

The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform

Keywords: public health, outbreak investigation, infectious disease surveillance, data integration, whole genome sequencing

Abbreviations: AMR, antimicrobial resistance; API, application programming interface; CARD, Comprehensive Antibiotic Resistance Database; cgMLST, core genome multi-locus sequence typing; CPHLN, Canadian Public Health Laboratory Network; DDBJ, DNA Database of Japan; EMBL-EBI, European Molecular Biology Laboratory – European Bioinformatics Institute; GI, genomic islands; GMI, Global Microbial Identifier; INSDC, International Nucleotide Sequence Data Consortium; IRIDA, Integrated Rapid Infectious Disease Analysis; JSON, JavaScript Object Notation; MLST, multi-locus sequence typing; NCBI, National Center for Biotechnology Information; PFGE, pulsed-field gel electrophoresis; PHAC, Public Health Agency of Canada; PNG, Portable Network Graphics; REST, Representational State Transfer; QC, quality control; SNV, single nucleotide variant; SRA, Sequence Read Archive; SVG, Scalable Vector Graphics; wgMLST, whole genome multi-locus sequence typing; WGS, whole genome sequencing.

Authors: Thomas C Matthews^{1†}, Franklin R Bristow^{2†}, Emma J Griffiths^{3†}, Aaron Petkau¹, Josh Adam¹, Damion Dooley⁴, Peter Kruczkiewicz¹, John Curatcha¹, Jennifer Cabral¹, Dan Fornika⁴, Geoffrey L. Winsor³, Melanie Courtot⁵, Claire Bertelli⁶, Ataollah Roudgar⁷, Pedro Feijao⁸, Philip Mabon¹, Eric Enns¹, Joel Thiessen¹, Alexander Keddy⁹, Judith Isaac-Renton⁴, Jennifer L. Gardy¹⁰, Patrick Tang¹¹, The IRIDA Consortium, João A Carriço¹², Leonid Chindelevitch⁸, Cedric Chauve¹³, Morag R Graham^{1,14}, Andrew G McArthur¹⁵, Eduardo N Taboada¹⁶, Robert G Beiko¹⁷, Fiona SL Brinkman^{3*}, William WL Hsiao^{3,4,18*}, Gary Van Domselaar^{1,14*}

† These authors contributed equally to this work

*corresponding authors: gary.vandomselaar@canada.ca, william.hsiao@bccdc.ca, brinkman@sfu.ca

Affiliations:

¹ National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada

² Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada

³ Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada

⁴ Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada

⁵ EMBL European Bioinformatics Institute, Hinxton, Cambridge, UK

⁶ Institute of Microbiology, Lausanne University Hospital and Lausanne University, Lausanne, Switzerland

⁷ Academic Research Computing, Simon Fraser University, Burnaby, BC, Canada

⁸ School of Computing Science, Simon Fraser University, Burnaby, BC, Canada

⁹ GoSecure, NS, Canada

¹⁰ British Columbia Centre for Disease Control, Vancouver, BC, Canada

¹¹ Department of Pathology, Sidra Medicine, Doha, Qatar Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, Doha, Qatar

- ¹² Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina de Lisboa, Universidade de Lisboa, Lisbon, Portugal
- ¹³ Department of Mathematics, Simon Fraser University, Vancouver, Canada
- ¹⁴ Department of Medical Microbiology and Infectious Diseases, Max Rady College of Medicine, University of Manitoba, Winnipeg, MB, Canada
- ¹⁵ M.G. DeGroote Institute for Infectious Disease Research, Department of Biochemistry and Biomedical Sciences, DeGroote School of Medicine, McMaster University, Hamilton, ON, Canada
- ¹⁶ National Microbiology Laboratory, Public Health Agency of Canada, Lethbridge, AB, Canada
- ¹⁷ Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada
- ¹⁸ British Columbia Centre for Disease Control Public Health Laboratory, Vancouver, BC, Canada

Abstract

Whole genome sequencing (WGS) is a powerful tool for public health infectious disease investigations owing to its higher resolution, greater efficiency, and cost-effectiveness over traditional genotyping methods. Implementation of WGS in routine public health microbiology laboratories is impeded by a lack of user-friendly automated and semi-automated pipelines, restrictive jurisdictional data sharing policies, and the proliferation of non-interoperable analytical and reporting systems. To address these issues, we developed the Integrated Rapid Infectious Disease Analysis (IRIDA) platform (irida.ca), a user-friendly, decentralized, open-source bioinformatics and analytical web platform to support real-time infectious disease outbreak investigations using WGS data. Instances can be independently installed on local high-performance computing infrastructure, enabling private and secure data management and analyses according to organizational policies and governance. IRIDA's data management capabilities enable secure upload, storage and sharing of all WGS data and metadata. The core platform currently includes pipelines for quality control, assembly, annotation, variant detection, phylogenetic analysis, *in silico* serotyping, multi-locus sequence typing, and genome distance calculation. Analysis pipeline results can be visualized within the platform through dynamic line lists and integrated phylogenomic clustering for research and discovery, and for enhancing decision-making support and hypothesis generation in epidemiological investigations. Communication and data exchange between instances are provided through customizable access controls. IRIDA complements centralized systems, empowering local analytics and visualizations for genomics-based microbial pathogen investigations. IRIDA is currently transforming the Canadian public health ecosystem and is freely available at <https://github.com/phac-nml/irida> and www.irida.ca.

Impact Statement

Whole genome sequencing (WGS) is revolutionizing infectious disease analysis and surveillance due to its cost effectiveness, utility, and improved analytical power. To date, no

“one-size-fits-all” genomics platform has been universally adopted, owing to differences in national (and regional) health information systems, data sharing policies, computational infrastructures, lack of interoperability and prohibitive costs. The Integrated Rapid Infectious Disease Analysis (IRIDA) platform is a user-friendly, decentralized, open-source bioinformatics and analytical web platform developed to support real-time infectious disease outbreak investigations using WGS data. IRIDA empowers public health, regulatory and clinical microbiology laboratory personnel to better incorporate WGS technology into routine operations by shielding them from the computational and analytical complexities of big data genomics. IRIDA is now routinely used as part of a validated suite of tools to support outbreak investigations in Canada. While IRIDA was designed to serve the needs of the Canadian public health system, it is generally applicable to any public health and multi-jurisdictional environment. IRIDA enables localized analyses but provides mechanisms and standard outputs to enable data sharing. This approach can help overcome pervasive challenges in real-time global infectious disease surveillance, investigation and control, resulting in faster responses, and ultimately, better public health outcomes.

DATA SUMMARY

1. Data used to generate some of the figures in this manuscript can be found in the NCBI BioProject PRJNA305824.

INTRODUCTION

Infectious diseases continue to exact a substantial toll on health and health-care resources accounting for nearly a quarter of the estimated 52.8 million deaths annually, as well as hundreds of billions of dollars in lost productivity representing significant percentage of global GDP (1,2). Globalization of food networks increases opportunities for the spread of foodborne pathogens beyond borders and jurisdictions. Furthermore, new foodborne pathogens emerge driven by factors such as pathogen evolution or changes in agricultural and food manufacturing practices (3). In response to these challenges, public health microbiology surveillance programs employ molecular methods for routine typing and monitoring of foodborne pathogens (4–7). Current techniques such as Pulsed-Field Gel Electrophoresis (PFGE) and traditional Multi-Locus Sequence Typing (MLST) offer far less discriminatory power for distinguishing outbreak cases from sporadic cases compared with more recently developed Whole Genome Sequence (WGS)-based approaches such as core genome MLST (cgMLST) or Single Nucleotide Variant/Polymorphism (SNV/SNP)-based phylogenies (8). Furthermore, conventional typing methods are resource intensive and can require relatively long turnaround times. In contrast, WGS can inform investigators about numerous infectious pathogen traits in a single “assay”, often with reduced turnaround time and with improved analytical power. As such, public health and food safety authorities are undergoing a historic transition, as WGS-based applications increasingly replace many molecular and phenotypic assays, ranging from serotyping and other molecular level microbial characterization for pathogen identification and surveillance as well as

outbreak response, to antimicrobial resistance (AMR) and virulence prediction for risk assessment (9–17).

