

## Using E-iRODS in the Management of Human Genomic Data for Research and Clinical Use

Charles Schmitt<sup>1</sup>, Chris Bizon<sup>1</sup>, Phil Owen<sup>1</sup>, Joshua Sailsbery<sup>1</sup>, Jason Reilly<sup>1</sup>, Xiaoshu Wang<sup>1</sup>, Erik Scott<sup>1</sup>, Michael Shoffner<sup>1</sup>, Nassib Nassar<sup>1</sup>, Kirk Wilhelmsen<sup>1</sup>

(1) Renaissance Computing Institute, University of North Carolina, Chapel Hill

### Abstract

RENCI is working with groups at the University of North Carolina at Chapel Hill and their collaborators in developing infrastructure to support projects that are leveraging next generation genomic sequencing technologies. These projects are generating large volumes of genetic data used for both research and clinical purposes. We are using the Enterprise version of the Integrated Rule-Oriented Data System (E-iRODS) to manage data collections and workflows that span multiple data centers. The paper describes are usage of E-iRODS for genomics, current issues, and future directions.

**Index Keyword Terms**— *iRODS, genomics, data management*

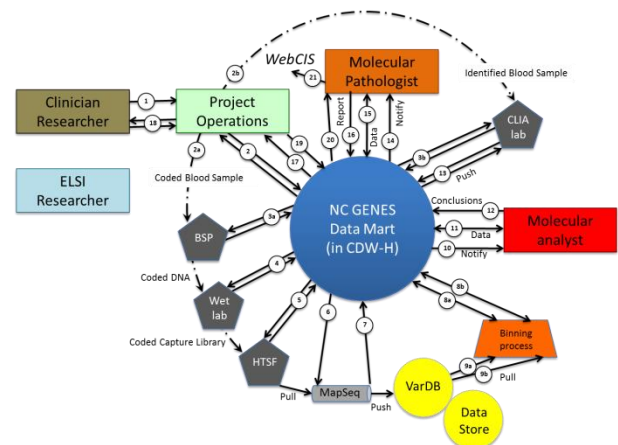
### 1. Introduction

Over the past three years RENCi has developed an informatics infrastructure for the management and analysis of human genomic data produced by next-generation sequencing technology. This infrastructure now supports the management of genomic data from several NIH sponsored research studies. These include principally a National Human Genome Research Institute study (project name NCGenes) aimed at evaluating the medical, ethical, legal, and societal implications of using high throughput genomic sequencing in clinical care and a National Institute on Drug Abuse sponsored study (project name NIDA) aimed at understanding the genetic factors associated with drug and alcohol dependencies. The infrastructure now includes next generation sequencing data from over a thousand human subjects sequenced at UNC for these projects plus genomic data on several thousand subjects pulled from other projects (including 1000 Genomes, NHLBI Exome Sequencing Project, The Cancer Genome Atlas, Complete Genomics public data sets). This paper reviews the architecture of the informatics infrastructure with a focus on our use of E-iRODS as a data management solution and the integration of E-iRODS with other data technologies.

### 2. Background

The developed infrastructure supports several purposes:

- Enabling research to uncover novel associations between genetic variants and clinical phenotypes;
- Enabling research to advance our understanding of the structure and evolution of clinically important variants and haplotypes;
- Research and development of new methodologies and tools to conduct genomic analysis;
- Providing genetic variant data on individual patients to clinicians for care decisions;
- Enabling research into better methods and approaches to present data on individual genetic variations for clinical utility.



**Figure 1: Informatics Workflow for the NCGenes Clinical Genomics Project**

Figure 1 presents an overview of the clinical workflow for the NCGenes Project. Starting from the upper left, the clinician researcher and subject determine if genomic sequencing is appropriate and desired. If so, multiple steps in a highly orchestrated workflow then occur as one progressively counter-clockwise around the figure. The major steps include: 1) blood processing and web lab work; 2) sequencing of the genome in the UNC high throughput sequencing facility (HTSF); 3) processing of the sequencing data through high performance analytical pipelines to reconstruct the

genome digitally and perform data cleaning (MapSeq); 4) detection of variants and linking of detected variants to databases of reported findings on variants (VarDB/Data Store); 5) assignment of variants into clinically-relevant categories based upon the medical phenotype of concern; 6) review of binned variants by clinical analysts; 7) validation of suspected variants in a CLIA-approved laboratory; and finally 8) review of validated variants by the clinical pathologists and clinicians to inform clinical care decisions.

Monitoring, reporting, and coordination of these stages is guided by a custom workflow environment that integrates with multiple laboratory information management systems (LIMS), analytical pipelines, web sites, and databases. The processing of a single patient through this entire workflow may take weeks, although the computational stages total no more than a few days.

The research infrastructure developed originally for the NIDA project overlaps with the clinical infrastructure in the initial stages (stage 1, 2, 3, and 4) although detailed sequencing procedures are significantly more complex to accommodate research goals of generating low coverage sequences with minimal variation in overall coverage across genomes.

A number of computationally intensive algorithms are run on populations of genomes in order to search for novel associations between variants and clinical phenotypes. These include algorithms to phase genotypes, to impute genotypes, and to discover long shared haplotypes. Such algorithms are typically run over hundreds to thousands of genomes and may have run times on the order of several weeks on high performance compute clusters. Ongoing research in our group continues to improve these algorithms and approaches.

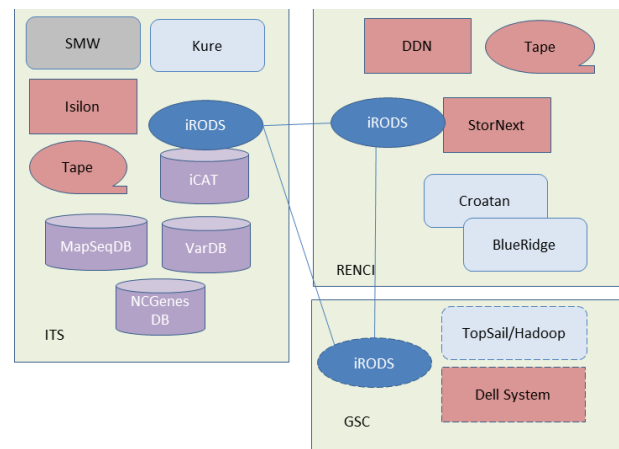
### 3. IT Infrastructure

Figure 2 depicts the IT infrastructure in use for the clinical and research genomics projects. Compute and data resources are shared with multiple other groups. This infrastructure includes:

- Kure: a 230 node HPC compute cluster in the UNC ITS Data Center (ITS) with 34 nodes dedicated to executing sequencing pipelines and smaller analysis jobs.
- BlueRidge: a 160 node HPC compute cluster in the RENCI Data Center (RENCI) for long running analysis.
- Croatan: a 30 core compute cluster at RENCI geared towards large data computations.
- TopSail: a 480 node compute cluster in the UNC Genomics Science Data Center (GSDC) being repurposed for data-intensive computing and Hadoop jobs
- A 1.7 Pb Isilon storage system in ITS

- A 1.7 Pb DDN storage system at RENCI
- 125 Tb of NAS storage at GDSC
- 1 Pb tape libraries at both ITS and RENCI
- A secured medical workspace (SMW) environment providing clinicians with access to patient genomes
- A database server, VarDB, which maintains information on known variants pulled from public data sources (e.g., NCBI, OMIM)
- A database server, NCGenes DB, for managing the clinical workflow
- A database server, MapSeq DB, for managing analytical pipeline processing and results

Data sets are generated on the Kure compute cluster and stored on the Isilon storage system. Archiving of primary data files (.fastq files) are done to tape at ITS. Data sets are copied to the RENCI, GSDC (under development), and partner sites for further analysis. The data center at RENCI provides an offsite backup of critical large data sets stored at ITS. A StorNext system has been introduced at RENCI to manage the automatic movement of data sets to tape.



**Figure 2: Primary IT resources located within three data centers. Red represent storage systems, light blue represents compute clusters, purple represents primary databases, and blue represents servers hosting E-iRODS. The GSC resources are installed but not configured. Networking between data centers varies from 1 to 10Gb/s and firewalls exist between all centers.**

### 4. Data Management with iRODS

The E-iRODS 3.0 Beta 1 release of iRODS was introduced early in 2012 to aid with management of the genomic data sets as the IT infrastructure needs began to grow beyond the systems deployed at ITS. Key challenges driving the decision to introduce iRODS were:

- The need to control access to and changes to data sets on a per project basis instead of on a per directory basis;
- The need to automate data management tasks, in particular making copies of data to other storage systems;
- A need for virtualization of data sets to allow production applications and users to query and access data sets independent of the physical location of the data.

