# The Phylogeny of Staphylinidae (Rove Beetles) from Whole Mitochondrial Genomes

April 5, 2024

Samuel Smith
Imperial College London
scs23@imperial.ac.uk

Superviser:
Alfried Vogler
Imperial College London
a.vogler@imperial.ac.uk

# 1 Keywords

Phylogeny, Rove Beetles, Staphylinidae, Diversification, Phylogenetic Tree, Mito-Genomes, Basal Relationships

# 2 Introduction

Staphylinidae (Rove Beetles) is the most species rich family of the diverse Coleopteran order. Staphylinidae consists of many subfamilies including the highly diverse Staphylininae, Paederinae, and Aleocharinae. Molecular phylogenetic studies of the family Staphylinidae have continued throughout the 21st Century, with the most accurate and successfull analysis' making use of Bayesian and Maximum Likelihood methods for tree building [1]. The development of modern sequencing techniques and an increased abundance of genetic data is a burden on traditional phylogenetic methods lacking computational power to process many and large sequences.

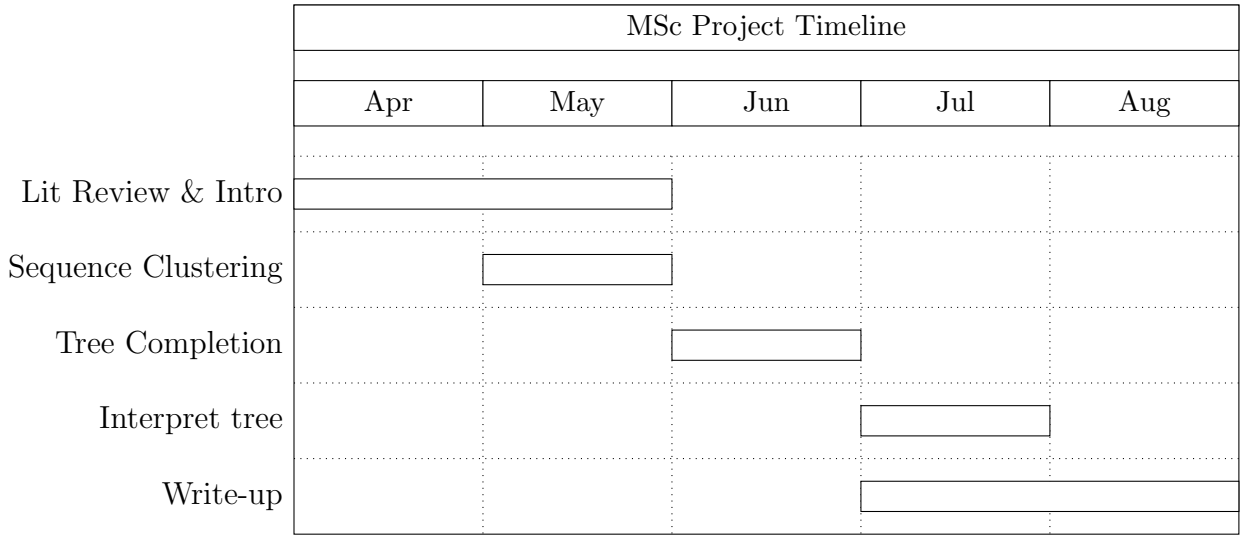
Thus, this project will overcome this challenge through the use of machine learning clustering algorithms to help create a phylogenetic 'backbone' for traditional phylogenetic methods to take place consequently. The outcome of this project will ultimately provide an efficient and automated pipeline that will take sequences as input and output an accurate tree based on the aforementioned methods. This project will also aim to make sense of the diversification of Staphylinidae on a geographical basis. Nuclear or mitogenomes can be used in this project, however, mitogenomes are useful due to their relatively high evolutionary rate and their lack of genetic recombination - thus numerous phylogenetic signals.

# 3 Methods

To construct an accurate phylogenetic tree using the mitogenomes of 1000 Staphylinidae species from around the world. Clustering (machine learning) methods will first be implemented to create a backbone, reference tree which can consequently be refined and added to, using traditional phylogenetic methods. k-means clustering can be applied to sequence similarity metrics (based on pairwise sequence comparisons, i.e. Damerau-Levenshtein Distance) to identify 'clusters' of similar species. Iterative k-means clustering

will be implemented to determine the most suitable clusters according to a silhoette coefficient score. Consequently, the 'centroid' species will be chosen from each cluster alongside the reference species, for traditional phylogenetic tree construction. Once this backbone is made, the remaining species in each cluster can be added to thier respecitve centroid to refine the tree. The tree will then be used to place existing (meta)barcode sequences for a phylogenetic framework and reference system. All methods likely to be carried out using Python and R.

# 4 Timeline




# 5 Budget

**1TB Hard Drive** for **£100** required for storing large amounts of mitogenomic and data. **HPC Computing Time** may be required for tree building or machine learning tasks but **cost unknown**. **Travel** fee (**c£50**) for commuting to the Natural History Museum via Transport for London.

# References

[1] Vladimir I. Gusarov. *Phylogeny of the Family Staphylinidae Based on Molecular Data: A Review*, pages 7–25. Springer International Publishing, Cham, 2018.