While the advantages of implementing WGS-based analytics are widely recognized, efforts to integrate WGS into regional and national foodborne pathogen surveillance programs have been hampered by the lack of easy-to-use and validated bioinformatics tools, particularly impacting implementation in regional public health microbiology laboratories. To date, no “one-size-fits-all” genomics platform has been developed or universally accepted owing to differences in national health systems, data sharing policies, computational infrastructures, prohibitive costs, lack of interoperability, and lack of qualified personnel to carry out analyses. Several international infectious disease monitoring initiatives are working on approaches to tackle these problems. PulseNet (www.pulsenetinternational.org/), a molecular subtyping network for foodborne disease surveillance, is building towards routine implementation of WGS for foodborne surveillance and cluster detection for coordinated outbreak response. PulseNet applies BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium), a widely-used commercial analysis platform that recently has incorporated whole genome sequence assembly and typing functionality into its suite of molecular data analysis, management, and reporting tools. The international GenomeTrakr network is the first distributed network of laboratories applying WGS to coordinate investigations of outbreaks of foodborne illnesses with compliance actions (18). The Global Microbial Identifier (GMI) pathogen tracking initiative collaboratively works with international health, regulatory and research organizations, as well as the International Nucleotide Sequence Database Collaboration (INSDC), to create a worldwide network of shared genomic information for bacterial, viral, and parasitic microorganisms for identification of and response to infectious disease clusters. The National Institutes of Health - National Center for Biotechnology Information (NIH-NCBI) Pathogen Detection pipeline and database offers centralized services for genome assembly, annotation, strain clustering and AMR prediction for foodborne pathogens. Such centralized services for analyses and data management, however, have limitations for public health agencies dealing with sensitive data under jurisdiction of other authorities, which require data and analyses to be performed and maintained locally. Commercial platforms can be extremely costly if multiple licenses are required to outfit multiple laboratories within a broader network. Furthermore, centralized services and proprietary systems have less flexibility than open-source software, resulting in the need for users to create in-house workarounds. As such, a number of in-house tools and platforms, have emerged in parallel, aiming to fill the void and address the distinct needs of different organizations. The Advanced Molecular Detection initiative of the US Centers for Disease Control and Prevention (CDC) supports the building of shared computing and laboratory capacity for genomics, bioinformatics, and next-generation diagnostic testing, and continues to rollout WGS technologies to state public health laboratories. Public Health England has implemented whole genome sequencing as a routine typing tool for public health surveillance of *Salmonella*, and has continued to rollout standardized WGS-based workflows via its Pathogen Genomics Service (15,19). INNUENDO is an integrated genomics-based foodborne pathogens surveillance project developed as a collaborative effort between several European institutions, co-funded by the European Food Safety Authority. The INNUENDO initiative aims to provide a lightweight framework for the integration of bacterial whole genome sequencing in routine surveillance and

epidemiological investigations supporting small European countries with limited resources (sites.google.com/site/theinnuendoproject). Similarly, the European COMPARE research network aims to provide an analytical framework and data exchange platform for sequence-based pathogen data, thereby contributing to containment and mitigation of emerging infectious diseases (compare-europe.eu) (20). The WGSa platform (wgssa.net), created by the Centre for Genomic Pathogen surveillance, provides standardized pipelines for users to analyze and visualize genomic assemblies. Nullarbor (Microbiological Diagnostics Unit Public Health Laboratory, University of Melbourne, Australia; github.com/tseemann/nullarbor) and the Center for Genomic Epidemiology (Danish Technical University, Denmark; genomicepidemiology.org), are microbial sequence analysis platforms containing pipelines for the assembly, annotation and analysis of sequenced isolates, and produce public health microbiology reports supporting research and foodborne outbreak investigations.

Similarly, Canada has developed its own frameworks for building bioinformatics and public health genomics capacity according to its policies and strategic goals. Canada's public health system exists as a decentralized, multi-jurisdictional network of provincial and federal laboratories collectively referred to as the Canadian Public Health Laboratory Network (CPHLN). In 2016, the CPHLN initiated a national public health genomics strategic goal with a special emphasis on accommodating the specific requirements and capabilities of each participating lab while adhering to published best practices. Best practices for implementing WGS in clinical settings for diagnostics of infectious disease include, as essential components for meeting regulatory standards: the provision of documentation for validation and quality assurance, high capacity data storage, version traceability, and data transfer confidentiality (21–23). In addition to clinical best practices, current best practices in bioinformatics for food safety authorities are being created to increase the transparency of food microbiology risk management decision making (24). These practices include the implementation of data standards, maintainability of software, demonstration of data integrity, traceability, and auditability, as well as the development of frameworks for reporting and interpretation. With these best practices in mind, we have developed the Integrated Rapid Infectious Disease Analysis (IRIDA) platform, a secure, decentralized, open-source, freely available, end-to-end public health genomics platform for microbial infectious disease investigations. IRIDA enables organizations to perform localized analyses, while simultaneously supporting data sharing and synchronization across jurisdictions, and sequence deposition to centralized repositories. The IRIDA project was originally developed as a trilateral collaboration between the Public Health Agency of Canada, the British Columbia Centre for Disease Control Public Health Laboratory, and Simon Fraser University, with key contributions from additional academic, federal, and provincial partners, representing national and provincial public health as well as research interests. Thereafter, the project has expanded to include additional collaborations with academic and government partners. Here, we describe the architecture and core functionality of the IRIDA platform, its current usage in Canada, and its emerging uptake by the global community.

Results

IRIDA is a web-based application that can be locally deployed as a stand-alone platform, designed to provide public health, clinical microbiology, and food regulatory authorities with the capability to incorporate next-generation sequencing into their surveillance, diagnostics, reference typing, and research programs. IRIDA's core functionality encompasses four main areas: 1) data management, 2) user management and data sharing, 3) data analysis, and 4) reporting and visualization. We describe each of these main functions here.

Data Management

IRIDA performs all aspects of sequence data and metadata management, including data import, export, storage, organization, tracking, and sharing (Fig. 1).

IRIDA's data structure is rooted by a Project. Projects contain a collection of Samples, which contain a collection of Sequence Data. Projects also contain Analyses generated from the sequence data within that project. Metadata can be associated with Projects, Samples, Sequence Data, and Analyses (Fig. 2). IRIDA's data structure is modeled after, and fully compatible with, the BioProject data structure used by INSDC databases (NCBI, EMBL-EBI and DDBJ). IRIDA's data structure facilitates deposition of reads, assemblies, and annotated contigs to the various public sequence repositories.

IRIDA currently provides a stand-alone sequence data Uploader Tool for Illumina sequencing instruments; uploaders for other sequencing platforms are under development. The sequence data Uploader Tool monitors the sequencer's output folders and will automatically import new sequence data and associated metadata into IRIDA. Users can also import sequence data and metadata into IRIDA directly via the web interface. To preserve data integrity, sequence data is stored in IRIDA's underlying file system with "read-only" access permissions.

IRIDA uses a familiar "online shopping cart" model to assist users in running their analyses: the user selects the data within their projects that they wish to analyze and adds them to the "cart" (Fig. 3). Users then choose their desired analysis pipeline, adjust the pipeline's runtime parameters (or select a pre-saved set of runtime parameters), and launch their analysis (Fig. 4). IRIDA manages the execution of their analysis and provides indicators for quality control (QC) and execution status.

One of IRIDA's principal functions is to manage the processing of sequence data and metadata through its collection of analysis pipelines. IRIDA incorporates an internal instance of the Galaxy bioinformatics workflow manager (25) to assist in executing the analysis pipelines, which are implemented in IRIDA as Galaxy workflows (galaxyproject.org). Galaxy manages the distribution of analysis jobs on high-performance computing systems. IRIDA provides a computational superstructure on top of its internal Galaxy system to shield users from the complexity of running the analysis pipelines and to provide additional functionality not available

in the Galaxy system such as automated processing of newly imported data, creation and selection of saved pipeline parameters, management of analysis results within a project, and incorporation of analysis outputs into the metadata tables contained within a project.

IRIDA provides several ways to export sequence data, metadata, and analysis results. Users can download their sequence data on to their local workstations through a web-browser. IRIDA also can export data to a locally installed “companion” installation of Galaxy. This feature is useful for running analytical pipelines that are not directly integrated within the IRIDA system. Advanced users with Linux workstations and having access to the IRIDA file system can export their sequence data to their user accounts using a command-line linking utility. Rather than copy the sequence data, both the Galaxy exporter and the command-line linking utility create symbolic links to the original sequence files avoiding unnecessary data duplication and ensuring that the integrity of the original data is preserved.

User Management and Data Sharing

IRIDA is developed as a multi-user application, providing industry standard user authentication and user access controls. Authorization to access data and operations are controlled throughout the application including web interface and REST API. A number of roles are in place to control what a user can and cannot do. The roles fall into two categories: User and Project. Within the User category there are four roles. The first role - User - is intended for regular users; this role grants authority to create and manage projects. The Sequencer role provides the authority to upload data to any project. This role has no login privileges and cannot perform any operations nor access any data on the system outside of data upload. Manager roles have the same privileges as Users, and have additional authority to invite other Users to the system and to modify User details. The Admin role has full system privileges. Admins can access all projects and see all analyses. Roles in the Project category include Project Manager and Project Collaborator. Project Managers “own” the project data. They can import new sequence data and metadata, and can modify existing metadata. Project Managers also can update project settings, invite other users to the project, and share data to another project. Project Collaborators have authority to view sequencing data and metadata for specified projects, and can share data to other projects with limited privileges. They also can run analysis pipelines on data contained within a project. Project Collaborators are not authorized to modify any existing data within a project.

IRIDA uses a project-centric data sharing model and features the ability to share data with other users within the same installation or share data across different installations. Changing of permissions is permitted transiently or permanently by IRIDA Administrators, who can manage their User Groups in such a way as to alter single profiles or assign permissions in bulk to any number of users. Project Managers can share their project data with other users on the same system by simply adding them to the project. Sharing data across IRIDA instances is accomplished through “synchronized projects” and uses a host-client model. The host IRIDA instance provides connection privileges to the client instance. A user on the client instance connects to the host and associates a project on the host instance with a local project; the client

instance then pulls the sequence data and metadata from the host project into a local project. As the host project acquires new data, or modifies existing data, the client will automatically synchronize the data. Each IRIDA instance can synchronize any number of projects with any number of remote IRIDA instances, thus enabling users to create *ad hoc* data sharing networks. Project synchronization can be revoked at any time by the client or the host, although the data shared among IRIDA instances cannot be revoked once it has synchronized across instances.

