How differences in genome annotation affect downstream analyses

Background:
The software used in the genome annotation pipeline can have a great impact on the quality of the final genome annotation data. There are many different annotation pipelines and within each of these pipelines the programs have multiple parameters which can influence the final genome annotation. These differences introduced in the annotation process may influence downstream analyses like genomic and evolutionary analyses.
The dataset used is a collection of 15 beetle genomes with high assembly quality. This dataset has already been used for some analyses and now we would like to find out, what impact the genome annotation pipeline might have on the analyses. The beetles in this dataset are weevils (Curculionidae), which are plant- and fungus-feeding beetles.
Additionally, there is a smaller dataset of burying beetles (Nicrophorus) available, for which we also have RNA-seq data. This dataset can be used to assess the effect of adding RNA-seq data to the genome annotation.
Both, weevils and burying beetles are very interesting beetle families as different species within these families evolved different social complexity levels, e.g. parental care. The evolution of these social characteristics within the beetle families can be analysed using the annotated genomes of these species.

Overall objective: In this project, different repeat masking pipelines and annotation settings will be compared. Annotation output like gene numbers, agreement between gene sets, domain occurrence will be analysed. The quality of all annotated genomes will be assessed and some basic analyses (GC content, gene/exon/intron length) will be performed.

Potential hypotheses, to be supplemented and further developed within the project

1. Genome annotation pipelines affect downstream analyses.
2. Using more data as input for genome annotation (e.g. including RNA seq data) increases genome annotation quality.

Materials & methods
The student will use two sets of genomes and annotate these with an established pipeline. Parameters of the pipeline will be changed and their effect on the genome annotation as well as downstream analyses will be analysed.

Requirements:

- Some experience with the command line, or interest to learn
- Interest in learning the usage of different programs
- Interest in genome annotation and genomic analyses

Supervision: Sarah Rinke-Stack, Room 17, s.rinke-stack@uni-muenster.de, Molecular Evolution and Bioinformatics Group (http://bornberglab.org/).

References

Biedermann, Nuotclà (2020). Social Beetles.
Gabriel et al. (preprint). BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA
Bell (2021). Transposable element annotation in non-model species - the benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines.