Managing Next Generation Sequencing Data with iRODS

Presented by Dan Bedard // danb@renci.org

at the 9th International Conference on Genomics
Shenzhen, China
September 12, 2014
Background: Problem Statement

• Next Generation Sequencing (NGS) results in lots of data
  – Several GB/genome, for each processing step.
  – cp feels safer than mv
  – We don’t know what we don’t know.
• Next Generation Sequencing (NGS) results in lots of data
  – Several GB/genome, for each processing step.
  – cp feels safer than mv
  – We don’t know what we don’t know.
    • We need to store *everything*.

• We need to know where it came from.

• We need to be able to find it.

• Well... “we” doesn’t include everybody.
  We need to secure it... but we still need to share it.
Background: Problem Statement

• Next Generation Sequencing (NGS) results in lots of data
  – Several GB/genome, for each processing step.
  – cp feels safer than mv
  – We don’t know what we don’t know.
    • We need to store everything.

  • We need to know where it came from.

  • We need to be able to find it.

  • Well... “we” doesn’t include everybody.
    We need to secure it... but we still need to share it.

And time is of the essence!
Background

• How do the world’s preeminent bioinformatics centers manage their data?
  – Beijing Genomics Institute (BGI)
  – The Wellcome Trust Sanger Institute (WTSI)
  – The Broad Institute
  – The International Neuroinformatics Coordinating Facility (INCF)
  – The iPlant Collaborative
  – UNC Lineberger Comprehensive Cancer Center
  – Uppsala Genome Center
  – Public Health England
  – “Life Science Industrial Users”
Agenda

• What is iRODS?
• How are People Using It?
• Reference Implementation for NGS
What is iRODS?

iRODS is open source data grid middleware for...

- Data Discovery
- Workflow Automation
- Secure Collaboration
- Data Virtualization
WAIT...

WHAT?
What is iRODS?

- free to use
- free to modify
- free to contribute

sits between
the files and the user

iRODS is open source data grid middleware for...

- Data Discovery
  - metadata
- Workflow Automation
  - policies: any condition; any action
- Secure Collaboration
  - sharing without losing control
- Data Virtualization
  - file system flexibility

iRODS is the underlying technology for the world’s preeminent genomic research institutes. iRODS is an infinitely configurable data janitor. iRODS is the kind of technology you need to host everyone’s unstructured data. iRODS is a powerful data migration tool. iRODS is the technology that underpins the iPlant Data Store. iRODS is a data preservation technology. iRODS is a fundamental technology for CineGrid. iRODS is a tool for providing fine-grained privacy and security controls. iRODS is extensible: iRODS has command-line clients, APIs for numerous programming languages, and web clients. iRODS supports new plug-ins for storage resources, authentication mechanisms, microservices, and network protocols.
iRODS: Ready for Enterprise

• Product of nearly 20 years of research and development, funded by DARPA, DOE, NASA, NSF, NARA, and NOAA.

• Starting with iRODS 4.0, the entire codebase has been reviewed and restructured for enterprise use.
  – Each change is verified with a test case in a continuous integration suite
  – Pre-compiled binary packages are available for several Linux distributions and multiple database management systems.
iRODS: The iRODS Consortium

• Founded to ensure that iRODS continues to be free open source software.
• Four levels of membership, with increasing levels of involvement
  – Participation in technical planning and governance
  – Contact, co-marketing, sales support
  – Discretionary staff hours

• Stakeholders who recognize the value of sustaining iRODS development.
• Currently:
  – RENCI
  – The DICE Center
  – DataDirect Networks
  – Seagate
  – The Wellcome Trust Sanger Institute
  – EMC Corporation

• Additionally, the iRODS Consortium provides professional integration services, training, and support on a contract basis to iRODS users.

• Learn more at [iRODS.org/consortium](https://irods.org/consortium) or contact us at [info@irods.org](mailto:info@irods.org)
Use Case: Wellcome Trust Sanger Institute

• Large Scale Genomics Research
  – Sequenced 1/3 of the human genome (largest single contributor)
  – Active cancer, malaria, pathogen, and genomic variation studies
  – All data publicly available through websites, FTP, direct database access, programmatic APIs

• 2 PB of data managed by iRODS
Use Case: Wellcome Trust Sanger Institute

Using iRODS for...

**Data Discovery**
Metadata for tracking origin and processing history

Example attribute fields →

- attribute: library
- attribute: total_reads
- attribute: type
- attribute: lane
- attribute: is_paired_read
- attribute: study_accession_number
- attribute: library_id
- attribute: sample_accession_number
- attribute: sample_public_name
- attribute: manual_qc
- attribute: tag
- attribute: sample_common_name
- attribute: md5
- attribute: tag_index
- attribute: study_title
- attribute: study_id
- attribute: reference
- attribute: sample
- attribute: target
- attribute: sample_id
- attribute: id_run
- attribute: study
- attribute: alignment

Users query and access data largely from local compute clusters

Users access iRODS locally via the CLI

Copyright Wellcome Trust Sanger Institute. Used with permission.
Use Case: Wellcome Trust Sanger Institute

Using iRODS for...

**Data Virtualization**

with Workflow Automation

Seamless data replication, automatic checksumming, policy-based data resource selection

- Data lands by preference on “green room” storage, when available.
- Replicated, with checksums, to “red room” storage.
- Read access served by both rooms.
- Integrity verified in flight.

Copyright Wellcome Trust Sanger Institute. Used with permission.
Use Case: Wellcome Trust Sanger Institute

Using iRODS for...

**Secure Collaboration**

Selectively sharing data between workgroups; isolation for maintenance operations; options for defining policy on a per-group basis

Sanger 1 Portal zone (provides Kerberised access)

Federation using head zone accounts

/seq  /uk10k  /humgen  /Archive

Copyright Wellcome Trust Sanger Institute. Used with permission.
Use Case: The Broad Institute

• Harvard-MIT collaboration focused on cross-disciplinary challenges in biology and medicine.

• Small pilot program using iRODS to archive 9TB of data.
Use Case: The Broad Institute

Using iRODS for...

**Data Discovery and Workflow Automation**

Metadata automatically generated from original file system, used to enforce policy and verify integrity.

**Policy 1** – Validate, checksum, replicate, compress

**Policy 2** – Users cannot delete files

**Policy 3** – Purge files by expiration date

---

<table>
<thead>
<tr>
<th>Read From Original File Attributes</th>
<th>User Parameter</th>
<th>Calculated</th>
</tr>
</thead>
<tbody>
<tr>
<td>broadUser</td>
<td>broadModifyTime</td>
<td>broadExpireDate</td>
</tr>
<tr>
<td>broadUid</td>
<td>broadModifyTimestamp</td>
<td>broadExpireTimestamp</td>
</tr>
<tr>
<td>broadGroup</td>
<td>broadCreateTime</td>
<td>broadEntryDate</td>
</tr>
<tr>
<td>broadGid</td>
<td>broadCreateTime</td>
<td>broadEntryTimestamp</td>
</tr>
<tr>
<td>broadFileMode</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copyright Distributed Bio LLC. Used with permission.
Use Case:
UNC Lineberger Comprehensive Cancer Center

• One of the leading cancer centers in the US
• Highly collaborative research between UNC departments
Use Case:
UNC Lineberger Cancer Research Center

Data Virtualization with Workflow Automation
Automatically staging data for HPC and interpretation; using hardware from multiple vendors; complex access control

Using iRODS for...

Hadoop computing, SAN scratch storage

Additional Processing, Secondary Archive

Primary Processing, Storage, and Archive

Secure Medical Workspace

Sequencing center

Isilon

UNC Kure HPC

Dell

Genomic Sciences TopSail Hadoop

iRODS Grid

RENCI Croatan ‘Big Data’

DDN, Quantum

RENCI BlueRidge HPC

i

Primary Processing, Storage, and Archive

Secure Medical Workspace

Sequencing center
Use Cases: The Upshot

iRODS is finding a permanent home at NGS sites because of:

• Metadata!

• Vendor neutrality
  – Not subject to storage vendor lock-in
  – Mitigates risk of vendor termination

• Open source
  – Mitigate risk of developer termination

• Flexibility
  – Policy enforcement: any trigger, any action
  – Storage virtualization: layers-deep replication; local<->cloud
  – User permissions

• Sharing between workgroups
What’s Next?
NGS Reference Implementation

- Initialization
- Sequencing
- Formatting and Cleaning
- Quality Control
- Standard Analytical Processing
- Querying
- Interpretation
- Consultation
- Additional Action (ex. Treatment)
- Archive/Replication
What’s Next?
NGS Reference Implementation

Initialization

Sequencing

Formatting and Cleaning

Quality Control

Standard Analytical Processing

Querying

Interpretation

Consultation

Additional Action (ex. Treatment)

Archive/Replication

iRODS will apply sample IDs and results (or links to results) of processing and interpretation.
What’s Next?
NGS Reference Implementation

- Initialization
- Sequencing
- Formatting and Cleaning
- Quality Control
- Standard Analytical Processing
- Querying
- Interpretation
- Consultation
- Additional Action (ex. Treatment)
- Archive/Replication

iRODS will apply sample IDs and results (or links to results) of automated processing.

iRODS will kick off each process in the pipeline, or launch a workflow engine for more complex tasks.
What’s Next? NGS Reference Implementation

- Initialization
- Sequencing
- Formatting and Cleaning
- Quality Control
- Standard Analytical Processing
- Querying
- Interpretation
- Consultation
- Additional Action (ex. Treatment)
- Archive/Replication

iRODS will apply sample IDs and results (or links to results) of automated processing.

iRODS will kick off each process in the pipeline, or launch a workflow engine for more complex tasks.

iRODS will automatically compile reports upon schedule or request.
### What’s Next?
#### NGS Reference Implementation