IRIDA is also capable of submitting sequence data and metadata to NCBI's sequence read archive (SRA), which is synchronized daily with EBI-EMBL's ENA and Japan's DDBJ. This functionality allows sequences to then be easily shared with global, centralized resources such as GenomeTrakr. Users select the datasets they wish to make public, the associated metadata they wish to publish, and the NCBI BioProject ID they have established prior to data deposition. IRIDA will then perform the data submission automatically. Once the data is received by NCBI, IRIDA retrieves and saves the accession numbers.

Sequence Analysis

IRIDA provides the ability to analyze genomic sequence data and metadata using its collection of analysis pipelines. Newly imported sequence data can trigger the execution of certain analysis pipelines, such as genome assembly or *in silico* serotyping. Combined with the automated sequence Uploader Tool, this gives IRIDA the ability to automatically generate analysis results without any user intervention. Analysis results can be incorporated back into the project's metadata and can be viewed in IRIDA as a line list. All results can be exported for use outside of IRIDA.

The sequence analysis pipelines and utilities currently included in IRIDA have been developed to serve the requirements of the Canadian public health system. We summarize them briefly in the following sections.

Quality Control

IRIDA automatically runs FastQC (26) to perform a quality check on newly imported sequence reads. Results of FastQC (quality scores, read duplication, overrepresented sequences, etc.) are captured by IRIDA and presented as charts and tables. Additional quality checks, such as sequence coverage, are calculated for samples within projects. Users can set their own quality score cutoffs for each project. IRIDA flags files that fail quality control before they are included in an analysis, and flags analyses that fail to run for any reason. In-depth information for QC and analysis failures is provided as "drill down" pages.

Genome Distance Calculation

IRIDA provides fast genome distance calculation through the RefSeqMasher pipeline (github.com/phac-nml/refseq_masher), which uses Mash (Mash v2.0+) (27) and a k-mer based Mash sketch database of 54,925 NCBI RefSeq genomes to enable users to quickly match the submitted isolate reads to the closest matching reference genomes. The RefSeq Masher

pipeline can be used to select appropriate reference genomes for use with SNVPhyl or to identify possible contamination. In IRIDA, RefSeqMasher results are presented in tabular form, sorted by the closest matching genome.

Single Nucleotide Variant Detection and Phylogenomic Inference

IRIDA uses SNVPhyl to detect single nucleotide variants (SNVs) contained in genomic sequence data, and applies these SNVs to infer a phylogenetic tree (28). Sequence reads generated from a collection of isolates are mapped against one user-selectable reference sequence using SMALT (<http://www.sanger.ac.uk/resources/software/SMALT/>), and high quality SNVs are detected from the resulting pileups using FreeBayes (29) and SAMtools/BCFtools (<https://github.com/samtools/bcftools>). The concatenated hqSNVs are organized into a multiple alignment, and a maximum likelihood tree is estimated from the alignment using PhyML (30). SNVPhyl provides a table of all identified SNVs and a SNV distance matrix. SNVPhyl can mask internally repeated sequences as well as regions of higher SNV density, indicating possible recombination. Additionally, users may mask user-selectable regions of the reference sequence, such as mobile elements.

Assembly and Annotation

IRIDA provides an Assembly and Annotation pipeline, which enables users to assemble and annotate a single genome or a collection of genomes in batch. Paired-end reads are first merged using FLASH (31), and then passed to SPAdes (32) to perform a *de novo* assembly. Contigs returned by SPAdes are filtered to remove small and low-coverage contigs; quality-filtered contigs are then annotated by Prokka (33). Output files for each sample include assembly statistics, a list of contigs, and Prokka-generated annotations. Batch assembly and annotation enables the user to download a single output data package for all submitted samples.

Serotype Prediction

IRIDA currently performs *Salmonella* serotype prediction with the *Salmonella In Silico* Typing Resource (SISTR), a validated bioinformatics platform for rapid *in silico* inference from draft *Salmonella* genome assemblies (34). SISTR performs highly-accurate serovar prediction based on genosero typing through sequence analysis of the *Salmonella* O and H antigens, with additional refinement of predictions based on population structure context via cgMLST analysis and genomic distance calculation using MASH (9,34,35). Results generated by SISTR are then incorporated into the Sample metadata and can be conveniently viewed in a single table.

Multi-Locus Sequence Typing

IRIDA facilitates cgMLST analysis through the integration of MentaLiST, a fast k-mer based MLST and cgMLST calculation engine enabling genotyping of bacterial samples directly from read data (36). MentaLiST's ability to call alleles directly from raw sequence reads bypasses time-consuming assembly, and is specifically designed and implemented to handle

large typing schemes (*i.e.* thousands of loci) while requiring minimal computational resources. A distance matrix can be constructed based on the Hamming distances between allelic profiles for each sample. The Bio.Phylo module from the Biopython library is used to calculate a neighbour-joining tree in Newick format (37). Within IRIDA, the Newick tree can be displayed using IRIDA's advanced visualization system, and allelic profiles are provided in downloadable tabular form.

Visualization and Reporting

IRIDA provides support for visualizing metadata and analysis results. Trees built by SNVPhyl or MentaLiST are viewable using a modified version of PhyloCanvas (phylocanvas.org) supporting real-time visualization of trees with thousands of taxa directly in the web browser (Fig. 5). Trees can be displayed in a variety of familiar layouts such as rectangular, circular, and radial views. Users can customize various tree attributes (e.g. colors, labels) and export trees to bitmap (PNG) or vector (SVG) formats.

Sample metadata contained within a Project can be viewed in an interactive tabular format, modeled on the “line list” format commonly used to coordinate outbreak investigations (Fig. 6). Columns corresponding to metadata categories can be rearranged or hidden to best present the relevant data for a given investigation. Users can save their customized table formats as templates, which can be applied to any number of projects. Metadata tables can be exported as a spreadsheet in Microsoft® Excel or CSV format. IRIDA's Advanced Phylogenomic Visualization tool integrates phylogenetic trees displayed by PhyloCanvas with epidemiological metadata. The integration of genomic relatedness information with epidemiological data can assist in outbreak investigations and enhances decision making by visualizing the covariation of genomic data and associated epidemiological, clinical, and biological trait data (Fig. 7).

IRIDA generates automated quality reports for all WGS data uploaded into the system using FastQC. Users may examine quality reports on a sample-by-sample basis, or they may examine reports in bulk across many samples - primarily used for sequencing coverage, with low-coverage samples flagged by IRIDA. Analysis pipelines produce differing reports - such as genome length and contig N50 for the assembly pipeline, or MLST results for MentaLiST - but all pipelines include a report on the input data files, tools, and parameters used to generate the results. Organism-specific reporting may be enabled on a project-by-project basis. Currently, this is only supported for *Salmonella* serotyping, generated using SISTR and conforming to the existing report format used within the Canadian Public Health Laboratory Network (CPHLN).

As most public health, clinical microbiology, and food safety laboratories have specific requirements for reporting the results of their analytical tests, IRIDA provides a REST API for external report-generating applications. One application to make use of the REST API is the IRIDA SISTR Results software (github.com/phac-nml/irida-sistr-results), which exports *Salmonella* serotyping information, in bulk to an Excel spreadsheet. IRIDA's web interface also provides the ability to export metadata, genome sequence data, visualizations, and analysis results in standard file formats to facilitate integration into custom reporting systems that cannot directly interact with IRIDA via its REST API.

Linkage to External Tools

IRIDA provides a REST-API which enables authorized external programs to interact with IRIDA and enhance its analytical capabilities. One such external program is GenGIS, a free and open-source bioinformatics application that allows users to combine digital map data with information about biological sequences collected from the environment from multiple sample sites (38). GenGIS provides a connector to IRIDA to download the results of phylogenetic analyses and geographic information stored within IRIDA, integrating this information into a phylogeographic map. Another external program making use of the IRIDA REST API is the BioNumerics software suite, facilitating PulseNet Canada's surveillance activities.

An instance of IRIDA maintained by Simon Fraser University (Burnaby, Canada; sfu.irida.ca) also features integrated API access to IslandViewer (pathogenomics.sfu.ca/islandviewer/), a genomic island (GI) prediction and visualization tool (39). IslandViewer has been used for detection of GIs in *Listeria monocytogenes* isolates obtained from food, food processing environments and listeriosis patients in Canada and Switzerland, as well as the detection of *Streptococcus pneumoniae* isolates obtained from an outbreak in British Columbia (40,41). Results from IslandViewer analyses are returned to IRIDA in tab-delimited or GenBank formats, and include links to interactive circular and linear layouts of the results on the IslandViewer website.

