Integration of Galaxy with IRIDA, a Genomic Epidemiology Platform

Aaron Petkau[1], Franklin Bristow[1], Thomas Matthews[1], Josh Adam[1], Damion Dooley[2], Emma Griffiths[4], Geoff Winsor[4], Matthew Laird[4], Melanie Courtot[2,4], William Hsiao[2,3], Gary Van Domselaar[1], Fiona Brinkman[4]

[1]National Microbiology Laboratory, Public Health Agency of Canada, Canada, [2]BC Public Health Microbiology and Reference Laboratory, Canada, [3]University of British Columbia, Canada, [4]Simon Fraser University, Canada

Abstract

The continuing decrease in the cost of genomic sequencing and the development of new data analysis methods has led to the increasing usage of whole genome sequencing as an epidemiological tool. Whole genome sequencing can provide a high-resolution snapshot of the relationship among pathogens and lead to a greater ability to identify and track infectious disease outbreaks. Initiatives, such as Global Microbial Identifier, have already started the discussion on developing a system and standards for genomic epidemiology. In our project, IRIDA (Integrated Rapid Infectious Disease Analysis), we propose a platform for genomic epidemiology which provides secure storage of whole genome sequence data, epidemiological metadata, data analysis pipelines, visualization of results, a REST API, and a federated data sharing model. Galaxy has already proven to be a useful application for integration of common bioinformatics tools and data, execution of data analysis pipelines, collection of results, and data sharing. In addition, Galaxy provides a REST API for programmatic access to running instances of Galaxy. We intend to leverage Galaxy as much as possible by interacting with locally installed Galaxy instances via the API to execute pre-defined data analysis pipelines, store data results and Galaxy histories, and manage installed bioinformatics tools. Direct export of whole genome sequencing data to instances of Galaxy will be provided for more complicated analysis. IRIDA will be released as free and open-source software and make use of common data standards to facilitate sharing with other genomic epidemiology platforms. More information will be made available at http://irida.ca.

Introduction

Modern epidemiology is making a greater use of genomic information as a valuable resource during an investigation. The wealth of information gained through whole genome sequencing of pathogens offers the potential to enhance the information from existing labratory-based techniques. This has been demonstrated during investigations such as the 2008 Listeriosis outbreak in Canada[1], and the 2010 Haiti Cholera[2,3] outbreak. The decreasing cost of whole genome sequencing has lowered the barrier of entry for generating whole genome sequence data, but the complexity of storage, management, analysis, and sharing of this data has limited the use of whole genome sequencing during a real-time outbreak investigation.

During an epidemiological investigation, data is gathered from many different sources and stored in a variety of formats including paper forms, spreadsheets, electronic databases, and files. This data may be dispersed among many institutions, each of which has their own policies for data sharing. Proper interpretation often requires access to this disperse set of data and any of the analysis results being generated.

Analysis of whole genome sequence data, in particular, often requires the usage of complex bioinformatics tools running within a high-performance computing environment. Rapidly changing software and the lack of analysis standards further limit the use of whole genome sequence data to experts who spend a great deal of time on interpretation of the results.

Galaxy[4,5,6] has reduced some of this complexity by providing a storage area for whole genome sequence data and access to the necessary tools to perform analysis on this data. IRIDA is an in-development platform attempting to further address some of these complexities by providing a centralized repository for whole genome sequence and epidemiological data, standardized analysis pipelines implemented using Galaxy, and a method to securely share data among other epidemiological platforms.

Data Analysis

The Data Analysis component of IRIDA will provide a set of standard data analysis pipelines that can be executed on data stored within IRIDA. Galaxy will be used for performing data analysis and managing provenance information for each analysis.

9. Galaxy Tool Shed

Galaxy workflows will be maintained and distributed using a locally installed Galaxy Tool Shed[8].

