IRIDA: A Genomic Epidemiology Platform Built on top of Galaxy

Aaron Petkau¹, Franklin Bristow¹, Josh Adam¹, Josh Adam¹, Josh Adam¹, Jennifer Cabral^{1,2}, Joel Thiessen^{1,2}, Cameron Sieffert¹, Natalie Knox¹, Jennifer Cabral^{1,2}, Joel Thiessen^{1,2}, Cameron Sieffert¹, Natalie Knox¹, Intersecon Sieffert¹, Natalie Knox¹, Jennifer Cabral^{1,2}, Joel Thiessen^{1,2}, Cameron Sieffert¹, Natalie Knox¹, Intersecon Sieffert¹, Inters Alex Keddy⁷, Robert G. Beiko⁷, William Hsiao^{3,4}, Gary Van Domselaar^{1,2}, Fiona Brinkman⁵

¹National Microbiology Laboratory, Winnipeg, Canada, ³BC Public Health Microbiology and Reference Laboratory, Vancouver, Canada, ³BC Public Health Microbiology and Reference Laboratory, Vancouver, Canada, ³BC Public Health Microbiology and Reference Laboratory, Vancouver, Canada, ⁴University, Intersity of British Columbia, Vancouver, Canada, ⁵Simon Fraser University, Intersity, Int Burnaby, Canada, ⁶Laboratory for Foodborne Zoonoses, Lethbridge, Canada, ⁷Dalhousie University, Halifax, Canada

Abstract

Whole-genome sequencing (WGS) is revolutionizing epidemiological methods for identification and investigation of infectious disease outbreaks. However, the routine use of WGS has been hindered due to the complexity in data management and the lack of pipelines supporting quality control and data analysis standards. While an increasing number of pipelines for genomic epidemiology are being developed, each typically has different installation and execution requirements. This leads to a difficulty in the integration of these pipelines into a single genomic epidemiology system.

Galaxy offers a solution by providing a system to integrate, execute, and maintain data analysis pipelines. In addition, Galaxy provides a community of developers who contribute and maintain the bioinformatics tools used for genomic epidemiology. Our project, IRIDA (Integrated Rapid Infectious Disease Analysis), builds on top of Galaxy a platform for genomic epidemiology. IRIDA provides a system for the storage and management of sequencing data and sample metadata, an interface for the execution of data analysis pipelines, and the storage, auditing and visualization of results. Within IRIDA, we provide standard pipelines for genomic epidemiology including SNVPhyl, our SNV (Single Nucleotide Variant) phylogeny pipeline. These pipelines are executed using a Galaxy instance internal to IRIDA and additional support is provided for exporting genomic sequence data to external Galaxy instances.

By building on top of Galaxy we hope to simplify the process of pipeline integration, to share our pipelines with the bioinformatics community, and to contribute to the development of standards for genomic epidemiology. More information can be found at <u>http://irida.ca</u>.

Data Management A Platform Projects DA Platform Projects - Analysis - Users Sequencing Runs Details Files 6 Project Detail Project Sequencing Run 1 Sequencing Run 1 - Files Details Sequencer Type ncc23 S1 L001 R2 001.fasto Sample Samples Filte Date Created 15 Jun 201 ncc23_S1_L001_R1_001.fastq Items to Show 08-5923 S1 L001 R2 001.fastg Sequence File Sequencing Runs Samples RIDA Platform Projects - Analysis - Users Metadata Projects Sequence Deta Run Total Sequences Total Bases 30325000 Min. Length Other Exam Max. Length GC Content Showing 1 to 3 of 3 entries Sequence Files Data Model Projects

Data in IRIDA is organized by **Projects**. Each project contains samples, each of which contain sequence reads. A **Sample** may also store metadata, such as the collection date, geographic location, and organism. Sequence Files are uploaded via the provided web interface, or directly from a genomic sequencer using an IRIDA upload tool and stored as a **Sequencing Run**. FastQC is used to assess the quality of sequence files. Organising and searching for samples is provided by the IRIDA web interface.



Agence de la santé publique du Canada













III UNIVERSITY of MARYLAND SCHOOL OF MEDICINE INSTITUTE FOR GENOME SCIENCES

LISBOA UNIVERSIDADE DE LISBOA



a place of mind THE UNIVERSITY OF BRITISH COLUMBIA

PHSA Laboratories Public Health Microbiology & Reference Laboratory