Methods

Architecture

Development Technologies

IRIDA is developed as a Java servlet-based web application built using the Java Spring Framework. The web interface is built using AngularJS (angularjs.org), Bootstrap (getbootstrap.com), Webpack (webpack.js.org) and Thymeleaf (thymeleaf.org). IRIDA's core database is versioned with Liquibase (liquibase.org) to ensure updates to the database are applied consistently during an upgrade. IRIDA uses the public GitHub software code repository for hosting, versioning, bug tracking, feature requests, task management, and software documentation (github.com/phac-nml/irida).

Analysis Pipelines

IRIDA uses the Galaxy workflow management system to carry out the execution of IRIDA's analytical pipelines on high-performance computing systems (25). IRIDA prepares sequencing data, analysis parameters, and other workflow information, and provides these to Galaxy, which manages the execution of the pipeline – implemented as a Galaxy workflow. If installed on a computer cluster, Galaxy also will manage the distribution of analysis jobs across

the cluster's available resources. IRIDA communicates with Galaxy via its API (galaxyproject.org/develop/api/) and monitors the progress of the analysis as it proceeds through the pipeline. Upon completion, IRIDA retrieves the analysis results along with details about the pipeline such as component software versions and database versions.

IRIDA is designed to simplify the integration of additional pipelines by external developers. External developers can include pipelines not provided by IRIDA by building the pipeline as a Galaxy workflow, then integrating the pipeline into IRIDA via an internal API designed to expose the pipeline's analysis parameters. IRIDA automatically generates web forms to present the analysis parameters to the user. The decoupling of the analysis pipelines from the IRIDA base system allows for easy modification using the Galaxy workflow editor. Thus, IRIDA's analysis capabilities can be tailored to the needs of individual labs.

IRIDA provides a REST API to facilitate interoperability by external programs, including scripts, desktop applications, and web applications. The API supports authentication via OAuth2.0. Once authenticated, external applications can download sequence data, metadata, and analysis results for the projects in which they have been granted access. Existing Project, Sample, and File metadata can be modified and new metadata added, depending on the access privileges granted by IRIDA to the external application.

Best Practices

IRIDA aims to comply with best practices for public health labs, clinical microbiology labs, and food regulatory authorities (21,24). Areas of particular focus include data security, data integrity, reproducibility, transparency, validation of software performance, and validation of analysis results.

Security

IRIDA has been designed with data security as a priority. IRIDA uses the Spring Security framework to control access to all data in IRIDA. Spring Security ensures that the roles assigned to users of the system have the correct authorization level to view or modify any data within IRIDA. Security controls are applied across all access points including the web interface and the REST API. The REST API uses the industry-standard OAuth2.0 protocol to delegate access to external applications. Access via the REST API can be granted or revoked on an application-by-application basis by IRIDA administrators. IRIDA enforces strong password policies including a configurable password expiry period.

Software Testing and Development

The IRIDA core development team follows a strict software development process to ensure the application performs as expected. All code written for IRIDA undergoes intensive review by the development team to ensure proper documentation and testing guidelines are followed. A comprehensive testing protocol, including unit tests and integration tests, is carried out to verify the integrity of the code base. User acceptance tests are conducted to verify the

application is fit for use within the Canadian Public Health Laboratory Network. Before deployment, IRIDA must undergo testing in multiple environments, including a development environment, an integration testing environment, and a mock production environment.

Data Integrity

IRIDA implements the Java Hibernate Envers (hibernate.org/orm/envers) module for data auditability. All data created, modified, or deleted in IRIDA is audited with the timestamp, user, and tool which was used to modify the data. Reports can be generated for tracing the history of samples and sequencing information. Deleted or modified data can be recovered and restored to prior versions. All sequence data imported into IRIDA is stored in a read-only format in its underlying file system. IRIDA additionally supports reproducibility by capturing the provenance for all analyses performed by its integrated pipelines. Component software versions, parameters, database versions, and ancillary workflow information are recorded and provided as a viewable report within the web application, or downloadable as a JSON-formatted file. All versions of the pipelines are stored within IRIDA to allow users to re-run and reproduce their results if desired.

Availability

IRIDA is open source software released under the Apache 2.0 license. Participation in developing IRIDA's codebase by the research community is encouraged. IRIDA can be installed on workstations or high-performance computing clusters. IRIDA requires a Linux, Java, Tomcat, and MySQL/MariaDB environment for installation. Analyses require a configured instance of Galaxy. IRIDA is intended to be installed on a network-accessible filesystem for operation on a compute cluster. Alternatively, a pre-built VirtualBox virtual machine is also available for evaluation purposes.

Training and Support

All IRIDA software is supported by documentation and user guides available online (irida.corefacility.ca/documentation/). Technical support for installing and operating IRIDA and its associated pipelines comes in two forms. For agencies operating within Canada's Public Health Laboratory Network, the Public Health Agency of Canada (PHAC) offers direct support through the IRIDA development team operating within the National Microbiology Laboratory in Winnipeg, Manitoba. For support outside the Canadian public health system, IRIDA requests are serviced through the open-source community support model (available at github.com/phac-nml/irida/issues).

DISCUSSION

The bioinformatics skills and computational sophistication required to perform data analysis of large genomics data sets are currently beyond the reach of many public health, food

safety, and clinical microbiology laboratories. Critical gaps still exist to integrate WGS into laboratory workflows. These gaps include the lack of laboratory capacity in decentralized healthcare systems to perform localized genomic data processing and analyses, and the lack of easy and automated methods to share data across jurisdictions without uploading to a centralized repository. The IRIDA platform reduces the expertise required for bioinformatics analysis, and provides technical solutions for multi-jurisdictional data sharing in accordance with jurisdictional data governance policies. Moreover, as an open-source platform, additional and customized functionalities can be added by third parties to address individual laboratory needs. The design principles of the IRIDA platform were developed to specifically address these critical gaps, as well as to better integrate bioinformatics into public health environments, to strengthen clinical and public health interfaces, and to facilitate local-to-global information exchange. A number of other bioinformatics tools and platforms have been developed to better enable laboratories to utilize genomics data, each of which may have different associated costs and best practices (Table 1). Although IRIDA is primarily designed to serve the needs of Canada's multi-jurisdictional public health system, it can be used by anyone wishing to analyze microbial genomic sequence data, and instances have been installed in countries spanning four continents.

IRIDA is an analysis environment for large-scale microbial genomics

IRIDA's web interface guides users through microbial genomics data analysis, from creating projects, uploading WGS samples, and analyzing the samples, to visualizing the analysis results and sharing these results with other users. IRIDA uses a "shopping cart" model to assist users in selecting data for analysis pipelines. IRIDA presents the progress of each pipeline and completed results in the web interface; results can also be exported for use in third-party applications. IRIDA records detailed information for each analysis pipeline and generates a provenance report for auditing purposes.

IRIDA performs analyses of sequence data and metadata with validated pipelines and integrated visualizations

Currently, IRIDA provides validated, user-friendly, fit-for-purpose, pipelines for analyses and integration of WGS and metadata. The SNVPhyl pipeline has been validated for analysis of foodborne disease outbreaks, and correctly distinguishes outbreak-related isolates from non-outbreak isolates across a range of parameters, sequencing data qualities, and in the presence of contaminating sequence data (28,42–44). Since 2010, SNVPhyl has been used for analysis by hundreds of public health analysts at Canada's National Microbiology Laboratory to support research and provincial laboratory services. The pipeline is currently being used for outbreak investigations and has been validated as part of a suite of tools used by PulseNet Canada for all genomic epidemiological investigations of foodborne outbreaks since 2012.

In addition to SNV-based methods, IRIDA provides functionality for performing MLST. Current MLST approaches that extend classical MLST schemes to include an organism's core genome (cgMLST) or whole genome (wgMLST) have been validated for surveillance of foodborne disease, and are quickly being adopted by surveillance programs such as PulseNet

International, and used for a variety of nosocomial, zoonotic, and tuberculosis outbreaks (45–49). IRIDA has adopted the fast, accurate, and computationally efficient MentaLiST system as its main pipeline for cgMLST-based analysis. The MentaLiST pipeline generates MLST allelic calls directly from read data, avoiding slow and computationally expensive genome assembly. MentaLiST's allele calling concordance with other popular genome-scale MLST programs is greater than 99% (36).

The computational resources required to characterize and compare genomes can be immense, with billions of bases generated per sequencing run and petabytes of publicly available sequence data (50,51). Modern data reduction techniques are extremely valuable. Mash is one such general-purpose toolkit that applies the MinHash technique to efficiently reduce large sequence read sets to compressed sketch representations (27). IRIDA offers a Mash-based pipeline called RefSeq Masher to quickly identify the closest genome matches, select appropriate reference genomes for SNVPhyl, as well as identify possible contamination.

