12 years of iRODS at Sanger: *What we've learned and what's next*

iRODS UGM Wellcome Sanger Institute
2021

Peter Clapham
Who are we?

Genome research Campus ~ 10 miles south of Cambridge (UK)
Established 1992
Part of the world wide human genome project

https://www.sanger.ac.uk/about/who-we-are/history-of-the-sanger-institute/
Science doesn’t stand still

Nature draft of the first human genome 2001:
https://www.nature.com/articles/35057062
Second generation sequencers.
Massive parallelisation meets sequencing.

https://www.enterprise.cam.ac.uk/case-studies/solexa-second-generation-genetic-sequencing/

Huge potential for increasing the speed, capacity and sequencing scale that could be delivered.

In 2015 it was possible to sequence a human genome for $1000 / sample in ~1 day vs 10 years and £10M on across a world wide consortium.

The Solexa Illumina sequencers were arriving!
One good idea, and a little innovation led to

https://www.ch.cam.ac.uk/collaboration-and-impact/solexa-sequencing
New tech drove “massive data growth”
New technology: new opportunities
Weekly calls with the Broad tech teams.

“So how many Sun Thumpers do you have next to the sequencers at the moment?”

We had a system in place, that worked, but it was handcrafted and scale was becoming an issue

How should we be searching for our data?

How should data be organised?

*What does good look like?*
What had we learned up to this point?

- Finding data can be hard
- Managing data can be hard
- We need something that will scale
- We need a tool/service that manages the data and does not become its own monster

KISS!
Think about a plan B
In early 2009 / 2010 not too many options available.

Dcache
SRM / SRB
General Atomics (Nirvana)
iRODS ?

Mostly out of the High energy Physics world.

Very helpful people !
Where’s my data ?!
Managed data delivery and curation.
Defined data workflows.
Strong iRODS rule engine.
Hardware agnostic.
Data lives on disk, meta-data in a database.
Checks for data status at rest.
No cul de sacs

NEW SEQ TECH INCOMING !
Version 1.0

Our first release was described here:

https://www.researchgate.net/publication/51637790_Implementing_a_genomic_data_management_system_using_iRODS_in_the_Wellcome_Trust_Sanger_Institute
Things then developed pretty fast..
Meta-data became really important

Structured data-curation

Example attribute fields →

Users query and access data largely from local compute clusters

Users access iRODS locally via the CLI

- 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
Today

Total sequencing yield by DNA pipelines
32.36 Peta-bases
Which in iRODS translates to

If 2021 had matched 2020, ~40% of the data in the archive would be < 2 years of age.
Automated data replication
Managed metadata
New tools to manage data ingress at scale, Baton, tears
BUT there are some challenges
Eventually scale hurts

- Meta-data searches become slow.
- An expectation for a strong GUI interface
- Ensuring rules are run effectively across all zones
- Managing many small systems is OK IF THEY ARE ALL THE SAME.
- Build systems change
- Upgrades and roll outs slow
- Risk that the data beast takes charge!

See John Constable's talk on tech debt later.
Did we keep it simple?
And the world is changing too

- **Politically**, UK no longer a part of Europe
- New **economic** opportunities, UKB
- **Social**, COVID has hit us all, vaccines, fake news, who do I / you believe?
- **Technology** bringing more opportunities daily, Oxford Nanopore? Real time sequencing?
- **Environmental** impacts, and dealing with compute at peta and exascale?

LEGAL!!
GDPR and Data governance Needs YOU!
Max fine limits for a GDPR infringement as set in 2018: £17.5m or 4% total annual global turnover (whichever is greater)

Who wants to be the first test case?

https://www.itgovernance.co.uk/dpa-and-gdpr-penalties
Data Management

- No longer just “Where’s my data ?!”
- Securing the data
- Data auditing
- Data classification
- Data encryption
- **Performance, can we continue to scale ?**
- Resilience
- Data linking
- **Cost**
More data-more data-lead science opportunities:
Bring in the microscopes!

Oh and our science is growing too

https://www.10xgenomics.com/spatial-transcriptomics
Spatial transcriptomics is potentially a huge area.

No longer “just” sequencing. More data types, greater insights and large scale projects:

[Link](https://www.sanger.ac.uk/collaboration/human-cell-atlas/)

---

Linking the data together!
And then..

World wide response to covide has been science lead.

Sequencing and informatics at the center of the fastest ever vaccine development!

https://www.cogconsortium.uk/genomics-in-a-pandemic-shedding-light-on-the-invisible/

Massive call to action.
Project organisation

- Governance & Advisory Group
- Steering Group
- Research Focus Areas
  - Mutational Analysis & Tracking
  - Sentinel Genomic Surveillance
  - Transmission & Outbreak Investigation
  - National Projects
- COG-UK Executive Director & Chair
- Management & Administration
  - Communications
  - Publications
  - Project Management
  - Internal Policies
- Operations
  - Sampling Centres
  - Sequencing Sites
  - Working Groups
  - Sampling Strategy
- Data
  - Data Flow

https://www.cogconsortium.uk/cog-uk/how-we-work/
You can view progress here:
### COVID-19 surveillance (as of 7 June 2021)

<table>
<thead>
<tr>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>323,765</td>
<td>Coronavirus genomes sequenced by the Sanger Institute</td>
</tr>
<tr>
<td>511,351</td>
<td>Coronavirus genomes sequenced by the COVID-19 Genomics UK (COG-UK) consortium</td>
</tr>
<tr>
<td>1,874,015</td>
<td>Coronavirus genomes sequences available worldwide</td>
</tr>
</tbody>
</table>
So what do we do next?

Can iRODS alone manage

- The data swell
- Meta-data performance
- Provide 2nd / 3rd gen GUI interfaces
- Provide audit logs and investigative tools
- Simplify customer engagement
- Reduce management churn
- Data security management

Perhaps include some sort of metrics?

*(managers like metrics)*
New data types and services are developing rapidly:

e.g.: 
Genestack 
Gen-3

- Would these fill the gap?
- What would this look like?
- Will there be best practices for use or is it a DiY suck and see?
- How would we manage interoperability long term?
- What does this mean for QA?
- What does this mean for our teams and developers?
Plugins are coming

For:
- Auditing
- Gen-3
- Indexing
- Improvements to the current GUI

But will they be a part of core?

*Can we bring new plugins up to speed more rapidly to meet demand while keeping the data-boat afloat?*
So what else have we learned?

Everything changes.
BUT...

- **Stay flexible.**
- Stable API’s are essential
- Strong auditing and security tools are the future
- Performance is important!
- Reliability, Resilience and Reproducibility are really really important!!
- Remain consistent and avoid surprises!
- *(sys-admins don’t like surprises, unless it’s a pay rise!)*
- A trusted partnership is essential!
Our original informatics testers!
Thomas Kean
Jim Stalker

NPG
David Jackson and his team past and present!

ISG
See John Constable’s presentation later.

Our scientists and informatics teams.

RENCI
Reagon Moore, Leesa! Terrell, Jason and of course Mike and team!

The community!