Development of a high throughput gene, environment and epigenetics database and analysis system for international ALS research

@ ENCALS 2017
Alfredo Iacoangeli
Al-Chalabi Group (Clinical Neuroscience)
Dobson Group (Health Informatics)

alfredo.iacoangeli@kcl.ac.uk
Issues

• Big Data
  • how to store it
  • how to manage it
  • how to analyse it

• Collaboration
  • ownership heterogeneity
  • data sharing

• Audience
  • accessibility
  • impact
Data Management

Data Virtualisation Layer
Data Management
iRODS is an open source software for:

- Working with data distributed across storage technologies
- Annotating and searching data with rich metadata
- Implementing access control, auditing, preservation, organisation, and data movement policies
- Providing a single interface to share data between organisations
How would we interact with all of this?

- through any iRODS zone of the network
- with the terminal command line: icommands
- web-browser: search, download/upload, write rules, add metadata, more to come…
Web Browser

- Creation of new files
- File Editing
- Rule Execution
- Main Navigation Bar
- Metadata Search
- Recent Query Collections
- Docker deployment

Data Accessibility and Analysis

Virtualised Data
Data Accessibility and Analysis

Virtualised Data
Data Accessibility and Analysis

Virtualised Data
Data Accessibility and Analysis

Virtualised Data
Docker is an open-source project that automates the deployment of Linux applications inside software containers. Docker provides an additional layer of abstraction and automation of operating-system-level virtualisation on Linux. Docker can be integrated into various infrastructure tools, including Amazon Web Services, Google Cloud Platform, OpenStack Nova, etc. Containers running on a single machine share the same operating system kernel; they start instantly and use less RAM. Images are constructed from layered filesystems and share common files, making disk usage and image downloads much more efficient.
Install it:
```
~$ git clone https://github.com/KHP-Informatics/ngseasy.git
~$ cd ngseasy
~$ make INSTALLDIR="/media/Data" all
~$ sudo make install
```

Run it:
```
~$ cd /media/Data/ngs_projects/config_files
~$ ngseasy -c my_config.tsv -d /media/Data/ngs_projects
```

https://github.com/KHP-Informatics/ngseasy
What is on the plate

- An iRODS system tailored for ALS research needs

- Dockerized analysis pipelines

https://github.com/KHP-Informatics/MND-DataManagementAnalysis-System
• open-source, community-driven knowledge management platform for translational medicine

• organises clinical and research data on per patient base allowing:
  
  • Compare data from proteomics, metabolomics and other “omics” studies
  
  • Contrast patterns of gene expression in healthy and diseased individuals and in human tissue samples
  
  • Investigate correlations between genotype and phenotype in clinical trial data
  
  • Mine pre-clinical data for insights into the biology of human disease
  
  • Study genetic and environmental factors involved in human disease
  
  • Display data visually using a graphical interface
  
  • Stratify clinical data into molecular subtypes of a specific disease
  
  • Collaborate across academic, government and corporate research sectors
Ok! All good but who is going to make this happen? and how?

<table>
<thead>
<tr>
<th>SERVICE</th>
<th>Deployment time</th>
<th>Needed experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>iRODS local</td>
<td>hours</td>
<td>basic</td>
</tr>
<tr>
<td>Docker iRODS</td>
<td>minutes</td>
<td>basic</td>
</tr>
<tr>
<td>Docker</td>
<td>minutes</td>
<td>basic</td>
</tr>
<tr>
<td>Docker container (r)</td>
<td>minutes</td>
<td>basic</td>
</tr>
<tr>
<td>Docker image (r)</td>
<td>minutes</td>
<td>basic</td>
</tr>
<tr>
<td>iRODS rules (r)</td>
<td>hours</td>
<td>intermediate</td>
</tr>
<tr>
<td>iRODS rules (w)</td>
<td>hours</td>
<td>needs training</td>
</tr>
<tr>
<td>iRODS administration</td>
<td>minutes</td>
<td>needs training</td>
</tr>
<tr>
<td>Docker container (w)</td>
<td>hours</td>
<td>needs training</td>
</tr>
<tr>
<td>Docker image (w)</td>
<td>hours</td>
<td>needs training</td>
</tr>
</tbody>
</table>
Summary

A. Data management iRODS
   • data sharing
   • data accessibility
   • data curation
   • automatise workflow
   • exploit metadata potentialities

B. docker, standardised pipelines, etc
   • docker files
   • docker images
   • shared scrips
   • Github

C. Community driven project
   • tailor the system according to our needs
Future Plans

• TranSMART platform interactive analysis

• Develop portable analysis pipelines

• iRODS management system refinement

• User friendly interface
Clinical Neuroscience
- Ammar Al-Chalabi
- Ahmad Al Khleifat
- Aleksey Shatunov
- Anna Kulka
- Anand Pandit
- Ashley Jones
- Sarah Martin
- William Sproviero

Health Informatics
- Stephen J Newhouse
- Richard Dobson

All our collaborators!!!!!!!!!!
<table>
<thead>
<tr>
<th>Attribute</th>
<th>sample bam</th>
<th>sample vcf</th>
<th>sample XYZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>file type</td>
<td>bam</td>
<td>vcf</td>
<td>XYZ</td>
</tr>
<tr>
<td>lane</td>
<td>2</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>sample ID</td>
<td>LP00192</td>
<td>LP00192</td>
<td>LP00192</td>
</tr>
<tr>
<td>reference</td>
<td>hg19</td>
<td>hg19</td>
<td>—</td>
</tr>
<tr>
<td>study</td>
<td>MyFavStudy</td>
<td>YourFavStudy</td>
<td>TheirFavStudy</td>
</tr>
<tr>
<td>aligned</td>
<td>1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>alignment</td>
<td>MyFavPipeline</td>
<td>MyFavPipeline</td>
<td>—</td>
</tr>
<tr>
<td>Genotyping</td>
<td>—</td>
<td>MyFavPipeline</td>
<td>—</td>
</tr>
<tr>
<td>user defined attribute 1</td>
<td>—</td>
<td>—</td>
<td>yellow</td>
</tr>
<tr>
<td>user defined attribute 2</td>
<td>—</td>
<td>—</td>
<td>bitter</td>
</tr>
</tbody>
</table>
• **Background:**
We are now collecting huge amounts of multilayered genetic, epigenetic, environmental and clinical data, much of which is difficult to share and therefore to analyse.

• **Project aim:**
This project is a collaboration between ALS researchers and bio
Data Management and Analysis System

- Data Management
- Data Sharing
- Data Curation
- Data Analysis
- Analysis Reproducibility
- Data Visualisation
Data Flow

1 - Request (*iget*) for a data object.

2 - Server to server request for data object.

3 - Server to server request for data object.

4 - Parallel transfer of data object.

5 - Connection closed.
Ok! All good but who is going to make this happen and how?

**who?**
- on a central system -> King’s IT + the community
- on your institution facilities -> ask your admins
- on local servers -> we can make it together

**how?**
- manual deployment -> www.iRODS.com
- Docker/guided deployment -> our [GitHub](https://github.com/KHP-Informatics/MND-DataManagementAnalysis-System)