8. Results Storage

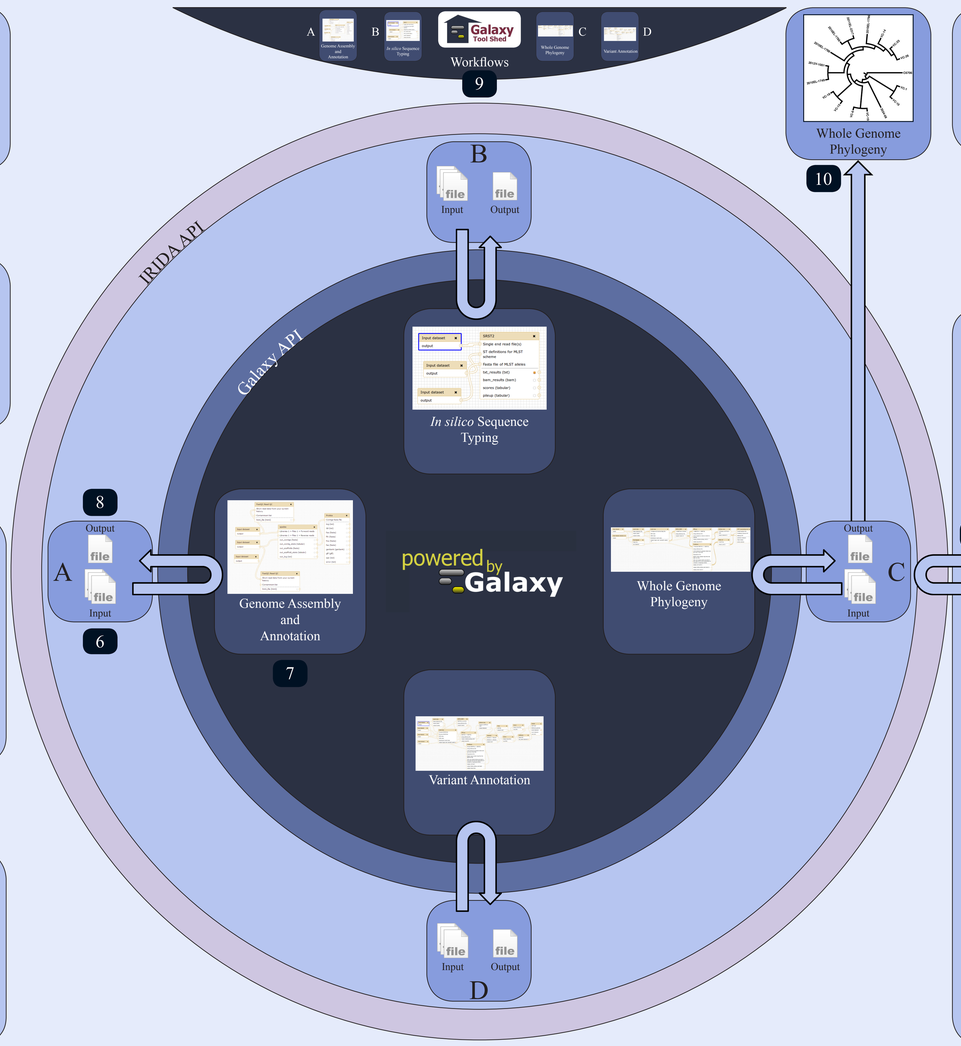
Output files from Galaxy workflows will be uploaded back into IRIDA for storage. Provenance information for each analysis will be kept for later access.

7. Galaxy Workflows

Workflows within Galaxy will be used to implement the different analysis types. Input data will be gathered by IRIDA and sent to an instance of Galaxy using the Galaxy API[7]. A pre-installed workflow within Galaxy will be executed on this data to perform the analysis.

6. Analysis Types

Pre-defined analysis types will be available within IRIDA. Some examples include:
A. Genome Assembly and Annotation
B. *In silico* Sequence Typing
C. Whole Genome Phylogeny
D. Variant Annotation



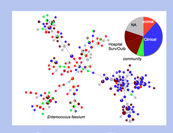
10. Visualization

Visualization of the results and metadata will be provided both within the web interface as well as by external software.

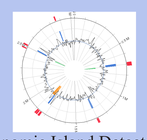
11. External Tools

Results can be further explored by transferring data to external tools. This can be accomplished either by using the IRIDA API or by exporting the data from the IRIDA web interface.

Phylogeography (GenGIS[9])



Sequence Type Visualization (PHYLOViZ[10])



Genomic Island Detection (IslandViewer[11])

Conclusion

The use and availability of whole genome sequence data for epidemiological investigations will continue to increase as genomic sequencing becomes cheaper. This will increase the demand for methods to store, process, and manage this vast amount of information. IRIDA will attempt to meet this demand by providing a central storage area for genomic and epidemiological data, standard data analysis pipelines implemented using Galaxy, and a mechanism for securely sharing data.

