples, but only from the *Canis* and *Urocyon* genera (possibly due to the lower size and more recent classification of *Cuon* and *Lycaon*). For the identity of each morphological character, see the SI for the aforementioned study. They

focus on mandibule and teeth morphology, though span a reasonable range of characters that are important for canid phylogenetics. Given we chose only nuclear data sites where *Canis lupaster* and *Canis aureus* overlap, it is not an issue to have to "delete" *lupaster* from our analysis (note the study did the same with *aureus* instead - their position in the topology should be interchangeable). I will also use the *Cuon javanicus* morphological data as the data for *Cuon alpinus*, since the former is now considered a subspecies of the latter. Both of these would require more justification for a published study, of course. I further treated the data by deleting characters for which we had the same value for all species in the data set, since these would not be informative. This leaves us with **62** characters for the analysis. Note that while our data set only has 23 species, this file will contain data for 129 taxon (119 fossils + 10 extant), since each fossil sample also has to be represented here (with a character vector equal to the vector for its species)

As a summary, our data will be: a .nex file containing 499353 bases of SNP data from 10 different species, being 8 extant wolf-like canids, the dire wolf, and the gray fox as an outgroup; a .nex file containing 62 morphological character data from 129 taxon spanning 23 different species, being our 10 DNA sequenced species, 11 extinct *Canis* and 2 extinct *Urocyon* species; a .tsv file for our taxon data frame, which contains 1. a taxon column, 2. a species column (since we have more than one occurrence for a couple of species, this is necessary to link fossils and species), 3. minimum age and 4. maximum age from each fossil occurrence. More details about the necessary inputs can be found at the FBD tutorials in the RevBayes website FBD tutorials in the RevBayes website.

3 Methods

As mentioned, I will use the FBD model in RevBayes for this analysis. This model jointly estimates the posterior probability of a set of topology, branch lengths, divergence times, sampled ancestor sampling times, and fossil tips, from a set of sequence, morphology, and fossil occurrences data. This is a highly complex model, leading to a high number of parameters in comparison to the data. As such, I plan to use simpler molecular evolution models than usual, with low/no partitioning. I am not set on the details of the analysis yet. I will run a simple analysis on the nuclear data alone to get some tree priors (e.g. trees with > 90% posterior probability). From this analysis, I will also test the tree's sensitivity to priors of diversification-sampling rates, molecular evolution, and topology. Given this knowledge, I will run a simple model selection analysis using BIC and/or posterior probabilities. The models and tree prior selected here will then be used on a full estimate, using the morphology and fossil data. The results of this analysis will be summarized with consensus trees, MAP and MCC trees, and credible intervals of divergence times for the speciation times relevant to our phylogenetic question. As mentioned, the details of this analysis might change as I proceed in the project, since FBD is a complex model and it might take some figuring out to set it up such that we get a tree with low enough uncertainty to answer our question.

References

[1] A. R. Perri, K. J. Mitchell, A. Mouton, S. Álvarez Carretero, A. Hulme-Beaman, J. Haile, A. Jamieson, J. Meachen, A. T. Lin, B. W. Schubert, C. Ameen, E. E. Antipina, P. Bover, S. Brace, A. Carmagnini, C. Carøe, J. A. Samaniego Castruita, J. C. Chatters, K. Dobney, M. dos Reis, A. Evin, P. Gaubert, S. Gopalakrishnan, G. Gower, H. Heiniger, K. M. Helgen, J. Kapp, P. A. Kosintsev, A. Linderholm, A. T. Ozga, S. Presslee, A. T. Salis, N. F. Saremi, C. Shew, K. Skerry, D. E. Taranenko, M. Thompson, M. V. Sablin, Y. V. Kuzmin, M. J. Collins, M.-H. S. Sinding, M. T. P. Gilbert, A. C. Stone, B. Shapiro, B. Van Valkenburgh, R. K. Wayne, G. Larson, A. Cooper, and L. A. F. Frantz, "Dire wolves were the last of an ancient New World canid lineage," *Nature*, vol. 591, pp. 87–91, Mar. 2021. Number: 7848 Publisher: Nature Publishing Group.

[2] C. Zhang, T. Stadler, S. Klopfstein, T. A. Heath, and F. Ronquist, "Total-Evidence Dating under the Fossilized Birth–Death Process," *Systematic Biology*, vol. 65, pp. 228–249, Mar. 2016.

[3] T. A. Heath, J. P. Huelsenbeck, and T. Stadler, "The fossilized birth-death process for coherent calibration of divergence-time estimates," *Proceedings of the National Academy of Sciences*, vol. 111, pp. E2957–E2966, July 2014.

[4] S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist, "RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language," *Systematic Biology*, vol. 65, pp. 726–736, July 2016.

[5] G. J. Slater, "Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 4897–4902, Apr. 2015. Publisher: National Academy of Sciences Section: Biological Sciences.