This website is free and open to all users and there is no login requirement. The code for this webserver and all third party software used by this website are available under the open-source Apache 2.0, BSD 3-clause, or similar licenses.

The website is available at `https://lfz.corefacility.ca/superphy/spfy/`. Spfy's code is provided at `https://github.com/superphy/backend` and documentation at `https://superphy.readthedocs.io/en/latest/`.

Our proposal covers an update to Superphy [6], an online predictive genomics platform targeting *E. coli*. The update, called Spfy, adds real-time subtyping options and uses graph datastructures to store and retrieve results for additional analyses. Many of the comparative analyses that are run on current workflows chain different software, but lack storage methods which understand the relations between results. By making the storage and retrieval of results part of the platform, and effectively linked to the organisms of interest with a standardized ontology, we can mitigate the recomputing of analyses. We can also perform comparitive analyses between any data points generated in the pipeline. Integrated data storage will be necessary as whole genome sequencing (WGS) data for bacterial pathogens have accumulated in public databases in the tens of thousands, with hundreds of thousands set to be available within the next few years.

Existing scientific workflow technologies such as Galaxy [2], and pipelines such as the Bacterium Analysis Pipeline (BAP) [4] and the Integrated Rapid Infectious Disease Analysis (IRIDA) platform `http://www.irida.ca/` help automate the use of WGS data for public-health surveillance. Like IRIDA and BAP, Spfy automates workflows for users, and like Galaxy, Spfy uses task queues to distribute selected analysis. File uploads begin through the ReactJS-based website, where user-defined analyses options are selected. To these concepts, we add the use of Docker containerization for task queue workers, thus allowing anaylsis software to safely run in parallel. The main graph database uses annotations from the GenEpiO [3], FALDO [1], and TypOn [5] ontologies which describe biological data. The entire platform is packaged using Docker-Compose, and can be recreated with a simple command.

Spfy was tested with 59,5323 public *E. coli* assembled genomes, 5,353 genomes from GenBank and 54,181 genomes from Enterobase ($\tilde{5}$96 GB), storing both the entire sequences and results for all included analysis modules. The resulting database had XYZ million nodes and XYZ million edges, with XYZ object properties, which worked out to X TB of data stored.