

Genedata Profiler & iRODS

An Open & Collaborative Enterprise Software
Platform for Patient and Compound Profiling

Marc Flesch, Tamas Rujan

Roots

Established in 1997 | Privately owned | Headquartered in Switzerland

Global Reach

~ 200 employees | Offices in Europe (Basel, Munich), North America (Boston, San Francisco) & Asia (Tokyo)

Dedicated to Drug Discovery & Biotechnology

Innovative portfolio of enterprise systems increasing productivity of data rich & complex research processes

Domain Expertise

Experienced Ph.D. level experts coupled with efficient software engineering processes

Marquee Customer Base

Leading pharmaceutical, biotechnology, and other life science organizations

Customer Base – Pharma



Supporting the Patient Profiling Process

Patient cohorts



NGS

ATCTCTTGGCTCCA
TCATTTAGAGGAAG
GAACGTCAAACCT
TGTTGCTTCGGCGG
GGCCTGCCGTGGCA
TCTCTTGGCTCCAG
CAGCATCGATGAAT
CGATACTTCTGAGT
CGGATCTCTTGGCT
ACAACGGATCTCTT
CGGATCTCTTGGCT
GATGAAGAACGCAG

Patient stratification
Drug response prediction



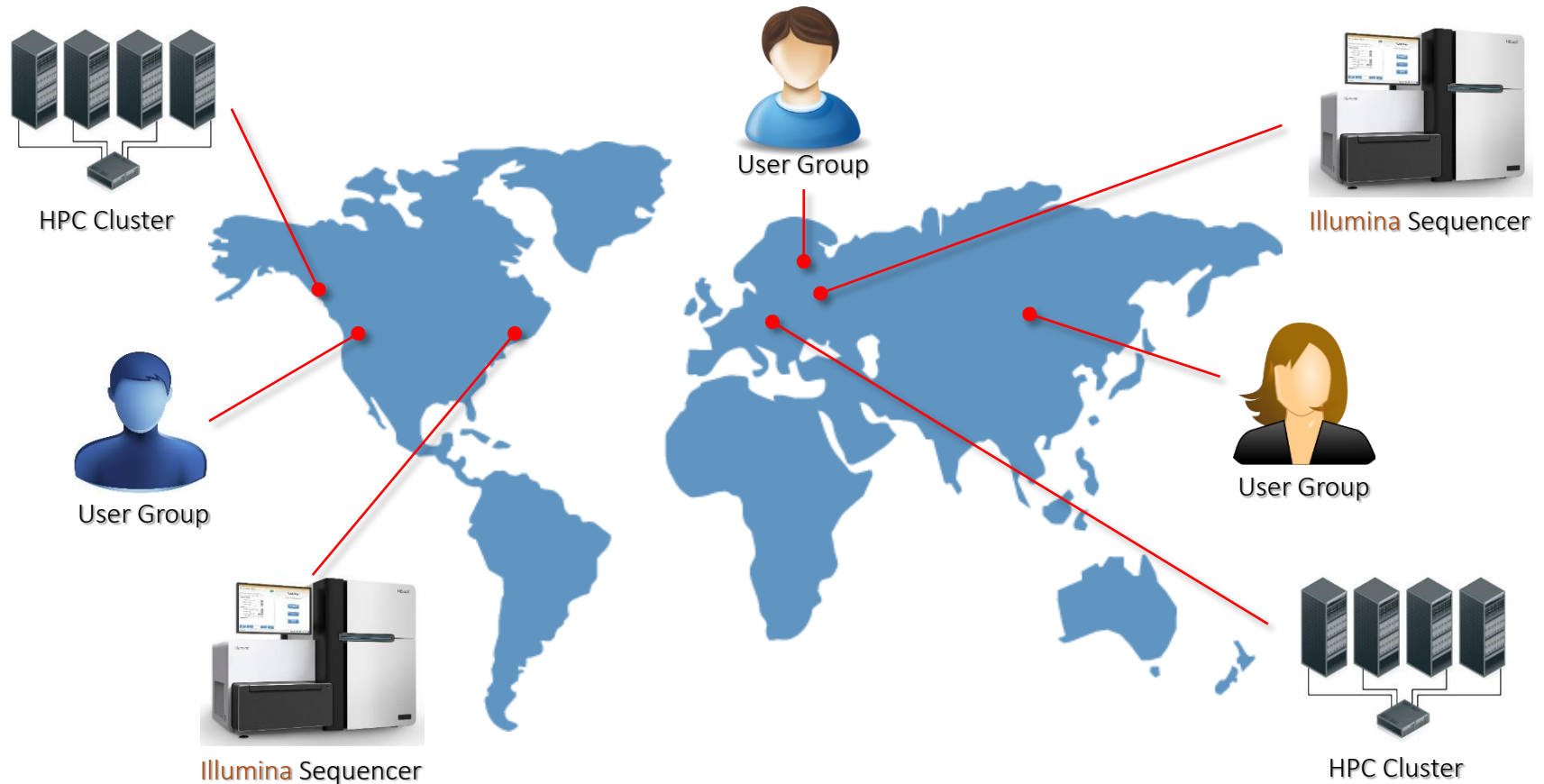
- Efficiently managing, processing, and analyzing data
 - Huge & complex datasets containing patient related omics data
 - Integrating disease & genomic information from different studies
- Facilitating collaboration within interdisciplinary teams
 - Enabling easy data, method & result sharing
 - Global distribution of data generators & data consumers
- **Working with data from human samples in research environments**
 - **Ensuring privacy of patient information**
 - **Maintaining chain of custody**



