IRIDA: Canada’s Integrated Rapid Infectious Disease Analysis Platform

Introduction

As whole genome sequencing replaces traditional typing for pathogen surveillance and outbreak investigations, there is a continuing need for easy-to-use tools which take advantage of this technology. While the bioinformatics community is efficient at producing analysis pipelines to better leverage sequencing data, organizing and analyzing this data remains difficult for researchers and epidemiologists. To address this issue we present the Integrated Rapid Infectious Disease Analysis (IRIDA) platform, a system that facilitates storage, management, analysis, and secure sharing of sequencing data and associated epidemiological metadata.

Security & Integrity

Data security and integrity is a priority for the IRIDA platform. IRIDA manages data in a project-centric structure. Permissions and settings can be managed on a per-project basis so that data owners can share data with individual users without making it public. All data created or modified in IRIDA is audited with a timestamp, user, and tool which modified the data to ensure traceability, and to help recover accidentally modified data. IRIDA also leverages Galaxy’s data provenance abilities to capture all aspects of how results were generated from an analysis pipeline.

Decentralization and Data Sharing

IRIDA is intended to be installed locally at public health institutions rather than hosted as a centralized service. IRIDA allows institutions to synchronize sequence data and metadata between them to enhance distributed investigations. IRIDA also assists users with uploading data to public archives such as NCBI’s SRA, allowing IRIDA users to contribute to global investigations outside of the IRIDA ecosystem. For any tools or visualization systems that cannot be built into the platform, IRIDA implements a comprehensive REST API that allows external tools to securely access any data stored within the system. IRIDA includes tools to assist with custom analysis by exporting data to Galaxy, or to the command line.

Acknowledgements

IRIDA project funding is graciously provided by Genome Canada, Genome BC, and the Genomics Research and Development Initiative (GRDI) with additional support from Simon Fraser University and Cystic Fibrosis Canada. Compute resources provided by Compute Canada and the Public Health Agency of Canada. We thank additional project advisors for constructive comments.

Sequencer Integration

IRIDA includes a sequence uploader tool that directly interacts with sequencing platforms to quickly import your sequence data into IRIDA. The uploader tool monitors the progress of a sequencing run to transmit the data as soon as the run completes. Once the data is uploaded analysis tools can be automatically launched on the IRIDA server.

Data Management

IRIDA’s data management tools make it easy to collect and keep track of your sequencing data and associated contextual metadata. Quality checks are run an all uploaded data and reported to the user. The web application uses a familiar e-commerce style “cart” interface to manage sample data. The cart interface allows users to easily launch analysis pipelines, create projects, and export data.

Pipeline Management

IRIDA’s analysis engine works closely with Galaxy to launch pipelines, monitor progress, and collect analysis results. Pipelines are built as Galaxy workflows in order to promote easy sharing and integration into the IRIDA system. Galaxy manages distribution of analysis pipelines across available computing resources and sends results to IRIDA to report to the user.

Reports and Visualization

Once results are received from Galaxy, IRIDA handles the reporting aspects required for public health genomics investigations. IRIDA includes a number of visualization tools to present results to the users. Results are fed into IRIDA’s metadata system for bulk investigations. Provenance is saved for all pipelines. All analysis results are exportable for further study in offline analysis tools.