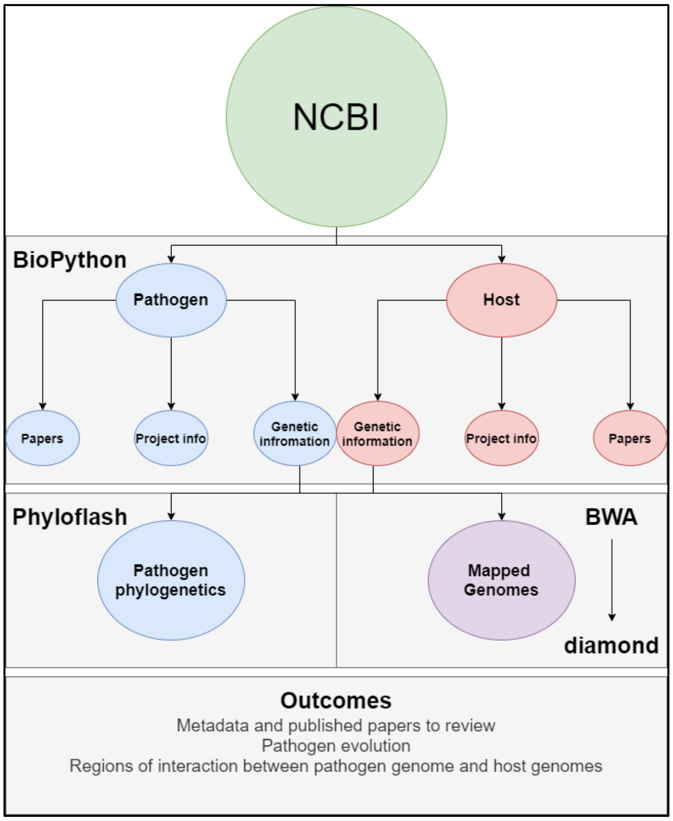
Project Update: Building a NCBI web scraper and deep-learning model for inferring host pathogen genomic interactions

Overall: Not going as planned



Issues: My original goal was to carry out many of the data mining function in the Rentrez package in Rstudio as this is a nice interface. But the big issue I was running into was updating R and having package conflict problems. After beating my head against this problem for multiple days I decided to switch gears. There is a similar program in Python called BioPython that I will be using. I think this is going to provide a better was to utilize arguments, other anaconda packages, and connecting results. From here I designed a small diagram to map out the information that I am going to be accessing, storing, and using in this pipeline. First, I will use BioPython to access the NCBI database by allowing the user to define the scientific name of the pathogen and host that they are interested in. Using these names, the script will pull relevant papers, the bioproject information for every "organism" used, and the accession information for the genetic information. Using phyloflash we will reconstruct the phylogenetic composition of our pathogen dataset. This will show us how related our pathogens are. Using BWA we will map the pathogen and the host genomes, identifying areas that might be shared or have introgressed. These regions will then be annotated with diamond using the ncbi blast database.

To do: There is a lot to do and I am concerned about being able to finish this project in time. I have had significant setback fighting with R but will continue to work with this new framework.