We foresee many genomic epidemiology platforms co-existing in the future, with Global Microbial Identifier providing a common ground for the sharing of data and analysis methods. By applying Galaxy and the Galaxy Tool Shed as the mechanism for the definition, execution, and sharing of workflows, we will make use of Galaxy as another common ground among researchers for the development of standards for whole genome sequence analysis. We plan to share the data analysis methods and code we develop and to make use of analysis methods developed by others. This will help benefit the larger scientific community in moving towards standards for the integration of whole genome sequencing data with epidemiological investigations.

References

Web

IRIDA - http://irida.ca
Global Microbial Identifier (GMI) - http://www.globalmicrobialidentifier.org/
GenGIS - http://kiwi.cs.dal.ca/GenGIS/Main_Page
IslandViewer - http://www.pathogenomics.sfu.ca/islandviewer/
PHYLOViZ - http://www.phyloviz.net/wiki/

Publications

1. Gilmour MW, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel KM, et al. **High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak.** BMC Genomics. 2010;11:120.
2. Reimer AR, Van Domselaar G, Stroika S, Walker M, Kent H, Tarr C, Talkington D, Rowe L, Olsen-Rasmussen M, Frace M, Sammons S, Dahourou GA, Boncy J, Smith AM, Mabon P, Petkau A, Graham M, Gilmour MW, Gerner-Smidt P, and Vibrio cholerae Outbreak Genomics Task Force. **Comparative genomics of Vibrio cholerae from Haiti, Asia, and Africa.** Emerg. Infect. Dis. 2011 November; 17(11):2113–2121.
3. Katz LS, Petkau A, Beaulaurier J, Tyler S, Antonova ES, Turnsek MA, Guo Y, Wang S, Paxinos EE, Orata F, Gladney LM, Stroika S, Folster JP, Rowe L, Freeman MM, Knox N, Frace M, Boncy J, Graham M, Hammer BK, Boucher Y, Bashir A, Hanage WP, Van Domselaar G, and Tarr CL. **Evolutionary dynamics of Vibrio cholerae O1 following a single-source introduction to Haiti.** mBio 2013 July; 4(4):e00398-13.
4. Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** Genome Biol. 2010 Aug 25;11(8):R86.
5. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. **"Galaxy: a web-based genome analysis tool for experimentalists".** Current Protocols in Molecular Biology. 2010 Jan; Chapter 19:Unit 19.10.1-21.
6. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. **"Galaxy: a platform for interactive large-scale genome analysis."** Genome Research. 2005 Oct; 15(10):1451-5.
7. Clare Sloggett, Nuwan Goonasekera and Enis Afgan, **"BioBlend: automating pipeline analyses within Galaxy and CloudMan"**, BMC Bioinformatics 2013
8. Daniel Blankenberg, Gregory Von Kuster, Emil Bouvier, Dannon Baker, Enis Afgan, Nicholas Stoler, the Galaxy Team, James Taylor and Anton Nekrutenko, **"Dissemination of scientific software with Galaxy ToolShed,"** in Genome Biology 2014, 15:403, doi:10.1186/gb4161
9. Parks DH, Mankowski T, Zangooei S, Porter MS, Armanini DG, Baird DJ, Langille MGI, Beiko RG. 2013. **GenGIS 2: Geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework.** PLoS One, 8: 7, e69885.
10. Francisco, A.P., C. Vaz, P. T. Monteiro, J. Melo-Cristino, M. Ramirez, and J. A. Carriço. 2012. **PHYLOViZ: Phylogenetic inference and data visualization for sequence based typing methods.** BMC Bioinformatics, 2012 May 8;13(1):87.
11. Langille, M.G.I. and F.S.L. Brinkman (2009). **"IslandViewer: an integrated interface for computational identification and visualization of genomic islands"**, Bioinformatics, Jan. 16. PMID: 19151094

Data Access and Management

The Data Access and Management component of IRIDA will provide mechanisms for the storage and access both of genomic sequence data as well as epidemiological data.

1. Web Interface

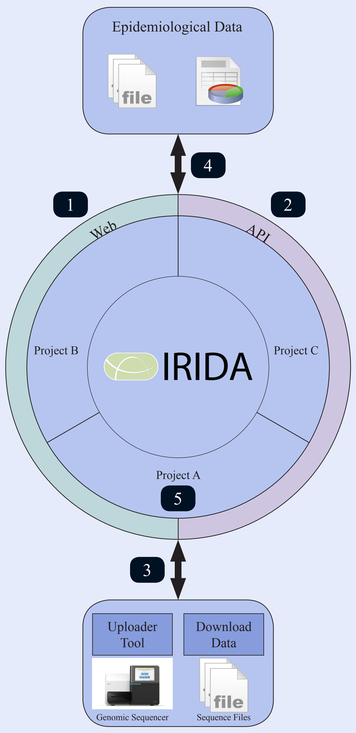
A user's main entry point to IRIDA will be a web interface. This provides a standard look and feel as well as the ability to manage the data uploaded through a variety of mechanisms and the analyses results generated from this data.

