Mechanizing our biologists with Galaxy and IRIDA platforms

GCC 2017, Montpellier
Philip Mabon
June 30, 2017
Outline

• Background of National microbiology laboratory (NML)
• History of bioinformatics core at NML
• Current overview of our bioinformatics lab
• Example of pipelines creation
Background
Background

Background

Bioinformatics

Outbreak

Research

Surveillance
Background: Outbreak & Surveillance

Clinical

Food

Environmental

Sample & Culture Bacteria

Fingerprint
Background: Outbreak & Surveillance

Database of fingerprints

Move towards using

Sample & Culture Bacteria

Whole Genome Sequencing

ATCATTC
CGGAATATG
GCCGAAT

Genomic fingerprint
Background: Research

• Support multiple different in house laboratories on general research
  – Bacteriology & Enterics
  – Viral Diseases
  – Zoonotic Diseases and Special Pathogens (ZDSP)
  – Science and Technology Core Services
    • Genomics
    • Proteomics and Mass Spectrometry
    • Bioforensics Assay Development and Diagnostics

http://www.phac-aspc.gc.ca/id-mi/index-eng.php
HISTORY OF BIOINFORMATICS
CORE
Dark times:
Pre Galaxy (-2011)
Dark times:
Pre Galaxy (-2011)
Dark times: 
Pre Galaxy (-2011)

• Accessibility: **Poor**
  – Command line only

• Data management: **Painful**
  – Where data located?
  – Which is the latest version?
  – Data duplication

• Provenance: **Difficult**
  • Tools, workflows, datasets

• Productivity: **Costly**
  – How much science can be done!
Dark times: Pre Galaxy (-2011)

Projects to Bioinformatician ratio

– One project to N
– One project to 1
– Multiple projects to 1
Age of scripting: Galaxy + API scripts (2011-2013)
Age of scripting :
Galaxy + API scripts (2011-2013)

Bulk importing of fastq files into a data library

usage: data_importer.py [-h] --api_key API_KEY --api_url API_URL --import_dir IMPORT_DIR --file_type FILE_TYPE --folder_dest FOLDER_DEST --library_id LIBRARY_ID [--link_data] [--verbose]

optional arguments:
-h, --help show this help message and exit
--api_key API_KEY Your API key from Galaxy
--api_url API_URL Url of the Galaxy instance to submit your request to.
i.e: http://galaxy
--import_dir IMPORT_DIR Based directory to be imported or linked to in Galaxy
--file_type FILE_TYPE
--folder_dest FOLDER_DEST
--library_id LIBRARY_ID
--link_data Link to the files instead of copying to Galaxy
--verbose

Looks like you don't know what you are doing... That's ok me too!
Age of scripting: Galaxy + API scripts (2011-2013)

Bulk workflow submission into multiple histories

usage: execute_workflow.py [-h] --api_key API_KEY --api_url API_URL
    --workflow_id WORKFLOW_ID --data_library DATA_LIBRARY
    --start_input START_INPUT
    [--reference REFERENCE] [--whitelist Whitelist] [--run_as RUN_AS]

Short sample app

optional arguments:
   -h, --help         show this help message and exit
   --api_key API_KEY
   --api_url API_URL
   --_workflow_id WORKFLOW_ID
   --data_library DATA_LIBRARY
   --start_input START_INPUT
   --reference REFERENCE
   --whitelist Whitelist
   --run_as RUN_AS
Age of scripting:
Galaxy + API scripts (2011-2013)

Bulk download of datasets from multiple histories

usage: extract_datasets.py [-h] --api key API_KEY --api url API_URL
--file name FILE_NAME --history name HISTORY_NAME
--outdir OUTDIR [--verbose]

Optional arguments:
-h, --help show this help message and exit
--api key API_KEY
--api url API_URL
--file name FILE_NAME
--history name HISTORY_NAME
--outdir OUTDIR
--verbose
Age of scripting: Galaxy + API scripts (2011-2014)

- **Accessibility**: **Worst**
  - Command line and web interface
- **Data management**: **Painful**
  - Where data located?
  - Which is the latest version?
  - Data duplication
- **Provenance**: **Good**
  - Tools, workflows, datasets
- **Productivity**: **Fair**
  - Copy and paste commands works
  - Dealt with dual interfaces
Enlightenment:
Galaxy Collection + IRIDA (2014-)
**Enlightenment:**

**Galaxy Collection + IRIDA (2014-)**

<table>
<thead>
<tr>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collection test</strong></td>
</tr>
<tr>
<td>6 shown</td>
</tr>
<tr>
<td>26.32 GB</td>
</tr>
</tbody>
</table>

- **6:** SRR1664628:reverse
- **5:** SRR1664628:forward
- **4:** SRR1664627:reverse
- **3:** SRR1664627:forward
- **2:** SRR1664626:reverse
- **1:** SRR1664626:forward

