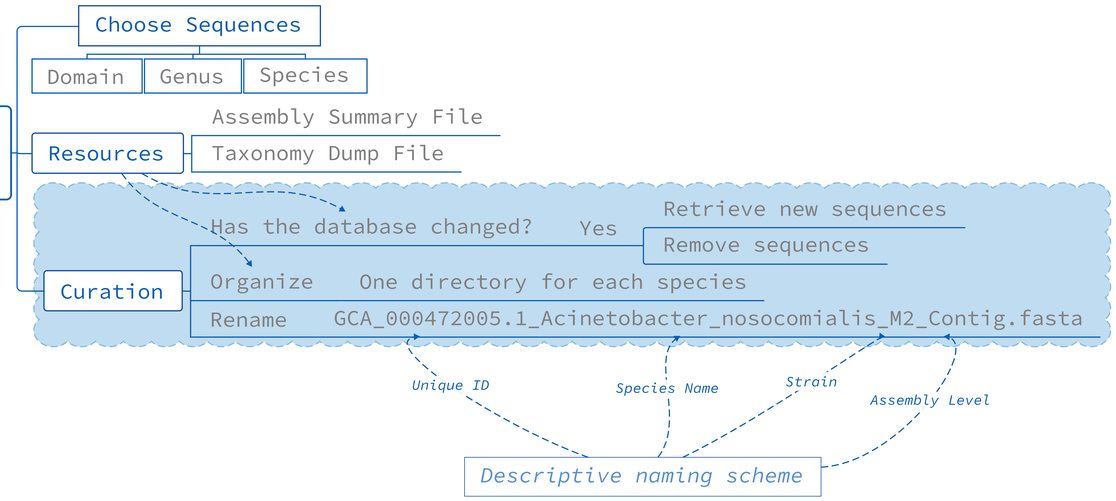
Quality Control and Curation of Genomic Databases

Andrew Sanchez [1]
Jason Sahl [2]

NORTHERN ARIZONA UNIVERSITY
College of Engineering, Forestry, and Natural Sciences

NCBITK Workflow

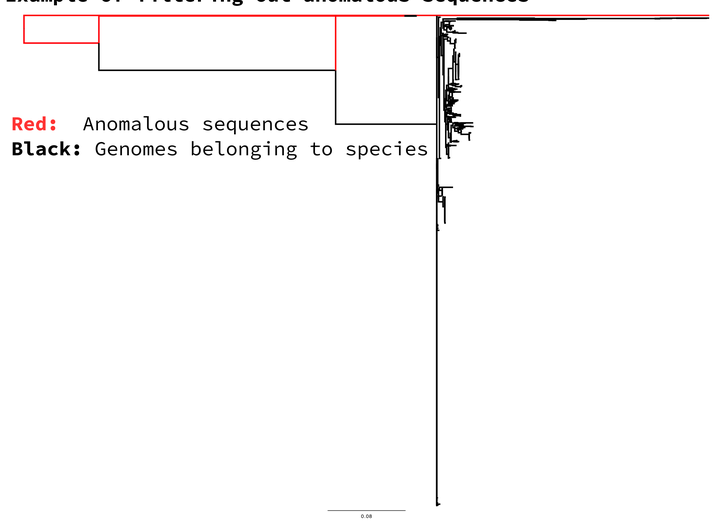


Motivations

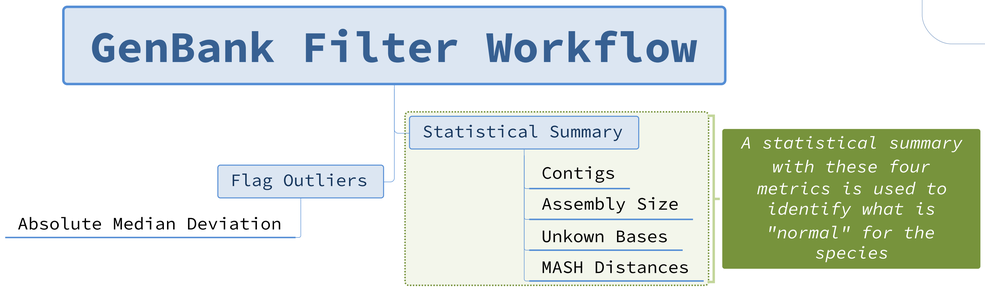
The National Center for Biotechnology Information (NCBI) hosts one of the most comprehensive genetic sequence databases known as GenBank. GenBank is "an annotated collection of all publicly available DNA sequences," and is a massively important resource in genetic research.[3] However, the problems outlined below make it difficult to utilize the data efficiently and comprehensively:

- Downloading sequences
- Staying up-to-date
- Human readable organization
- Quality of sequences

Example of filtering out anomalous sequences



GenBank Filter Workflow



Impacts

NCBITK[4] (NCBI Tool Kit) provides the following solutions:

- Seamless access to genetic sequences:
  - A complete copy of GenBank
  - Just sequences of interest based on taxonomic parameters, i.e. domain, genus, species
- Automatic synchronization ensures that researchers are working with the latest assembly versions.
- Meaningful file names replace the complicated identifiers assigned by NCBI.
- Semi-automated quality control warns researchers of potential bad data.

Conclusion

NCBITK eliminates bottlenecks related to accessing, organizing, and analyzing sequences so that researchers can focus on their work rather than the tedium of maintaining a curated database. Providing a user friendly interface to GenBank which lends itself to extensibility, automation, and integration with existing tools will enable and enhance future research that can benefit from the ability to efficiently utilize the vast amount of data hosted by GenBank.

References

1 Undergraduate Researcher Assistant, Northern Arizona University
2 Principle Investigator, Northern Arizona University
3 https://www.ncbi.nlm.nih.gov/genbank/
4 https://github.com/andrewsanchez/NCBITK

PATHOGEN & MICROBIOME INSTITUTE
NAU & TGen