In silico serotyping of *Salmonella* is an alternative approach to classical serotyping that identifies *Salmonella* serovars directly from genomic sequence data, which has the benefit of being cheaper, higher throughput, and automatable. IRIDA includes in its pipeline collection the *Salmonella In Silico* Typing Resource (SISTR), a bioinformatics application for rapidly performing simultaneous *in silico* analyses for several leading subtyping methods on draft *Salmonella* genome assemblies. SISTR has been extensively validated, yielding accuracies of ~95%, the highest among serotype prediction tools (9). SISTR is currently being used to generate serotype predictions for all genomes submitted to Enterobase, the largest global repository of *Salmonella* WGS data (enterobase.warwick.ac.uk/). SISTR has generated predictions for the 30% of *Salmonella* genomes that have been deposited at NCBI with missing serovar information (9). Implementation of SISTR also has led to the phasing out of antigen-based serotyping at Canada's National Microbiology Laboratory, which, as of May 2018, has moved to employ WGS as the primary means for characterizing *Salmonella* isolates from national surveillance programs (9).

IRIDA generates dynamic line lists based on uploaded or manually entered sample metadata and other contextual information. The line list is the primary tool used by epidemiologists to collect and organize preliminary information on cases under investigation. Epidemiologists can analyze and integrate contextual information with genomics data in the same platform—critical for the rapid identification of outbreaks, when the alternative is the difficult and time-consuming process of moving large WGS files between analytical platforms. This dynamic feature enhances decision making and can be used to identify isolate clusters fitting the criteria for triggering outbreak investigations, establishing case definitions, and developing response protocols.

IRIDA enables localized analyses, as well as data sharing

IRIDA instances behave as independent data management and analysis environments, with fine-grained data-access security controls. Through IRIDA's decentralized framework, data sharing across IRIDA instances can be fine-tuned through User and Project permissions; these

can be customized according to data governance guidelines, bypassing the need to transmit sensitive information by traditional means such as fax, phone, mail or email. Furthermore, sequence data, metadata, and analysis results can be exported for offline use by authorized users. IRIDA also provides native support for depositing sequence data and metadata to the public archives, supporting global data sharing and complementing centralized resources.

IRIDA is contributing to the creation and adoption of community data standards

The FAIR (Findability, Accessibility, Interoperability, and Reusability) Guiding Principles prescribe best practices for “metadata” (or contextual data) knowledge integration and reuse by the community (52). To make contextual information capture for infectious disease and food safety investigations more FAIR, the IRIDA team has created two new open-source and publicly available ontologies, the Genomic Epidemiology Ontology (GenEpiO) and the Food Ontology (FoodOn), in collaboration with two new international consortia (www.genepio.org; www.foodon.org). Ontologies are hierarchies of well-defined and standardized vocabulary interconnected by logical relationships (53). These logical interconnections provide a layer of intelligence to query engines, making ontologies much more powerful than simple flat lists of terms (54). The IRIDA team continues to work toward the integration of ontology-derived metadata specifications and ontology terms within the platform. Uptake of interoperable ontologies between different public health genomics platforms and tools will better enable the communication of results and information, which will be a crucial step towards realizing a truly responsive global infectious disease surveillance system.

IRIDA Deployment and Uptake

IRIDA is the official bioinformatics platform for public health genomics within the Public Health Agency of Canada. It has been in use since 2016 by Canada’s provincial and national public health laboratories for genomic investigations of foodborne disease outbreaks, as part of PulseNet Canada’s foodborne disease surveillance activities. Instances of IRIDA have been installed across the globe (e.g. the United States, Switzerland, Singapore, South Africa, Italy), including a demonstration version currently available in the Genomics Virtual Lab toolkit (55). Potential users wishing to test the functionality of the IRIDA platform can do so in a number of different ways: as a public demo instance maintained for demonstration (hosted by Simon Fraser University; sfu.irida.ca), private instance installation (requires a working and configured install of Galaxy), or install of an IRIDA virtual machine (downloadable, fully-configured virtual appliance).

Future Development

Future IRIDA developments include ontology integration, enhancements of privacy controls and data exchange mechanisms, the creation of tools for modeling gene transmission dynamics as well as expert systems for genome assembly and annotation. Future pipeline development and integration will aim to decouple individual pipelines from the platform, facilitating a plugin-style architecture to allow developers to individually package their pipelines. Ongoing efforts include the integration of IslandViewer as a core module to enable viewing of genome, annotation and genomic islands in a visually dynamic way, from the whole genome

down to individual genes and enabling side-by-side comparison with other publicly available genomes (39). The IRIDA consortium of developers will also work towards the creation of tools to enable the export of data to other external web applications through provided APIs, for example, the Comprehensive Antimicrobial Resistance Database (CARD) and its Resistance Gene Identifier (RGI) providing antimicrobial resistance and virulence determinant prediction (56). We are also working to enhance IRIDA's deployability, maintainability, portability, and support for cloud computing.

CONCLUSION

The IRIDA platform (irida.ca) is a user-friendly, distributed, open-source bioinformatics and analytical web platform developed to support real-time infectious disease outbreak investigations using whole genome sequencing data. IRIDA strives to shield public health, regulatory and clinical microbiology laboratory personnel from the computational challenges of genomics big data by providing solutions for data storage, management, analyses, and data sharing, to better incorporate WGS technology into routine operations. The functionality and control over data provided by IRIDA can help to overcome the many pervasive challenges for real-time global infectious disease surveillance, investigation and control, resulting in faster responses, and ultimately, better clinical outcomes and improved public health.

REFERENCES

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2012 Dec 15;380(9859):2095–128.
2. Fonkwo PN. Pricing infectious disease. *EMBO Rep*. 2008 Jul;9(Suppl 1):S13–7.
3. Meng J, Doyle MP. Emerging and evolving microbial foodborne pathogens. *Bull Inst Pasteur*. 1998 Jul 1;96(3):151–63.
4. Oniciuc EA, Likotrafiti E, Alvarez-Molina A, Prieto M, Santos JA, Alvarez-Ordóñez A. The Present and Future of Whole Genome Sequencing (WGS) and Whole Metagenome Sequencing (WMS) for Surveillance of Antimicrobial Resistant Microorganisms and Antimicrobial Resistance Genes across the Food Chain. *Genes*. 2018 May 22;9(5).
5. Timme RE, Rand H, Shumway M, Trees EK, Simmons M, Agarwala R, et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ*. 2017;5:e3893.
6. Rantsiou K, Kathariou S, Winkler A, Skandamis P, Saint-Cyr MJ, Rouzeau-Szynalski K, et al. Next generation microbiological risk assessment: opportunities of whole genome

sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. *Int J Food Microbiol.* 2017 Dec 1;

7. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull.* 2017 08;22(23).
8. Deng X, den Bakker HC, Hendriksen RS. Genomic Epidemiology: Whole-Genome-Sequencing-Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annu Rev Food Sci Technol.* 2016;7:353–74.
9. Yachison CA, Yoshida C, Robertson J, Nash JHE, Kruczkiewicz P, Taboada EN, et al. The Validation and Implications of Using Whole Genome Sequencing as a Replacement for Traditional Serotyping for a National Salmonella Reference Laboratory. *Front Microbiol.* 2017;8:1044.
10. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *J Clin Microbiol.* 2015 Aug;53(8):2410–26.
11. Carrillo CD, Koziol AG, Mathews A, Goji N, Lambert D, Huszczyński G, et al. Comparative Evaluation of Genomic and Laboratory Approaches for Determination of Shiga Toxin Subtypes in *Escherichia coli*. *J Food Prot.* 2016;79(12):2078–85.
12. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, et al. In silico serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genomics [Internet].* 2016 Jul 11;2(7). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343136/>
13. Whiteside MD, Gannon VPJ, Laing CR. Phylotyper: in silico predictor of gene subtypes. *Bioinforma Oxf Engl.* 2017 Nov 15;33(22):3638–41.
14. Lindsey RL, Pouseele H, Chen JC, Strockbine NA, Carleton HA. Implementation of Whole Genome Sequencing (WGS) for Identification and Characterization of Shiga Toxin-Producing *Escherichia coli* (STEC) in the United States. *Front Microbiol.* 2016;7:766.
15. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, et al. Whole-Genome Sequencing for National Surveillance of Shiga Toxin–Producing *Escherichia coli* O157. *Clin Infect Dis Off Publ Infect Dis Soc Am.* 2015 Aug 1;61(3):305–12.
16. Neuert S, Nair S, Day MR, Doumith M, Ashton PM, Mellor KC, et al. Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal *Salmonella enterica*. *Front Microbiol [Internet].* 2018 [cited 2018 May 3];9. Available from: <https://www.frontiersin.org/articles/10.3389/fmicb.2018.00592/full>
17. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, et al. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J Antimicrob Chemother.* 2013 Apr;68(4):771–7.

18. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, et al. Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. *J Clin Microbiol.* 2016 Aug;54(8):1975–83.
19. Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, et al. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* [Internet]. [cited 2018 May 3]; Available from: <https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty212/4961427>
20. Alleweldt F, Kara S, Osinski A, Van Baal P, Kellerborg K, Aarestrup FM, et al. Developing a framework to assess the costeffectiveness of COMPARE - a global platform for the exchange of sequence-based pathogen data. *Rev Sci Tech Int Off Epizoot.* 2017 Apr;36(1):311–22.
21. Kwong JC, McCallum N, Sintchenko V, Howden BP. Whole genome sequencing in clinical and public health microbiology. *Pathology (Phila).* 2015 Apr;47(3):199–210.
22. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med.* 2015 Apr;139(4):481–93.
23. Kozyreva VK, Truong C-L, Greninger AL, Crandall J, Mukhopadhyay R, Chaturvedi V. Validation and Implementation of Clinical Laboratory Improvements Act-Compliant Whole-Genome Sequencing in the Public Health Microbiology Laboratory. *J Clin Microbiol.* 2017 Aug;55(8):2502–20.
24. Lambert D, Pightling A, Griffiths E, Van Domselaar G, Evans P, Berthelet S, et al. Baseline practices for the application of genomic data supporting regulatory food safety. *J AOAC Int* [Internet]. 2017 [cited 2017 Feb 9];100(3). Available from: <http://www.ingentaconnect.com/content/aoac/jaoac/pre-prints/content-jaoacint.160269>
25. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005 Oct;15(10):1451–5.
26. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available [Httpwwwbioinformaticsbabraham.ac.uk/projects/fastqc](http://www.bioinformaticsbabraham.ac.uk/projects/fastqc). 2010;
27. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016 20;17(1):132.
28. Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J, Iskander M, et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microb Genomics* [Internet]. 2017 Jun 8;3(6). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5628696/>
29. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* [Internet]. 2012 Jul 17; Available from: <http://arxiv.org/abs/1207.3907>

30. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010 May;59(3):307–21.
31. Magoc T, Salzberg S. FLASH: fast length adjustment of short reads to improve genome assemblies | *Bioinformatics* | Oxford Academic [Internet]. [cited 2018 Jun 1]. Available from: <https://academic.oup.com/bioinformatics/article/27/21/2957/217265>
32. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*. 2012 May;19(5):455–77.
33. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl*. 2014 Jul 15;30(14):2068–9.
34. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, et al. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. *PloS One*. 2016;11(1):e0147101.
35. Robertson J, Yoshida C, Kruczkiewicz P, Nadon C, Nichani A, Taboada EN, et al. Comprehensive assessment of the quality of Salmonella whole genome sequence data available in public sequence databases using the Salmonella in silico Typing Resource (SISTR). *Microb Genomics*. 2018 Jan 17;
36. Feijao P, Yao H-T, Fornika D, Gardy J, Hsiao W, Chauve C, et al. MentaLiST - A fast MLST caller for large MLST schemes. *Microb Genomics*. 2018 Jan 10;
37. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma Oxf Engl*. 2009 Jun 1;25(11):1422–3.
38. Parks DH, Mankowski T, Zangooei S, Porter MS, Armanini DG, Baird DJ, et al. GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PloS One*. 2013;8(7):e69885.
39. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res*. 2017 Jul 3;45(Web Server issue):W30–5.
40. Hingston P, Chen J, Dhillon BK, Laing C, Bertelli C, Gannon V, et al. Genotypes Associated with *Listeria monocytogenes* Isolates Displaying Impaired or Enhanced Tolerances to Cold, Salt, Acid, or Desiccation Stress. *Front Microbiol*. 2017;8:369.
41. Miller RR, Langille MGI, Montoya V, Crisan A, Stefanovic A, Martin I, et al. Genomic Analysis of a Serotype 5 *Streptococcus pneumoniae* Outbreak in British Columbia, Canada, 2005–2009. *Can J Infect Dis Med Microbiol J Can Mal Infect Microbiol Médicale* [Internet]. 2016;2016. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904568/>

42. Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, et al. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Front Microbiol.* 2017;8:375.
43. Bekal S, Berry C, Reimer AR, Van Domselaar G, Beaudry G, Fournier E, et al. Usefulness of High-Quality Core Genome Single-Nucleotide Variant Analysis for Subtyping the Highly Clonal and the Most Prevalent *Salmonella enterica* Serovar Heidelberg Clone in the Context of Outbreak Investigations. *J Clin Microbiol.* 2016 Feb;54(2):289–95.
44. Usongo V, Berry C, Yousfi K, Doualla-Bell F, Labbé G, Johnson R, et al. Impact of the choice of reference genome on the ability of the core genome SNV methodology to distinguish strains of *Salmonella enterica* serovar Heidelberg. *PloS One.* 2018;13(2):e0192233.
45. Woksepp H, Ryberg A, Berglind L, Schön T, Söderman J. Epidemiological characterization of a nosocomial outbreak of extended spectrum β -lactamase *Escherichia coli* ST-131 confirms the clinical value of core genome multilocus sequence typing. *APMIS Acta Pathol Microbiol Immunol Scand.* 2017 Dec;125(12):1117–24.
46. Roisin S, Gaudin C, De Mendonça R, Bellon J, Van Vaerenbergh K, De Bruyne K, et al. Pan-genome multilocus sequence typing and outbreak-specific reference-based single nucleotide polymorphism analysis to resolve two concurrent *Staphylococcus aureus* outbreaks in neonatal services. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis.* 2016 Jun;22(6):520–6.
47. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, et al. Core Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of *Enterococcus faecium*. *J Clin Microbiol.* 2015 Dec;53(12):3788–97.
48. Kingry LC, Rowe LA, Respcio-Kingry LB, Beard CB, Schrieffer ME, Petersen JM. Whole genome multilocus sequence typing as an epidemiologic tool for *Yersinia pestis*. *Diagn Microbiol Infect Dis.* 2016 Apr;84(4):275–80.
49. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, et al. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. *J Clin Microbiol.* 2014 Jul;52(7):2479–86.
50. Eisenstein M. Big data: The power of petabytes [Internet]. *Nature.* 2015 [cited 2018 May 3]. Available from: <https://www.nature.com/articles/527S2a>
51. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? *PLoS Biol* [Internet]. 2015 Jul 7;13(7). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494865/>
52. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship [Internet]. *Scientific Data.* 2016 [cited 2018 May 8]. Available from: <https://www.nature.com/articles/sdata201618>
53. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform.* 2006 Sep;7(3):256–74.

54. Griffiths E, Dooley D, Graham M, Van Domselaar G, Brinkman FSL, Hsiao WWL. Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance. *Front Microbiol* [Internet]. 2017 Jun 26;8. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5483436/>
55. Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, et al. Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud. *PloS One*. 2015;10(10):e0140829.
56. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 2017 Jan 4;45(D1):D566–73.
57. Thomsen MCF, Ahrenfeldt J, Cisneros JLB, Jurtz V, Larsen MV, Hasman H, et al. A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance. *PloS One*. 2016;11(6):e0157718.

FUNDING INFORMATION

This work was supported by the Government of Canada Genomics Research and Development Initiative, the Public Health Agency of Canada, Genome Canada, Genome British Columbia, Genome Atlantic, Cystic Fibrosis Canada and Compute Canada, with the support of AllerGen NCE Inc. Andrew G. MacArthur holds a Cisco Research Chair in Bioinformatics, supported by Cisco Systems Canada, Inc.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the IRIDA Consortium (Marsha Taylor, Eleni Galanis, Linda Hoang, Natalie Prystajecy, Kim Macdonald, Lynn Schriml, David Aanensen, Michel Dumontier and William Klimke), as well as Celine Nadon, Aleisha Reimer, Lorelee Tschetter, Natalie Knox, Shaun Tyler, Cameron Sieffert, Sukhdeep Sidhu, Shane Thiessen, Paul Williams the GRDI-FWS Consortium, and SFU Research Computing, for their critical feedback and support.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

ETHICAL STATEMENT

There is no human or animal work in this study.

FIGURE LEGENDS

Table 1. Comparison of commonly used public health genomics software according to best practices.

The IRIDA platform is compared with INNUENDO, WGSa, the Centre for Genomic Epidemiology (CGE), BioNumerics, SeqSphere, Nullarbor, COMPARE and NCBI public health genomics software, using cost and best practices criteria. Only IRIDA, INNUENDO and Nullarbor are considered completely open-source. Only IRIDA, SeqSphere and BioNumerics enable independent installations to perform private analyses in a decentralized system with data management and storage features, however work is underway to enable this functionality in INNUENDO. While many software packages and platforms record data provenance and parameters, some are predefined and cannot be changed by the user. All data created, modified, or deleted in IRIDA is audited with the timestamp, user, and tool that was used to modify the data. All software offer quality control checks. Only IRIDA software currently enables users to share data with other instances, as well as public repositories, in an automated fashion. Most non-commercial software is free; in contrast, the licenses for the commercial packages SeqSphere and BioNumerics carry significant costs.

Figure 1. IRIDA platform schematic overview.

The IRIDA platform provides an end-to-end solution for sequence data management and analysis. This includes multiple options for importing data into the system including a custom Uploader Tool that runs on an Illumina MiSeq sequencer, a web interface for uploading data, or synchronization from other IRIDA installations. IRIDA's authorization and security layer validates that a user is able to access or modify the data they are attempting to access. Once authenticated, IRIDA's auditing layer creates an audit trail for all data uploaded, modified, or deleted. IRIDA contains an analysis engine to perform quality control on all uploaded sequencing data, provides automation tools, and stores data provenance. The analysis engine communicates with Galaxy to schedule and execute its analysis pipeline tools. IRIDA also provides multiple secure data sharing options, ranging from exporting data to Galaxy for further analysis, to command-line tools, to data synchronization between IRIDA installations.