2. REST API

An alternative method to access the data within IRIDA is via the REST API. Many tools under development for IRIDA will make use of this API for access to IRIDA.

3. Genomic Sequence Data

Genomic sequence data for a pathogen sample can be uploaded to IRIDA with the help of small uploader tools. These uploader tools, written to target specific sequencing platforms, parse the information generated by a genomic sequencer and upload the sequencing data via the REST API. The sequencing data can later be managed by users via IRIDA's web interface or accessed using the API.



4. Epidemiological Data

IRIDA will provide support for storing and managing metadata about each of the samples being uploaded. This type of data will include the date of collection, geographic information, organism name, and microbial typing information for a sample. Access to this metadata will be provided via the REST API and the web interface.

5. Projects

The genomic sequence data and epidemiological data will be grouped together into projects. Each project will have an associated set of users who have permission to access data within that project. Projects will also contain the results of data analyses run within the project.

Data Sharing and Security

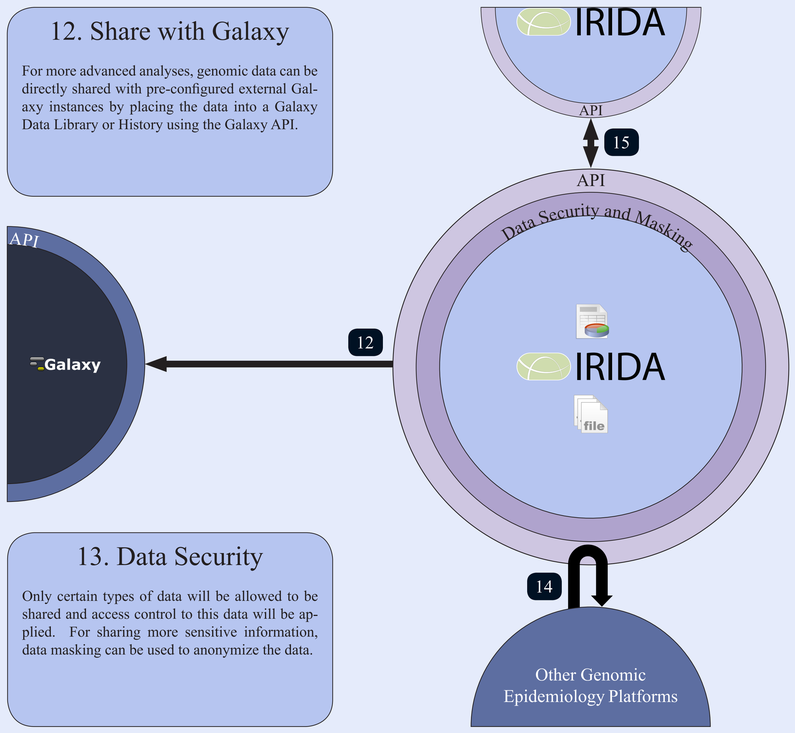
The Data Sharing and Security component of IRIDA provides mechanims for securely sharing data among other data analysis platforms.

12. Share with Galaxy

For more advanced analyses, genomic data can be directly shared with pre-configured external Galaxy instances by placing the data into a Galaxy Data Library or History using the Galaxy API.

13. Data Security

Only certain types of data will be allowed to be shared and access control to this data will be applied. For sharing more sensitive information, data masking can be used to anonymize the data.



14. Genomic Epidemiology Data Sharing

Sharing of data with other remote genomic epidemiology platforms will be supported. This will be facilitated by a common API which will be implemented by IRIDA.

15. Share With Remote IRIDA Instances

Data within an IRIDA project can be linked with remote IRIDA instances that have been pre-configured to allow sharing of data. Access control on the data will be handled through a federated authentication and authorization protocol.

Public Health Agency of Canada | Agence de la santé publique du Canada | GenomeCanada | GenomeBritishColumbia | University of Maryland School of Medicine Institute for Genome Sciences | Lisboa Universidade de Lisboa | Dalhousie University Inspiring Minds | UBC a place of mind The University of British Columbia | PHSA Laboratories Public Health Microbiology & Reference Laboratory | BC Centre for Disease Control | SFU Simon Fraser University Engaging the World