Last Chapter
Possible name: “Deciphering the Non-coding Conundrum: A Multi-Omics Analyse Approach Towards Constructing a Comprehensive Non-coding RNA Database”

We have curated a comprehensive database that houses metagenome, metatranscriptome, and metaproteome datasets. For more information, please visit our recent abstract at https://www.biorxiv.org/content/10.1101/2023.01.03.519475v1.abstract.

- Our Workflow Consists of Several Steps:
  - We start by executing the assembly of Metagenomes utilizing our MuDoGer tool.
  - Next, we aim to identify and extract potential non-coding regions from these assemblies:
    - As an initial approach, we plan to use Infernal, although we are also considering developing a custom pipeline for this purpose doing the follow steps:
      - Use the Prodigal software is used to identify possible protein-coding regions, which are subsequently removed from the genomes.
      - We are thinking of developing an algorithm that will sweep across the remaining genome, isolating potential non-coding regions.
  - We then apply our established pipeline for non-coding RNA classification:
    - The BioAutoML-Fast software is used for feature engeneering of these non-coding regions.
    - We use a Deep Learning model (need to parallelize) that we've submitted to RNA Biology, along with the existing classification method in BioAutoML, for classifying these non-coding regions.
  - We then proceed to map the RNA to DNA sequences (we are currently exploring various methodologies and considering consulting with Stefania for expert advice on this matter).
    - Next, we analyze the metatranscriptome to verify whether the non-coding regions identified do not transcribe (we do this by comparing Reads for the genome or using BLAST for comparison).
  - For metaproteome analysis, we intend to collaborate with Lorenz Adrian from UFZ.
  - Once the pipeline has been evaluated and validated, we plan to apply it on the CLUE-TERRA database, which comprises over 23k MAGs. We also plan to utilize another dataset (Possibly GEM? We'll confirm the name with Ulisses) which will bring our total genomes to around 50k.
  - The culmination of these efforts will result in an extensive database of annotated non-coding RNAs.
  - As a service to the scientific community, we are in the process of developing a user-friendly web application that will host the database, enabling users to perform searches and filter results by organism, among other functionalities.
  - Furthermore, we plan to incorporate a feature into the web application that will allow users to submit a genome for analysis, after which we will return the results of our comprehensive non-coding RNA analysis. This will necessitate the creation of a dedicated WebService.