In addition to these basic needs, we are interested in allowing for the following capabilities:

- Determining archival strategies based upon project and file-based metadata;
- Providing a common interface to access data, including data stored in HDFS which many biologists are unfamiliar with;
- Using iRODS to manage data workflow tasks ;
- Exploring iRODS as a means to manage our analytical workflows instead of the existing custom database.

We have incrementally rolled out iRODS into our informatics infrastructure. A key benefit of iRODS is the ability to overlay the technology on existing solutions, thereby minimizing the disruption to existing users and automated processes.

#### 4.1 Integration with Analysis Pipelines

RENCI and UNC have developed a pipeline technology termed MapSeq which is derived from the SeqWare pipeline system. MapSeq is a custom Java-based software solution built on top of the Pegasus and Condor job queuing systems. MapSeq has its own database, MapSeqDB for tracking workflow progress, status, and maintaining workflow products and metadata, and it integrates with Bio-specimen processing facility LIMS, the UNC's tissue tracking LIMS systems. This pipeline technology is now in use in most workflows run by the UNC High Throughput Sequencing Facility, including pipelines for whole genome, exomic genome, and RNA Seq processing.

MapSeq stores its data products on the Isilon system, so integration with iRODS can be easily done by registering pipeline-generated data sets into iRODS using the ireg command. Within iRODS, we organize data logically by /Project/Cohort/Sample\_Version where a sample folder contains all files generated by the pipeline. Each sample collection and files are tagged with common metadata attributes such as 'SampleName', 'ParticipantID', 'Run', 'Barcode', 'Pipeline'. Each file has a 'FileType' attribute and additional metadata attributes are dependent on the file type. For instance, FastQ files include metadata tags such as 'LibraryId' and 'LaneSampleId'.



In general, editing, moving, and deleting of pipeline generated data is not permitted by researchers and

versioning is used to deal with situations when pipeline results need to be regenerated. The benefit of this approach is that researchers can direct compute against data registered in iRODS without removing the data from iRODS, a requirement for HPC-based analysis.

#### 4.2 Managing data across data centers

As shown in figure 2, a primary requirement in our system is leveraging the computational and storage resources at the ITS, RENCi, and GDSC data centers. iRODS has been setup to allow for automation of data movements between sites and to allow users to access data independent of which site the data resides at. To this end, the iRODS genomic data grid includes an iRODS iCAT server in ITS, an iRODS resource server in ITS that mounts the Isilon storage system, an iRODS resource server at RENCi that mounts the StorNext storage appliance as a storage system, as well as various iRODS clients. Figure 3, a screenshot from the Scotty iRODS administration tools shows close to 2 Tb of genomic data has been registered to date, representing sequence data generated since the grid was deployed.

Total Zone (genomicsDataGridZone) Statistics

	5,068	Files in this Zone
	1,858,351,974,005	Total Size (bytes) of Files in this Zone
	317	Collections in this Zone
	5,068	Collections in this Zone containing Data Objects
	29	Files in this Zone that are in Trash
	5,825,616,455	Total Size (bytes) of Files in this Zone that are in Trash
	59	Collections in this Zone that are in Trash
	29	Collections in this Zone containing Data Objects that are in Trash

**Figure 3: Data grid statistics pulled from the prototype Scotty administration tool**

A copy of pipeline-produced data is automatically copied by iRODS to the storage system at RENCi. Once the GDSC is fully deployed, a similar copy of selected data files will be made to that system.

The Quantum StorNext appliance at RENCi provides capabilities for moving data to and from tape based upon policies, thus freeing up disk space. However, the governing policies can only be set on a directory basis. This approaches forces us to design directory structures to accommodate archiving policies, an undesirable situation in a research environment. As such, we are now investigating with Quantum using metadata attributes on iRODS collections and files to set policies for moving data files to tape.

