Development of multiple whole genome species alignments to investigate the radiation of *Bovidae*

Supervisors	Christian Gross[a,b], Dick de Ridder[b], Marcel Reinders[a]
Type	Data analysis
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Programming
Timestamp	December 20, 2018

Description

The genome database Ensembl[1] published several whole genome species alignments (WGSA) by using their EPO (Enredo-Pecan-Ortheus)[2,3] pipeline. These alignments are a great asset to trace the development of phenotypical traits across the evolution of species and they are the basis for CADD (Combined Annotation Dependent Depletion) methodology [4,5]. CADD is a methodology that enables the evaluation of short variants in the genomes of target species according to their deleteriousness (likelihood that the variant has a negative effect on fitness) by utilizing differences between evolutionary expected and observed variations.

The exact selection pressure is constantly changing and depends on environmental constraints and intra species dynamics. The idea is that by making use of differently deep phylogenies, CADD models could be generated that score periods of particular selection differently which would help to trace the development of phenotypical traits from the past to the current populations.

This project would be conducted by investigating the evolutionary history of *Bovidae*. The Family of *Bovidae* contain between ~140-~280 extant species[6] with several species of high economical value such as Goat, Sheep and Cattle. Despite the large number of high quality genomes and the economical value of some of their members, WGSAs have not been constructed for *Bovidae*, partly due to inherently NP-hard nature of WGSA that make them infeasible for any large number of species. The complexity of the problem is increasing by $O(n^k)$ for k sequences of length n. Therefore your task as a student would be to develop/deploy heuristics that would make the creation of WGSA feasible for the investigation of the *Bovidae* family history via the CADD methodology.

References

1. Daniel R. Zerbino et al. Ensembl 2018. PubMed PMID: 29155950. doi:10.1093/nar/gkx1098
2. Paten, B et al. “Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs” Genome Research, 18:1814-1828, 2008
3. B. Paten and et al.” Genome-wide nucleotide-level mammalian ancestor reconstruction” Genome Research, 18:1829-1843, 2008
4. M. Kircher e al. “A general framework for estimating the relative pathogenicity of human genetic variants,” *Nature Genetics*, vol. 46, no. 3, pp. 310-317, 2014.
5. C. Gross, D. de Ridder and M.J.T Reinders, “Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse”, *BMC Bioinformatics* 19(1):373, 2018.
6. Colin Groves et al.: Ungulate Taxonomy. Johns Hopkins University Press, 2011, S. 1–317 (S. 108–280)

a. The Delft Bioinformatics Lab

b. Bioinformatics – WUR (Wageningen University & Research)