Figure 2. Project structure and data model.

IRIDA's data structure is rooted by a "Project". A Project is a container for any number of related samples, their sequencing data, and analysis results. Metadata associated with the project, sequences and analyses are also stored within the Project. A "Sample" container stores contextual data (e.g. environmental, clinical, food source, collection dates, and geographical locations) alongside sequencing data, better enabling integrated analyses. Data access is provided on a per-project basis. Users who are added to a project are given access to all data contained within the project.

Figure 3. Project Sample page with "Shopping cart" functionality.

IRIDA's project sample page provides point-and-click access to all sample data stored within a project. Sample data can be sorted, filtered, and managed from within this interface. From this page samples can be selected and grouped into IRIDA's "shopping cart". The cart provides a

familiar means of collecting samples into a temporary storage space for users to later perform further actions. The cart is the entry point for all analysis pipelines in the IRIDA system.

Figure 4. Pipeline launch.

The pipeline launch page makes it as easy for users to launch advanced analysis tools. After adding samples to the cart and selecting a pipeline, users may adjust the parameters—optionally saving these adjustments for application to future pipelines. Clicking the “Launch” button initiates the pipeline.

Figure 5. Example of a SNVPhyl phylogenomics analysis.

Once an analysis from one of IRIDA’s different pipelines is completed, the results are stored in IRIDA and can be viewed within the browser, e.g. the SNVPhyl circular SNV-based tree presented here. A number of pipelines have custom visualization tools associated with them to quickly view results. All analysis results can be exported for further analysis in the user's preferred external tools. Provenance information also can be viewed from this page, which displays all information about how the analysis pipeline was executed, including all tools, versions, and parameters that were used to build an analysis result.

Figure 6. Example of line list displaying contextual information.

IRIDA's line list display provides users with a high-level view of all available contextual metadata for samples in a given project. This view allows users to sort, import, and export any terms available in their samples. Imports and exports of data from IRIDA's metadata system are in Microsoft® Excel or csv format, allowing users to easily update or share data outside of IRIDA. Visibility of columns in the table can be toggled on or off, and rearranged, enabling different projects or users to have customized views on the same datasets. These views can be saved as a metadata template, allowing users to quickly return to a given view. Metadata templates can be defined by project managers, configuring metadata terms for their project or investigation.

Figure 7. Example of integrated phylogenomic tree with contextual data.

IRIDA's Advanced Phylogenomic Visualization tool combines results of IRIDA's SNVPhyl phylogenomic pipeline with the contextual metadata from the samples used to build the phylogenetic tree. This visualization tool aligns the metadata from IRIDA's line list tool with the leaves of the tree. This example displays a phylogenetic tree of *Salmonella* isolates, labelled with attributes such as sample source, pulsed-field gel electrophoresis (PFGE) typing information, antimicrobial resistance genes, and cgMLST plus SISTR sequence type data. Metadata can be toggled similar to the line list tool, and metadata templates can be applied to the fields. Colour coding is displayed to assist users in grouping metadata terms. This visualization can be exported in SVG format.

Figures and Tables

Table 1.

Criteria	IRIDA	INNUENDO	Center for Genomic Epidemiology	WGSA	Nullarbor	COMPARE	SeqSphere	BioNumerics	NCBI Pathogen Detection
Open-Source	✓	✓	Partially	✗	✓	Partially	Proprietary	Proprietary	Partially
Free	✓	✓	✓	✓	✓	✓	✗	✗	✓
Centralized/ Localized	Localized	Centralized	Centralized	Centralized	Localized	Centralized Data Hubs	Centralized (Star Topology)	Centralized (Star Topology)	Centralized
Interface	Web interface & REST-API	Web interface	Web interface	Web interface	Command line/ Web interface	Web interface	Graphical User Interface	Graphical User Interface	Web interface
Private Data Storage & Management	Private Projects and Samples	Private Projects and Samples; accessible by all users of each installation	Temporary storage; Operates as service only	Private Projects and Assemblies	Private Projects and Samples	Consortia Data Hubs (contain Private Projects and Samples)	Private Projects and Samples	Private Projects and Samples	Public repository with embargo periods for submitted sequences
Auditability	Provenance & customizable parameters captured and archived; metadata changes recorded and tracked	Predefined / Custom installation workflows could enable provenance & customizable parameter capture	Predefined parameters	Predefined parameters	Provenance & customizable parameters captured and archived	Project-dependent parameters	Provenance & customizable parameters captured and archived	Provenance & customizable parameters captured and archived	Limited documentation
Quality Control Checks	✓	✓	✓	✓	✓	✓	✓	✓	✓
Data Sharing	Between IRIDA instances and with NCBI SRA	Between users of the same installation / Manual	✗	Assemblies can be shared with public WGSA collection	Manual	Between members of a Data Hub	Central Repository	Central Repository	Public repository
Publication/ Website	irida.ca	sites.google.co	(57)	wgsa.net	github.com/tsee	(20)	ridom.com/seqsph	applied-	ncbi.nlm.nih.gov/p

		m/site/theinnu endoproject	genomicepidemi ology.org		mann/nullarbor	compare- europe.eu	ere/	maths.com/bionum erics	athogens
--	--	-------------------------------	-----------------------------	--	----------------	-----------------------	------	---------------------------	----------

Figure 1

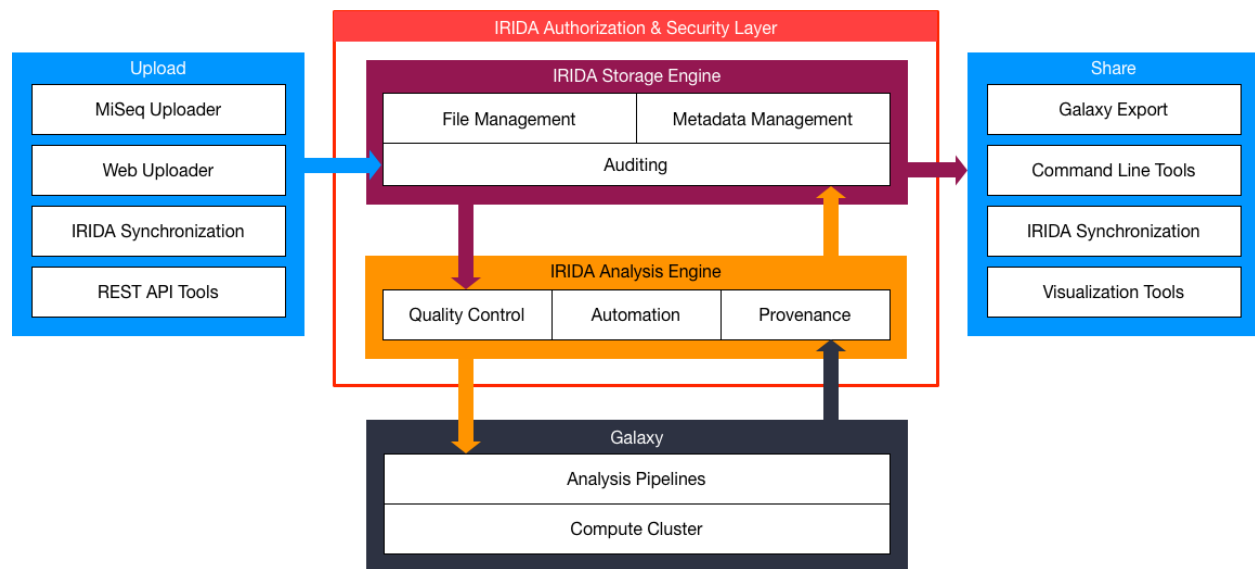


Figure 2

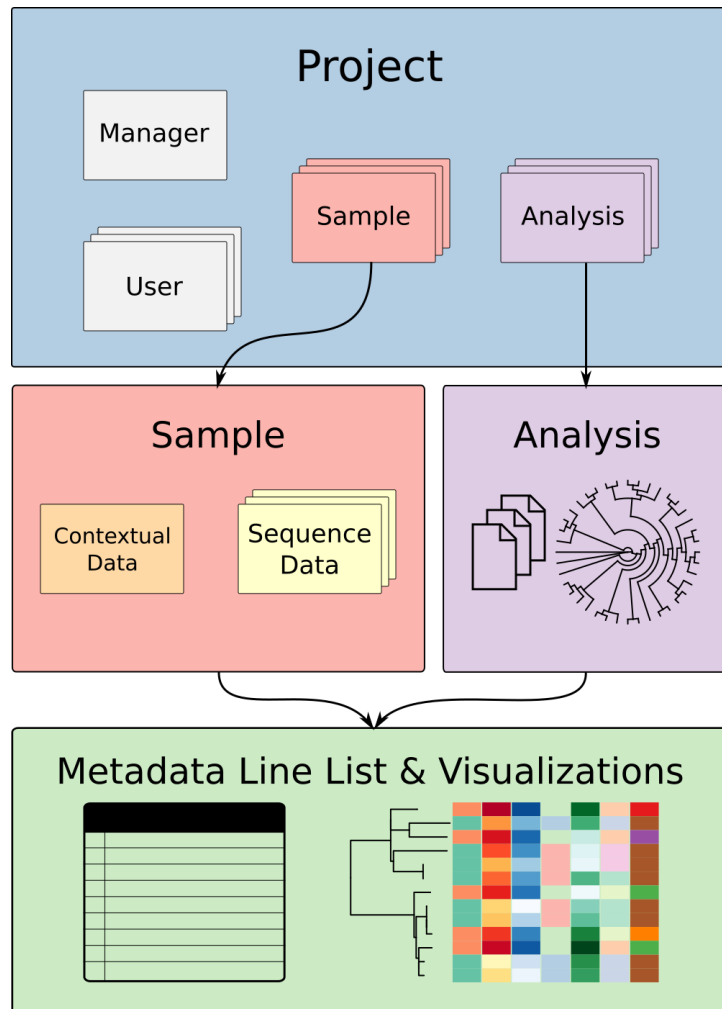


Figure 3

IRIDA Projects ▾ Analyses ▾ Search

Salmonella Surveillance ID 831

Samples Line List Analyses Details NCBI Exports Recent Activity Settings

Sample Tools ▾ Export ▾ Add to Cart

Select All Select None Search:

