Data Management Plan

1 Expected Data Type

Primary data from research activities will be digital DNA sequence reads generated from whole genome sequencing. Primary data from educational activities will consist of curriculum and the number of students attending the course, stored digitally.

2 Data Format

The primary data format for research activities will be gzip compressed FASTQ files containing both the sequence and quality scores. Secondary data formats include Makefiles, BASH scripts, Python scripts, and R scripts used for analysis in addition to text files describing the data processing. These will be accompanied by virtual machines (Docker containers) that will preserve identical computing environments for reproducibility.

The primary data format for educational activities will be markdown-formatted text files and rendered HTML files along with all the scripts used for website generation.

3 Data Storage and Preservation

All data will be stored free of charge in the Open Science Framework (`https://osf.io`) and additionally archived on the CERN-funded Zenodo (`https://zenodo.org/`). Both organizations store their data distributed across data centers with daily backups and quality checks to prevent data loss, degradation, or damage. Zenodo uses CERN data centers in Geneva, mirrored in Budapest, with retention guaranteed for the next 20 years. In the event of closure, Zenodo has outlined plans to transfer data to appropriate external repositories. THe Open Science Framework uses Rackspace and Amazon Glacier for data storage and have an established preservation fund of $250,000 to fund data storage for 50+ years in the case of termination of the Open Science Framework. In addition, sequencing reads and assembled genomes will be uploaded to the NCBI Short Read Archive and GenBank. As both Zenodo and the Open Science Framework are free of charge, no funds will be required for data storage and preservation.

4 Data Sharing and Public Access

All data and materials will be hosted or mirrored on the Open Science Framework.

Data from the research component will be available under the Open Data Commons (ODC) Open Database License 1.0 and code will be available under the Massachusetts Institute of Technology (MIT) License.

Data from the teaching component will be released to the Public Domain, hosted on GitHub (https://github.com), and archived on Zenodo.

5 Roles and Responsibilities

Zhian Namir Kamvar will be responsible for all aspects of data management, and Sydney E. Everhart will verify all appropriate actions are taken.