“Using data from clinical samples is challenging, because we need to take patient privacy very seriously”

*Henrik Seidel, Bayer

Data privacy within a global Organization



... how-to efficiently work with distributed data?

Common technologies applied include

- UNIX file permissions
- POSIX Access Control Lists (ACLs)
- CIFS Shares (SAMBA)

With the following shortcomings

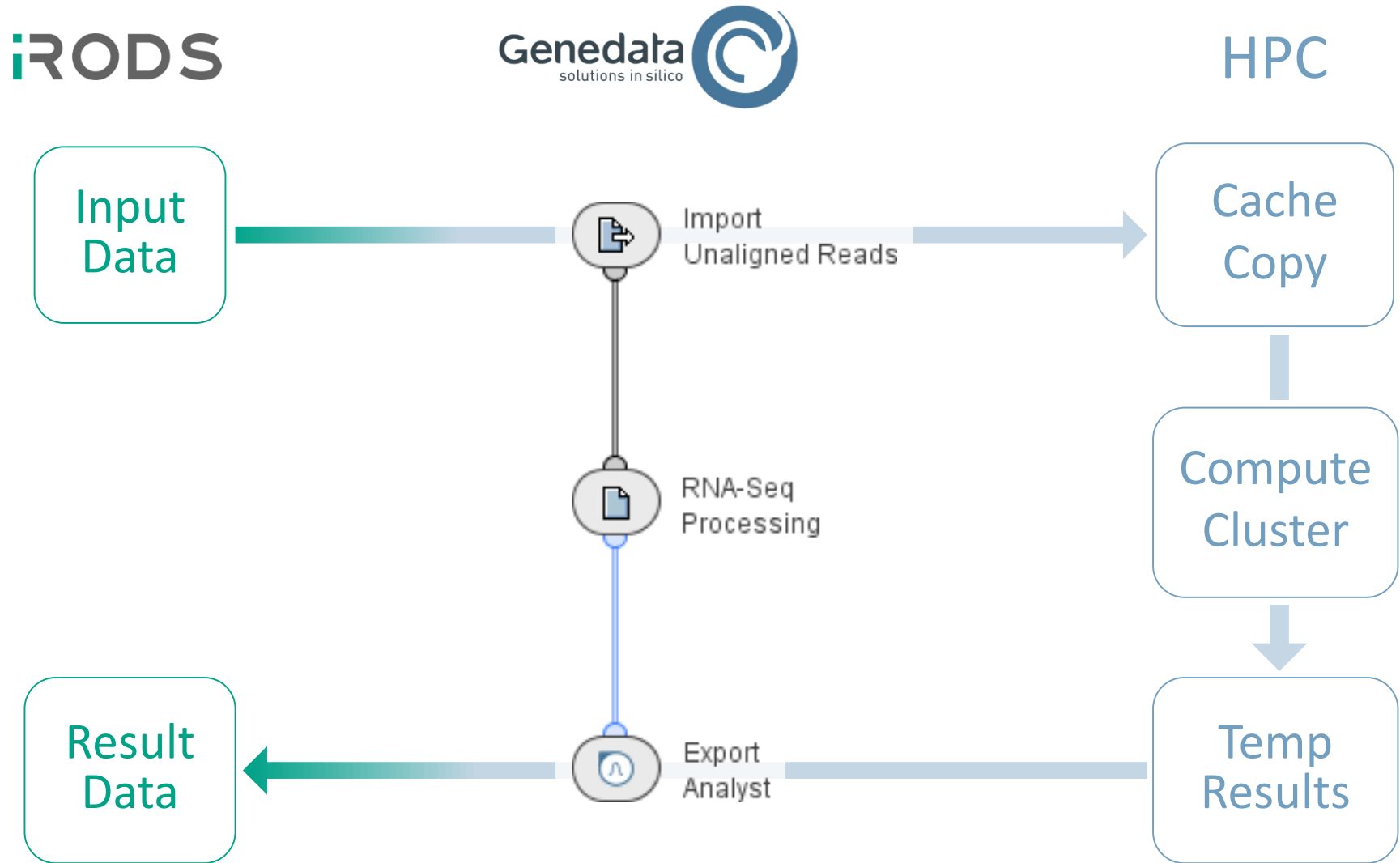
- UNIX permissions are too simple to model project centric access patterns
- paths on UNIX file systems can't replace data management systems
- permissions have to be maintained manually which is extremely cumbersome
- ACLs are hard to manage
- distributed storage problem stays unresolved