#### 4.2 Interfacing with the Secure Medical Workspace

In order to support research on data sets containing protected health information (PHI), RENCi and the UNC Medical School have developed the Secure Medical Workspace (SMW). The SMW is an IT system that allows system administrators to provision a virtual

machine to researchers working with PHI. The virtual machine is configured to comply with UNC security policies. Of significance, the virtual machine also runs a Data Leakage Protection (DLP) client. The client intercepts any attempt to transfer data off of the client, e.g., by file copy, as text in an instant message, or even by pasting text into an HTML text box. The client will then either allow the transfer, challenge the transfer (e.g., by asking the user to confirm that they have permission to perform the transfer), or disallow the transfer. In all cases, the transfer of data is logged allowing for auditing of removed data by a security compliance officer. The DLP client interfaces with a DLP server that manages policy settings for the client and logs all client actions. The system allows the medical school to provision PHI data for research while ensuring the data is not transferred outside of UNC administrative control.

Clinicians are not given direct access to the genetic data on individual patients; rather they view the list of clinically relevant variants through an access controlled NCGenes web site. However, there are cases in which the clinician may wish to view the aligned sequence reads associated with a variant call. In this case, the sequence data is provisioned to the clinician within the SMW. Operationally, the clinician hits a button on the web site requesting the data. The web server then makes a query request to iRODS using the iRODS PHP client API (pRODS) including the region of interest and the patient id. A rule in iRODS then packages the BAM file and the BAM index file into a zip file, sends the zip file to the web client, and the client opens the bam file in the Integrated Genomic Viewer software which is resident on the SMW virtual machine.

#### **4.3 Integration with Hadoop**

The Hadoop system is a data storage and computation technology well suited for large scale data analysis tasks. We have been exploring the use of Hadoop to deal with analysis across large number of genomic samples. To this end, we have developed a set of tools termed HadoopVCF that allows one to perform Map-Reduce jobs in Hadoop over sets of variant calls (i.e., VCF formatted files) to perform common analysis tasks such as counting allele frequencies, percent missing, and Hardy Weinberg Equilibrium p-values.

While the use of Hadoop and the HDFS file system provides scalability in certain genomic computations, Hadoop is a system that most users of our systems are unfamiliar with. Simple tasks such as retrieving data or adding data require some amount of training to perform. As such, we have developed a Hadoop-iRODS driver that allows iRODS users to access data within the Hadoop system.

Usage requires creating an iRODS storage resource using the HDFS file driver available currently from RENC1. Once created, the standard iRODS `iput`, `iget`,

and `ils` commands can be used to put, get and list files within Hadoop. A future goal will be to extend this driver to allow for exposing Hadoop based computations, such as those provided by HadoopVCF, to users.

#### **4.4 Future: An iRODS Genomics Science Kit**

A future goal for our use of iRODS in genomic data management is to integrate common analysis tools with iRODS, thus allowing analysis to be conducted on data that may have been physically moved (e.g., to tape). The group at the Wellcome Trust Sanger Institute has already done this with Samtools, a popular tool for manipulating sam and bam formatted genomic data.

A second goal is to provide capabilities for developing analysis that runs across a set of files stored in iRODS. For instance, being able to compute the average coverage across all genomes associated with a study, independent of where the data sets are stored, would be beneficial. The simple approach to this is to write and deploy microservices for such analysis that call out to facilitating code libraries.

To these ends, the development of a genomics science kit that extends iRODS seems worth pursuing. Such a kit would include extensions to iRODS-enable common tools plus it would include libraries that one would one present on an iRODS grid that manages genomic data (such as the pysam and pybedtool libraries).

#### **Conclusions**

Our efforts to integrate iRODS into an existing genomics informatics infrastructure have been successful to-date, although not without some problems. The lack of documentation that describe typical deployment scenarios and best practices have been a weakness. iRODS brings significant benefit for the management of backup and archiving of data across data centers. The value to end users is not as obvious in our case as most users are involved in large-scale analysis that is currently done outside of iRODS. The development of a genomics specific science kit would help in this regard. However, we have been able to deploy iRODS without negatively impacting analysis work.

#### **Acknowledgement**

This work has been supported in part by RENC1 and NIH grants 5UL1RR025747 and DA030976-01. Both the E-iRODS team and the iRODS team and community have been very helpful including Leesa Brieger, Casey Averill, Terrell Russell, Jason Coposky, and Peter Clapham and Keith James from the Wellcome Trust Sanger Institute for providing us with the samtools extension.