### Build List of Dataset Pairs

- Hide datasets
- Unhide datasets
- Delete datasets
- Undelete datasets
- Permanently delete datasets
- Build Dataset List
- Build Dataset Pair

### Collection test

<table>
<thead>
<tr>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collection test</strong></td>
</tr>
<tr>
<td>7 shown</td>
</tr>
<tr>
<td>26.32 GB</td>
</tr>
</tbody>
</table>

- **7:** Candida auris
  - a list of 3 dataset pairs
Enlightenment:

Galaxy Collection + IRIDA (2014-)

Results!
Getting data into IRIDA

- Manual web interface upload
- Automated instrument upload (Illumina MiSeq)

http://github.com/phac-nml/irida-miseq-uploader
Automatic analyses

- IRIDA
  - FastQC
    - On Upload
      - Default
    - SPAdes
      - Not Default
    - SISTR
      - Not Default
Data management
# Data management

## Other Metadata

<table>
<thead>
<tr>
<th>Location</th>
<th>first_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB</td>
<td>John</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>identifier</th>
<th>Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>Eggs, French Fries</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>last_name</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doe</td>
<td>Winnipeg</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ReceivedDate</th>
<th>HealthAuthority</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-07-03</td>
<td>WRHA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SourceType</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>34</td>
</tr>
</tbody>
</table>
Architecture

Sequencer

Researcher

Servlet Container

REST API

Application Logic

Web Interface

Central File Storage

Compute Cluster

Galaxy

Researcher

http://github.com/phac-nml/irida
Getting data out of IRIDA

- Downloading from browser
- Exporting to the command-line
  - Remote or locally
- Export to external Galaxy instance

Data Analysis

- Assembly and Annotation (Collection) Pipeline
- SNVPhyl Phylogenomics Pipeline
- SISTR Pipeline

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly and Annotation Pipeline</td>
<td>Generate an assembled and annotated genome from the reads within a sample using FLASH, SPAdes, and Prokka. Outputs analyzed and produced separately for each sample include: log files, assembly statistics, the contigs (all contigs, filtered contigs with repeats, filtered contigs without repeats), and annotations from Prokka.</td>
</tr>
<tr>
<td>Assembly and Annotation Collection Pipeline</td>
<td>The assembly and annotation collection pipeline provides the same results as the assembly and annotation pipeline, but all samples are analyzed together which allows you to download a single package for all samples submitted.</td>
</tr>
<tr>
<td>SNVPhyl Phylogenomics Pipeline</td>
<td>Generate a Whole Genome Phylogeny from a set of samples and a reference genome based on Single Nucleotide Polymorphisms (SNVs) using the SNVPhyl pipeline. This will provide a dendrogram as well as a table of all SNVs used and a SNV distance matrix between each sample.</td>
</tr>
<tr>
<td>SISTR Pipeline</td>
<td>Generates in silico typing results using the Salmonella In Silico Typing Resource (SISTR). This assembles a genome and runs the resulting contigs through <a href="https://github.com/peterk87/sistr_cmd/">https://github.com/peterk87/sistr_cmd/</a> to generate the final result.</td>
</tr>
</tbody>
</table>
Data Analysis + Metadata

Advance Phylogenetic Analysis

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Source Type</th>
<th>City</th>
<th>Location</th>
<th>Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP1032</td>
<td>Human</td>
<td>Prince Albert</td>
<td>SK</td>
<td>milkshake</td>
</tr>
<tr>
<td>CP1042</td>
<td>Human</td>
<td>Victoria</td>
<td>BC</td>
<td>spinach</td>
</tr>
<tr>
<td>VC19</td>
<td>Human</td>
<td>Brandon</td>
<td>MB</td>
<td>shrimp salad</td>
</tr>
<tr>
<td>VC6</td>
<td>Human</td>
<td>Winnipeg</td>
<td>MB</td>
<td>Eggs, French Fries</td>
</tr>
<tr>
<td>VC18</td>
<td>Human</td>
<td>Whistler</td>
<td>BC</td>
<td>roast beef</td>
</tr>
<tr>
<td>VC25</td>
<td>Human</td>
<td>Regina</td>
<td>SK</td>
<td>mayo chicken salad</td>
</tr>
<tr>
<td>VC26</td>
<td>Human</td>
<td>Vancouver</td>
<td>BC</td>
<td>egg salad</td>
</tr>
<tr>
<td>HC06A1</td>
<td>Human</td>
<td>Prince Albert</td>
<td>SK</td>
<td>cucumber sushi</td>
</tr>
</tbody>
</table>

reference
Data Analysis

1. IRIDA
2. Galaxy
3. Command line

Flexibility
Ease

Enlightenment:
Galaxy Collection + IRIDA (2014-)

- **Accessibility**: **Good**
  - Web interface only!
  - Command line available if needed
- **Data management**: **Enjoyable**
  - Where data located?: IRIDA
  - Which is the latest version?: IRIDA
  - Data duplication: linking
- **Provenance**: **Awesome**
  - Tools, workflows, datasets
- **Productivity**: **Very Good**
  - Bulk submission is easy!
  - More science can be done!