<table>
<thead>
<tr>
<th>Process</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization</td>
<td>iRODS will apply sample IDs and results (or links to results) of automated processing</td>
</tr>
<tr>
<td>Sequencing</td>
<td>iRODS will kick off each process in the pipeline, or launch a workflow engine for more complex tasks.</td>
</tr>
<tr>
<td>Formatting and Cleaning</td>
<td>iRODS will automatically compile reports upon schedule or request</td>
</tr>
<tr>
<td>Quality Control</td>
<td>iRODS will stage files for processing, evaluation on a secure workspace, and archiving</td>
</tr>
<tr>
<td>Standard Analytical Processing</td>
<td></td>
</tr>
<tr>
<td>Querying</td>
<td></td>
</tr>
<tr>
<td>Interpretation</td>
<td></td>
</tr>
<tr>
<td>Consultation</td>
<td></td>
</tr>
<tr>
<td>Additional Action (ex. Treatment)</td>
<td></td>
</tr>
<tr>
<td>Archive/Replication</td>
<td></td>
</tr>
</tbody>
</table>
## What’s Next?
### NGS Reference Implementation

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization</td>
<td>iRODS will apply sample IDs and results (or links to results) of automated processing</td>
</tr>
<tr>
<td>Sequencing</td>
<td>iRODS will kick off each process in the pipeline, or launch a workflow engine for more complex tasks.</td>
</tr>
<tr>
<td>Formatting and Cleaning</td>
<td>iRODS will automatically compile reports upon schedule or request</td>
</tr>
<tr>
<td>Quality Control</td>
<td>iRODS will stage files for processing, evaluation on a secure workspace, and archiving</td>
</tr>
<tr>
<td>Standard Analytical Processing</td>
<td>iRODS will search on metadata</td>
</tr>
<tr>
<td>Querying</td>
<td></td>
</tr>
<tr>
<td>Interpretation</td>
<td></td>
</tr>
<tr>
<td>Consultation</td>
<td></td>
</tr>
<tr>
<td>Additional Action (ex. Treatment)</td>
<td></td>
</tr>
<tr>
<td>Archive/Replication</td>
<td></td>
</tr>
</tbody>
</table>
## What’s Next?
### NGS Reference Implementation

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization</td>
<td>iRODS will apply sample IDs and results (or links to results) of automated processing</td>
</tr>
<tr>
<td>Sequencing</td>
<td>iRODS will kick off each process in the pipeline, or launch a workflow engine for more complex tasks.</td>
</tr>
<tr>
<td>Formatting and Cleaning</td>
<td>iRODS will automatically compile reports upon schedule or request</td>
</tr>
<tr>
<td>Quality Control</td>
<td>iRODS will stage files for processing, evaluation on a secure workspace, and archiving</td>
</tr>
<tr>
<td>Standard Analytical Processing</td>
<td>iRODS will search on metadata</td>
</tr>
<tr>
<td>Querying</td>
<td>iRODS will manage complex, dynamic user permissions across multiple workgroups</td>
</tr>
<tr>
<td>Interpretation</td>
<td></td>
</tr>
<tr>
<td>Consultation</td>
<td></td>
</tr>
<tr>
<td>Additional Action (ex. Treatment)</td>
<td></td>
</tr>
<tr>
<td>Archive/Replication</td>
<td></td>
</tr>
</tbody>
</table>
What’s Next?

• Create the reference genomics implementation

• Document it in a “cookbook,” so other NGS centers can adapt and implement systems
  – Examples: Replication, Policy-based storage selection, User interface API, Access control policies, Archiving policies
What’s Next?
NGS Reference Implementation

- **Initialization**
- **Sequencing**
- **Formatting and Cleaning**
- **Quality Control**
- **Standard Analytical Processing**
- **Querying**
- **Interpretation**
- **Consultation**
- **Additional Action (ex. Treatment)**
- **Archive/Replication**

**iRODS will apply sample IDs and results (or links to results) of automated processing.**

**iRODS will kick off each process in the pipeline, or launch a workflow engine for more complex tasks.**

**iRODS will automatically compile reports upon schedule or request.**

**iRODS will stage files for processing, evaluation on a secure workspace, and archiving.**

**iRODS will search on metadata.**

**iRODS will manage complex, dynamic user permissions across multiple workgroups.**
What’s Next?

We need partners
- LIMS integration
- System integrators
- Data processing with native I/O
- Storage/computing appliance vendors
- Hosts for training
- Users to shape the problem space and evaluate our solution

Get involved
- Contact info@irods.org
- Follow us on Github: https://github.com/irods/irods
- Read our blog: http://irods.org/controlyourdata/
- Join the conversation on iRODS Chat: https://groups.google.com/forum/#!forum/iROD-Chat
Acknowledgments

Thank you to:

- DARPA, DOE, NASA, NSF, NARA, and NOAA for supporting the creation of iRODS
- iRODS Consortium Members, for their continued support
- John Constable, WTSI
- Chris Smith, Distributed Bio
- Sai Balu, Lineberger Cancer Research Center
- Genomics researchers, for coming up with interesting problems

Except where specified otherwise, this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
Thank you!

Dan Bedard
iRODS Market Development Manager
danb@renci.org