Pr

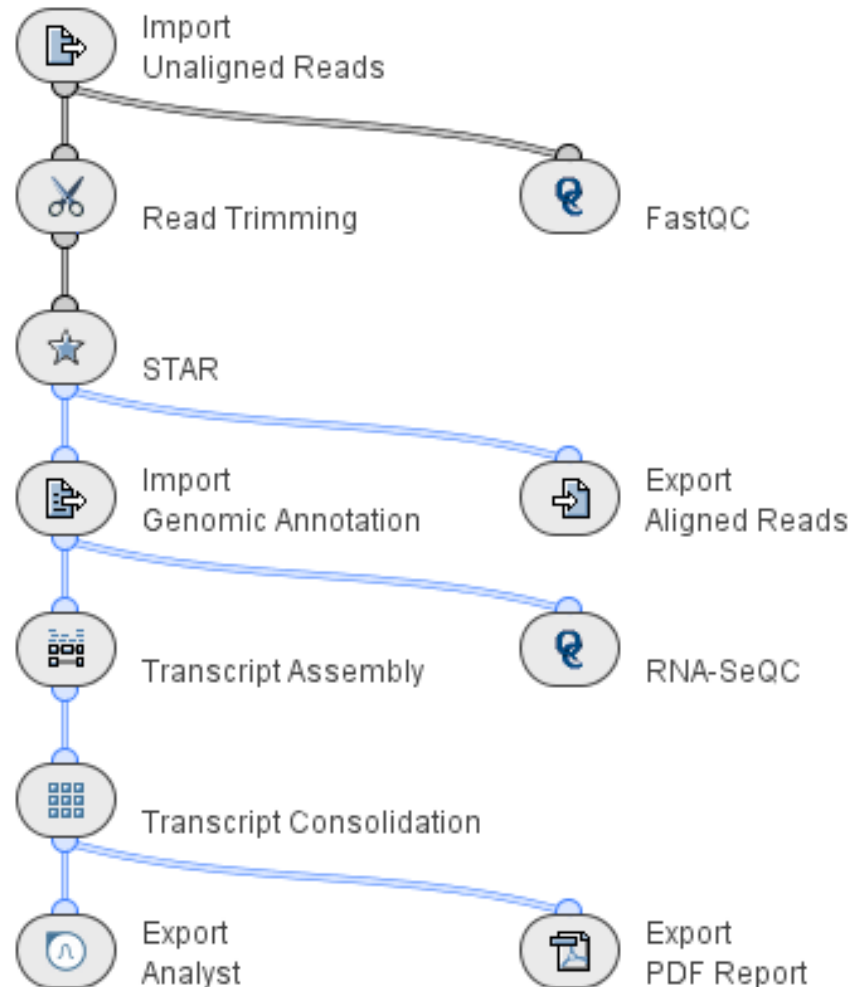


Our Solution

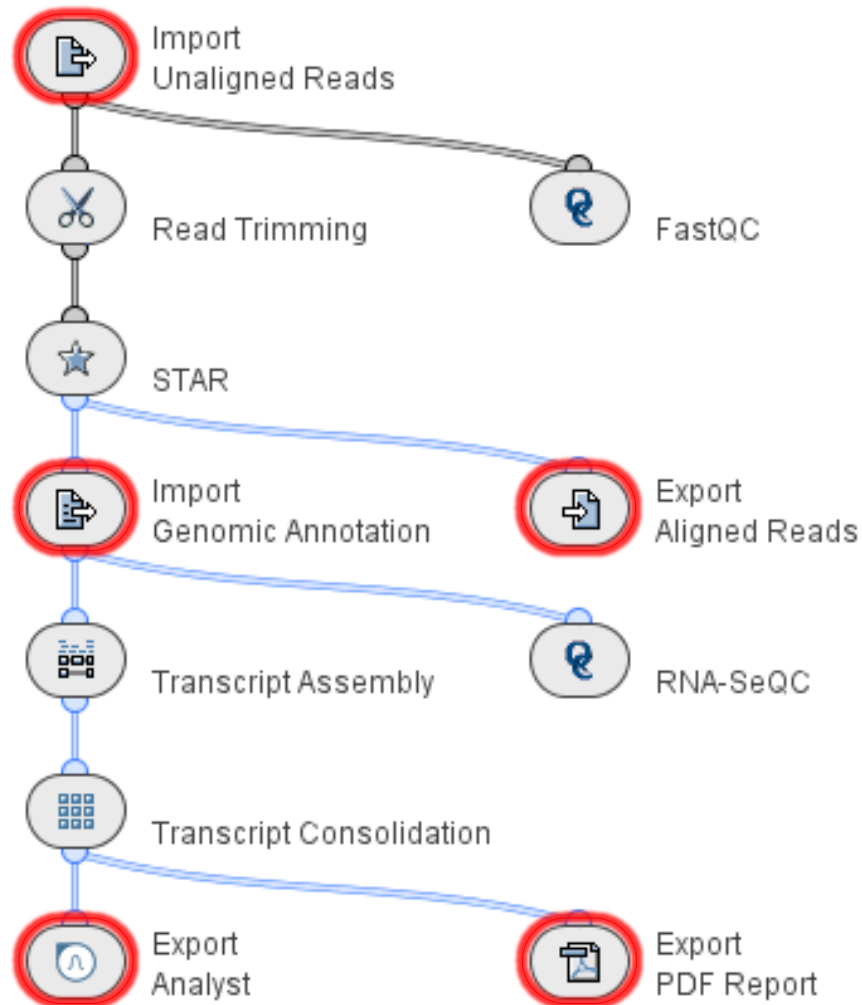
Marrying Security with Performance



RNA-Seq Data-Processing Pipeline



and Interaction Points with iRODS

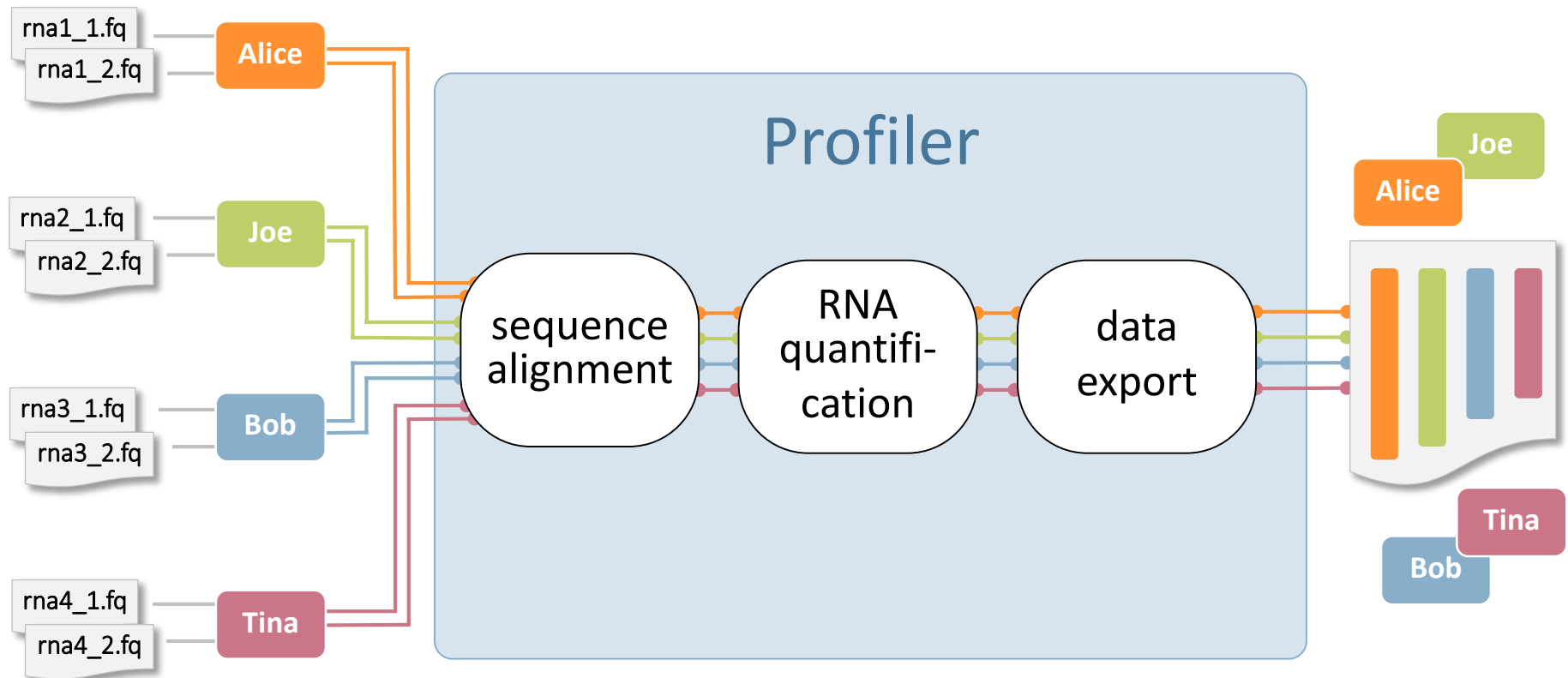


Chain-of-Custody

iRODS



iRODS



1. Visualization of clinical sample annotation together with corresponding raw data
2. Flexible search functionalities across the whole database
3. Powerful annotation curation capabilities including bulk editing and annotation information protection