![Self-sufficient users](chart)
![Bioinformatician](chart)
Current bioinformatics environment (2017)
Current bioinformatics environment (2017)

**Production: IRIDA**
- Semi-static
- Regular updates (monthly & yearly)
- Documentation (SOPs)

**Development: Non-IRIDA**
- 100% flexible
- Irregular updates (when needed)
- Less comprehensive documentations

---

**IRIDA**
- REST API
- Web Interface
- Application Logic
- Galaxy

**Common**
- Central File Storage
- Compute Cluster

**Dev**
- Galaxy
- $\sim >_-$
• 2013-2015 - 10 projects/month
• 2016 - 19 projects/month
IRIDA Samples

- Average 75 samples/project (max 1095)
IRIDA Analyses

SNVPhyl 1.0 added to IRIDA
Dev Galaxy Jobs

Average 100k jobs per month
Current bioinformatics environment (2017)
Process : the Request

From: Technician@Canada.ca
To: BioInfo-Support@Canada.ca
Subject: Help me please!
Is there a quick way of doing present or absent of specific region?

E-mail request

Internal biohub

Kraken databases
Is it possible to get some of the PPP databases configured for Kraken. It would be handy to have organism type specific databases like bacterial, viral, fungi, etc. for targeted read searching.
Process: the search

Search and Install
• Main Galaxy ToolShed
• Bioconda

Implement & Share
• Planemo
• Bioconda
• Sharing to Main Galaxy ToolShed

http://maxpixel.freegreatpicture.com/static/photo/1x/Smiley-Pray-Religion-Christianity-Church-Prayer-2003386.png

https://pixabay.com/p-1294361
Process: implementation & testing

- Initial design of workflow and publish to user
- Designate workflow as **BETA**

**BETA: K-mer search with KAT v1.1.1 Subtyping**
Process: implementation & testing

Initial implement

Local Publish

Refine

Test by user

End product

Changes suggested

Cry

Go home and rethink my life
Process: Documentation and training

• Tri-level documentations:
  – Developer documentation
  – Workflow documentation
  – High level description of workflow by end-user

• Training as needed
  – By e-mail
  – Walkthrough over the phone
  – In person
Process: Integration 2 IRIDA platform

• Tools & workflows publicly available
• Write some java code
  – Indicate inputs & workflow
  – Which files to pull from Galaxy
  – Post processing
    • Interactive tabular data
    • Phylogenomics tree
Process: Documentation again?

Formal documentation and in-class training course

Certificate of Completion

THIS ACKNOWLEDGES THAT

Billy Bishop

HAS SUCCESSFULLY COMPLETED IRIDASNVPHYL PIPELINE COURSE

PHILIP MASON, INSTRUCTOR       JUNE 30, 2017

https://www.flickr.com/photos/12873321@N05/19620861981
Acknowledgements

Project Leaders
Fiona Brinkman – SFU
Will Hsiao – PHMRL
Gary Van Domselaar – NML

University of Lisbon
João Carriço

National Microbiology Laboratory (NML)
Franklin Bristow
Aaron Petkau
Thomas Matthews
Josh Adam
Adam Olsen
Tara Lynch
Shaun Tyler
Philip Mabon
Philip Au
Celine Nadon
Cameron Sieffert
Natalie Knox
Morag Graham
Chrystal Berry
Lorelee Tschetter
Eduardo Toboada
Peter Kruczkwiewicz
Chad Laing
Vic Gannon
Matthew Whiteside
Ross Duncan
Steven Mutschall
Jennifer Cabral

Simon Fraser University (SFU)
Emma Griffiths
Geoff Winsor
Julie Shay
Matthew Laird
Melanie Courtot
Bhav Dhillon
Raymond Lo

BC Public Health Microbiology & Reference Laboratory (PHMRL) and BC Centre for Disease Control (BCCDC)
Judy Isaac-Renton
Patrick Tang
Natalie Prystajecky
Jennifer Gardy
Linda Hoang
Kim MacDonald
Yin Chang
Eleni Galanis
Marsha Taylor
Damion Dooley
Cletus D'Souza
Ana Paccagnella

University of Maryland
Lynn Schriml

Canadian Food Inspection Agency (CFIA)
Burton Blais
Catherine Carrillo
Dominic Lambert

Dalhousie University
Rob Beiko
Alex Keddy
Questions?