4 samples selected

Name	Organism	Project	Created On
☒ SH14-019	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☒ SH14-020	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☒ SH14-021	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☒ SH14-022	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☐ SH14-023	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☐ SH14-024	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☐ SH14-025	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☐ SH14-026	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☐ SH14-027	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM
☐ SH14-028	Salmonella	Salmonella Surveillance	Mar 12, 2018 3:54 PM

Show 10 ▾ entries

Previous 1 2 3 4 5 6 Next

Close Clear

Search Cart

- Salmonella Surveillance
 - SH14-022
 - SH14-021
 - SH14-020
 - SH14-019

Select a Pipeline ▾

Figure 4



Projects

Analyses

Q Search

Help

5

Settings

Profile

SNVPhyl Phylogenomics Pipeline

Pipeline Name

SNVPhyl_20180606

Reference File

S_HeidelbergSL476.fasta (Salmonella Surveillance)

Upload New

Parameters

Default Parameters

Customize

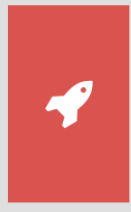
Description

Enter description of the analysis (Optional)

Share Results with Projects

☐ Salmonella Surveillance

Ready to Launch?



Files

Salmonella Surveillance / SH14-023

Remove

→	SH14-023_S1_L001_R1_001.fastq	210.2 MB File Size	12 Mar 2018 Date Created
←	SH14-023_S1_L001_R2_001.fastq	211.8 MB File Size	12 Mar 2018 Date Created

Salmonella Surveillance / SH14-019

Remove

→	SH14-019_S1_L001_R1_001.fastq	206.4 MB File Size	12 Mar 2018 Date Created
←	SH14-019_S1_L001_R2_001.fastq	208.9 MB File Size	12 Mar 2018 Date Created

Salmonella Surveillance / SH14-021

Remove

→	SH14-021_S1_L001_R1_001.fastq	227.2 MB File Size	12 Mar 2018 Date Created
←	SH14-021_S1_L001_R2_001.fastq	229.0 MB File Size	12 Mar 2018 Date Created

Salmonella Surveillance / SH14-022

Remove

→	SH14-022_S1_L001_R1_001.fastq	227.9 MB File Size	12 Mar 2018 Date Created
←	SH14-022_S1_L001_R2_001.fastq	229.5 MB File Size	12 Mar 2018 Date Created

Salmonella Surveillance / SH14-020

Remove

→	SH14-020_S1_L001_R1_001.fastq	224.3 MB File Size	12 Mar 2018 Date Created
←	SH14-020_S1_L001_R2_001.fastq	226.1 MB File Size	12 Mar 2018 Date Created

Figure 5

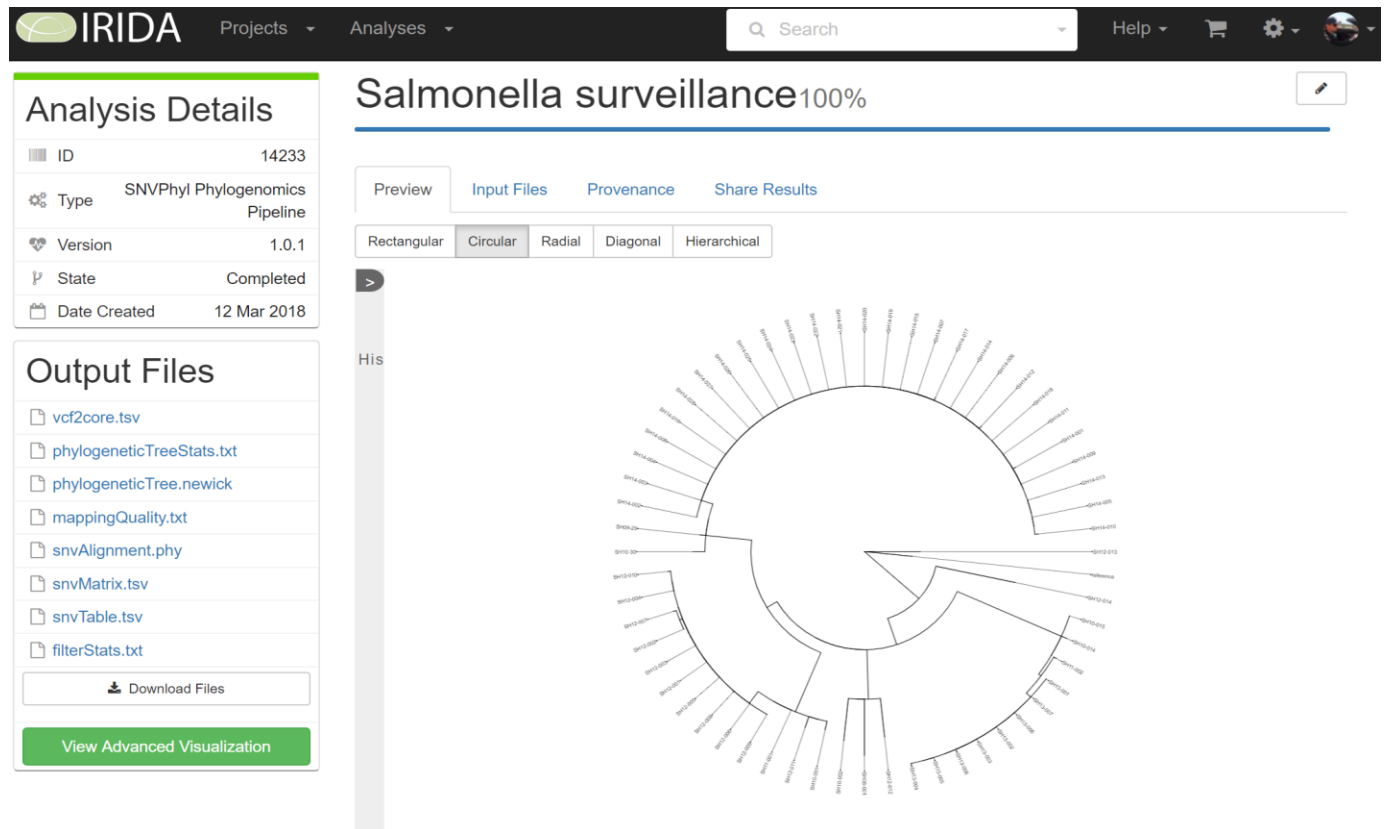


Figure 6


Projects ▾ Analyses ▾

Help ▾

Salmonella Surveillance ID 831

Samples
Line List
Analysis
Details
NCBI Exports
Recent Activity
Settings

Toggle Metadata Columns
Templates
Salmonella Surveillance ▾
Save Template
Upload Bulk Metadata

Export
Search:

Sample Name	Canadian PFGE pattern	Outbreak num	AMR Gene	SISTR cgMLST Sequence Type	Source
SH08-001	SHXAI.0001/SHBNI.0001	NA	blaCMY-2, fosA7	3016918627	Human
SH09-29	SHXAI.0001/SHBNI.0001	NA	fosA7	1860662990	Human
SH10-001	SHXAI.0001/SHBNI.0001	NA	fosA7	2243726273	Human
SH10-002	SHXAI.0001/SHBNI.0001	NA	fosA7	1372851835	Human
SH10-014	SHXAI.0001/SHBNI.0001	NA	fosA7	2550372349	Human
SH10-015	SHXAI.0001/SHBNI.0001	NA	fosA7	1326888805	Human
SH10-30	SHXAI.0001/SHBNI.0001	NA	blaCMY-2, fosA7	1266334878	Human
SH11-001	SHXAI.0001/SHBNI.0001	NA	fosA7	1860662990	Human
SH11-002	SHXAI.0001/SHBNI.0001	NA	fosA7	2990194624	Human
SH12-001	SHXAI.0001/SHBNI.0001	1	fosA7	1860662990	Human
SH12-002	SHXAI.0001/SHBNI.0001	1	fosA7	1725085330	Human
SH12-003	SHXAI.0001/SHBNI.0001	1	fosA7	1860662990	Human

Showing 1 to 59 of 59 entries

Figure 7