Marrying Raw Data with Sample Annotation

Server File System

iRODS [homezone]/home/marc/demo_data

Name	Size	Last Modified	Type	Assembly	Condition	Control	Type	Factor
experiment_annotation.txt	305B	01/19/2015 10:13	TXT					
sample1.bam	1.5MB	01/19/2015 10:13	BAM		NoPi	sample5	ChIP	Pho4
sample1.bam.bai	1.2KB	01/19/2015 10:13	BAI		NoPi	sample5	ChIP	Pho4
sample2.bam	1.4MB	01/19/2015 10:13	BAM		NoPi	sample5	ChIP	Pho4
sample2.bam.bai	1.2KB	01/19/2015 10:13	BAI		NoPi	sample5	ChIP	Pho4
sample3.bam	1.7MB	01/19/2015 10:13	BAM		NoPi	sample6	ChIP	Pho4
sample3.bam.bai	1.2KB	01/19/2015 10:13	BAI		NoPi	sample6	ChIP	Pho4
sample4.bam	3.9MB	01/19/2015 10:13	BAM		HighPi	sample7	ChIP	WT
sample4.bam.bai	1.2KB	01/19/2015 10:13	BAI		HighPi	sample7	ChIP	WT
sample5.bam	9.9MB	01/19/2015 10:13	BAM		NoPi			
sample5.bam.bai	1.2KB	01/19/2015 10:13	BAI		NoPi			
sample6.bam	6.4MB	01/19/2015 10:13	BAM		HighPi			
sample6.bam.bai	1.2KB	01/19/2015 10:13	BAI		HighPi			
sample7.bam	9.2MB	01/19/2015 10:13	BAM		HighPi			
sample7.bam.bai	1.2KB	01/19/2015 10:13	BAI		HighPi			

Choose Metadata

Select the metadata you want to display for the items in this folder.

Filter

- ☒ Name
- ☒ Condition
- ☒ Control
- ☒ Type
- ☒ Factor
- ☐ method
- ☐ parent
- ☐ Concentration
- ☐ Coverage [Mode]
- ☐ Covered Bases [%]
- ☐ Duplicates Removed (%)
- ☐ Experiment
- ☐ Mapped Read Count
- ☐ Mapped Reads [%]
- ☐ Mapping Quality [Mode]
- ☐ MQ0 [%]
- ☐ Paired Sequenced
- ☐ Primary Aligned
- ☐ Read Length [Mode]
- ☐ Total Read Count

Raw Data

Sample Annotation

Providing 'Google-Like' Search

complex search

The screenshot displays the iRODS Server File System interface. At the top, there's a navigation bar with icons for back, forward, refresh, and upload/download actions. Below this, a search bar contains the query: `Not control=* factor=Pho condition=HighPi`. The search results are shown in a table below the search bar, listing files in the `iRODS [homezone]/home/marc/demo_data` directory. The table has columns for Name, Size, Last Modified, Relative Path, Type, Assembly, Condition, and Control. Two files are listed: `sample6.bam` (6.4MB, 01/19/2015 10:13, BAM, HighPi) and `sample6.bam.bai` (1.2KB, 01/19/2015 10:13, BAI, HighPi).

Server File System

← → ↻ ↻ Upload Files Download Files

⊗ Not control=* factor=Pho condition=HighPi

iRODS [homezone]/home/marc/demo_data

Name	Size	Last Modified	Relative Path	Type	Assembly	Condition	Control
sample6.bam	6.4MB	01/19/2015 10:13		BAM		HighPi	
sample6.bam.bai	1.2KB	01/19/2015 10:13		BAI		HighPi	

search result

Sample Annotation Curation

The screenshot displays the Genedata Sample Annotation Curation interface. The main window shows a list of files in the 'Server File System' (iRODS [homezone]/home/marc). A context menu is open over the file list, with the 'Edit Metadata' option highlighted. The 'Metadata Editor' dialog is also open, showing a list of attributes and their values. The 'Concentration' attribute is highlighted, and its values are listed in a table. The 'parent' attribute is highlighted with a lock icon, indicating it is locked down. The 'Browse sequence' button is also highlighted.

