

# Proposal

Amyre

2025-03-17

## Final Project Proposal - BIOL 432

### Group 2 Members:

Bonnie Rule (20295738),  
Mackenzie Calhoun (20265644),  
Amyre Wells (20320047),  
Jennah Levac (20169998),  
Taylor Pontet-Dasilva (20310351),  
Nithila Sathieskumar (20279386)

### Git Repository

#### ***Link:***

[https://github.com/amyrewells/432\\_Final](https://github.com/amyrewells/432_Final)

#### ***Username:***

jennah2882  
mackenziecalhoun22  
rulbon24  
amyrewells  
Nithy3640  
Taylorpd4

### Chosen Dataset

#### ***Article Citation:***

Gut microbiota of dung beetles correspond to dietary specializations of adults and larvae. [Shukla SP, Sanders JG, Byrne MJ, Pierce NE. Gut microbiota of dung beetles correspond to dietary specializations of adults and larvae. Mol Ecol. 2016;25(24):6092-6106. doi:10.1111/mec.13901]

#### ***Dataset link:***

<https://datadryad.org/dataset/doi:10.5061/dryad.vq744>

#### ***NCBI ascension numbers:***

(found in “Data Availability” section of original article)

KX459698- KX460822

***Files used:***

Mapping file (mapping.txt): Contains metadata about each sample, including species, diet, and experimental conditions.

OTU table (OTU\_table.xls converted to OTU\_table.csv): Contains operational taxonomic unit (OTU) abundances per sample, based on 16S rRNA sequencing.

OTU sequences (Deposited in NCBI): Can be used for phylogenetic analysis.

## **Biological Questions & Methods**

***Question 1:***

Does microbiome composition differ significantly between dung beetle species (irrespective of life stage or sex), and can these differences be used for species classification?

**Method:**

- Utilize a distance matrix in conjunction with non-metric multidimensional scaling (NMDS) as a means to effectively visualize the clustering of microbiome samples based on species classification.
- Perform a linear model to test composition significance of each NMDS axes
- Random Forest classification techniques should be implemented to identify the most effective operational taxonomic units (OTUs) differentiating among the various species present within the microbiome dataset. Must balance out OTUs among training and test data to ensure more abundant OTUs are not overrepresented.
- Assess model performance via confusion matrix and Cohen's Kappa calculation

***Question 2:***

How do microbial communities differ between dung beetle samples?

**Method:**

- Perform NMDS:
- Relative abundance-based NMDS to examine quantitative differences.
- Identify dominant bacterial families for each species using barplots (extract families using regex from OTU taxonomic lineages).

***Question 3:***

How do microbial communities of each dung beetle species differ in terms of taxonomic diversity and phylogenetic structure?

**Method:**

- Using the normalized data from previous analyses, multiple sequence alignment is meticulously carried out utilizing the MUSCLE algorithm.
- A phylogenetic tree is constructed to elucidate the evolutionary relationships among the sequences.
- The distances between various taxa are visualized using a heatmap, clearly representing interspecies relationships.

**Overall,** we aim to investigate whether certain microbial taxa exhibit species-specific characteristics, even when subjected to identical rearing conditions. This indicates the presence of host-dependent symbionts, which may play a crucial role in the ecological dynamics of these microbial communities.

## Dataset Structure & Pre-processing

### Overview of dataset structure

#### *Number of samples:*

The data contains details of various kinds of dung beetles. It contains a fixed number of samples (for more information, refer to the metadata file). Every sample is from a single dung beetle under various experimental conditions

#### *Metadata variables:*

Species: Indicates what type of dung beetle the sample was taken from

Diet: What beetles eat, which can alter the kind of germs that are present

Life stage: Whether the sample is from an adult or larval beetle

Experimental condition: Various circumstances under which the samples were collected (such as environmental or diet alterations).

Sequencing depth: The quantity of sequencing reads obtained from each sample (crucial for normalization)

#### *OTU table:*

Rows = OTUs, columns = samples, values = abundance counts.

#### *Taxonomic resolution:*

family, genus, species (depends on classification confidence).

### Pre-processing steps

#### *Normalize the data:*

- Sum across the OTU ID column
- Each value in that OTU ID column then gets divided by that sum
- That number must then be multiplied by  $10^3$  (to the thousand) or sum of columns. As long as method is consistent across samples and be rational (annoying if the numbers to be too small)

#### *Data transformation:*

- Log-transform OTU abundance for NMDS.
- Convert OTU presence/absence to a binary matrix for compositional analysis

## Timeline

Date	Tasks	Responsible Members
March 11-16	Finalize dataset choice	All
March 17	Set up Github	Amyre
March 17	Finalize and submit proposal	All
March 18-22	Conduct NMDS & Random Forest Analysis	Amyre and Nithila
March 18-22	Perform normalization & phylogenetic tree construction	Jennah and Taylor
March 23-27	Create visualizations (barplots, heatmap)	Bonnie and Mackenzie
March 23-27	Interpret results & draft discussion, methods summary	All
March 23-27	Methods summary	Nithila
March 28- 30	Finalize report & GitHub documentation	All
March 28- 30	Make final presentation	All

Date	Tasks	Responsible Members
March 31	Submit final project, ready to present!	All