Server File System

Upload Files Download Files sample

iRODS [homezone]/home/marc

Name	Size	Last Modified	Relative Path	Type	Assembly	Condition	Control	Type	Factor
sample2.bed	554B	04/21/2015 14:22	ChIPseq (Results)/beds(1)	BED		NoPi	sample5	ChIP	Pho4
sample4.bed	697B	04/21/2015 14:22	ChIPseq (Results)/beds(1)	BED					
sample6.bed	122B	04/21/2015 14:22	ChIPseq (Results)/beds(1)	BED					
sample1.bed	557B	04/21/2015 14:23	ChIPseq (Results)/beds(2)	BED					
sample2.bed	554B	04/21/2015 14:23	ChIPseq (Results)/beds(2)	BED					
sample3.bed	316B	04/21/2015 14:23	ChIPseq (Results)/beds(2)	BED					
sample4.bed	697B	04/21/2015 14:23	ChIPseq (Results)/beds(2)	BED					
sample5.bed	122B	04/21/2015 14:23	ChIPseq (Results)/beds(2)	BED					
sample6.bed	122B	04/21/2015 14:23	ChIPseq (Results)/beds(2)	BED					
sample7.bed	122B	04/21/2015 14:23	ChIPseq (Results)/beds(2)	BED					
sample1.bam	1.5MB	01/19/2015 10:13	demo_data	BAM					
sample1.bam.bai	1.2KB	01/19/2015 10:13	demo_data	BAI					
sample2.bam	1.4MB	01/19/2015 10:13	demo_data	BAM					
sample2.bam.bai	1.2KB	01/19/2015 10:13	demo_data	BAI					
sample3.bam	1.7MB	01/19/2015 10:13	demo_data	BAM					
sample3.bam.bai	1.2KB	01/19/2015 10:13	demo_data	BAI					
sample4.bam	3.9MB	01/19/2015 10:13	demo_data	BAM					
sample4.bam.bai	1.2KB	01/19/2015 10:13	demo_data	BAI					
sample5.bam	9.9MB	01/19/2015 10:13	demo_data	BAM					
sample5.bam.bai	1.2KB	01/19/2015 10:13	demo_data	BAI					
sample6.bam	6.4MB	01/19/2015 10:13	demo_data	BAM					
sample6.bam.bai	1.2KB	01/19/2015 10:13	demo_data	BAI					
sample7.bam	9.2MB	01/19/2015 10:13	demo_data	BAM					
sample7.bam.bai	1.2KB	01/19/2015 10:13	demo_data	BAI					

39 out of 39 selected.

Find... Ctrl+F
Select All Ctrl+A
Invert Selection
Clear Selection
Open
New Folder
Cut Ctrl+X
Copy Ctrl+C
Paste Ctrl+V
Delete
Rename F2
Open location
Edit Metadata

Metadata Editor - 39 files/folders selected

Attribute	Value	Unit
Concentration	12	μmol/ml
Condition	15	μmol/ml
Control	18	μmol/ml
Coverage [Mode]	20	μmol/ml
Covered Bases [%]	22	μmol/ml
Duplicates Removed [%]		
Experiment		
Factor		
Mapped Read Count		
Mapped Reads [%]		
Mapping Quality [Mode]		
method		
MQ0 [%]		
Multiline		
Paired Sequenced		
parent		
patient_id		
Primary Aligned		
Proper Paired		
Proper Paired [%]		
Read Length [Mode]		
study_id		
Total Read Count		
Type		
Unmapped Read Count		

multiple values including units

locked down attribute

browse sequence

- The smooth integration of **Genedata Profiler** with **iRODS** enables scientists to **preserve their research eco-system** when working with confidential data
- **Genedata Profiler's** data processing and management capabilities together with **iRODS'** metadata and security concepts are a unique combination to establish the **chain-of-custody** for analyzing personalized medicine data