2022-07-06

Managing high-throughput sequencing and other -omics data with RODEOS and rodeos-ingest

Clemens Messerschmidt, Marten Jäger, Mathias Kuhring, Dieter Beule and Manuel Holtgrewe

BIH Berlin Institute of Health
Charité & MDC

What are -omics data?

Genomics

Transcriptomics

Proteomics

Phosphoproteomics

Metabolomics

…

High-throughput sequencing

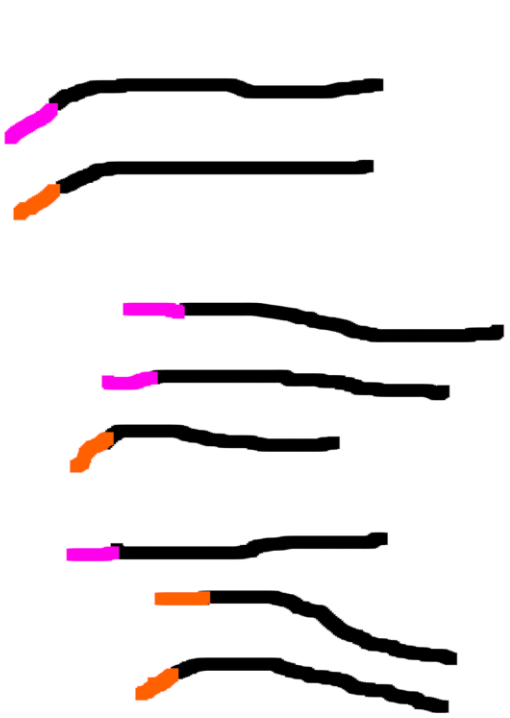
Sequencing on Illumina machines

- Running times: 4 – 60 hours
- Millions to billions of "reads"
- Few GB to TB sized output data

Images: www.illumina.com/systems/sequencing-platforms.html

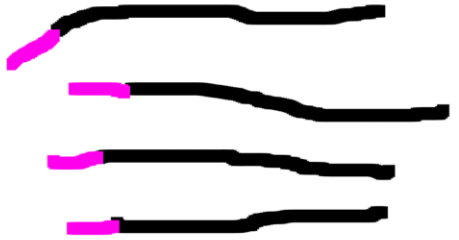
Multiplexing and Demultiplexing

Raw data: flowcell / BCL

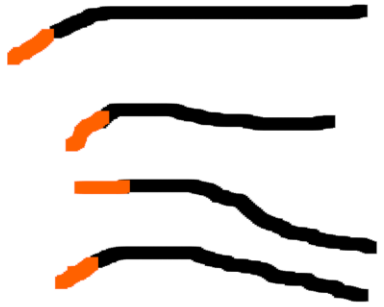


Demultiplexing

(per sample) FASTQ files

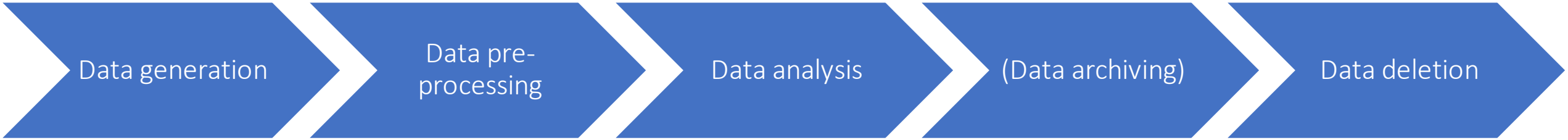


SampleA_ProjectY.fastq.gz



Sample1_ProjectM.fastq.gz

The life cycle of data

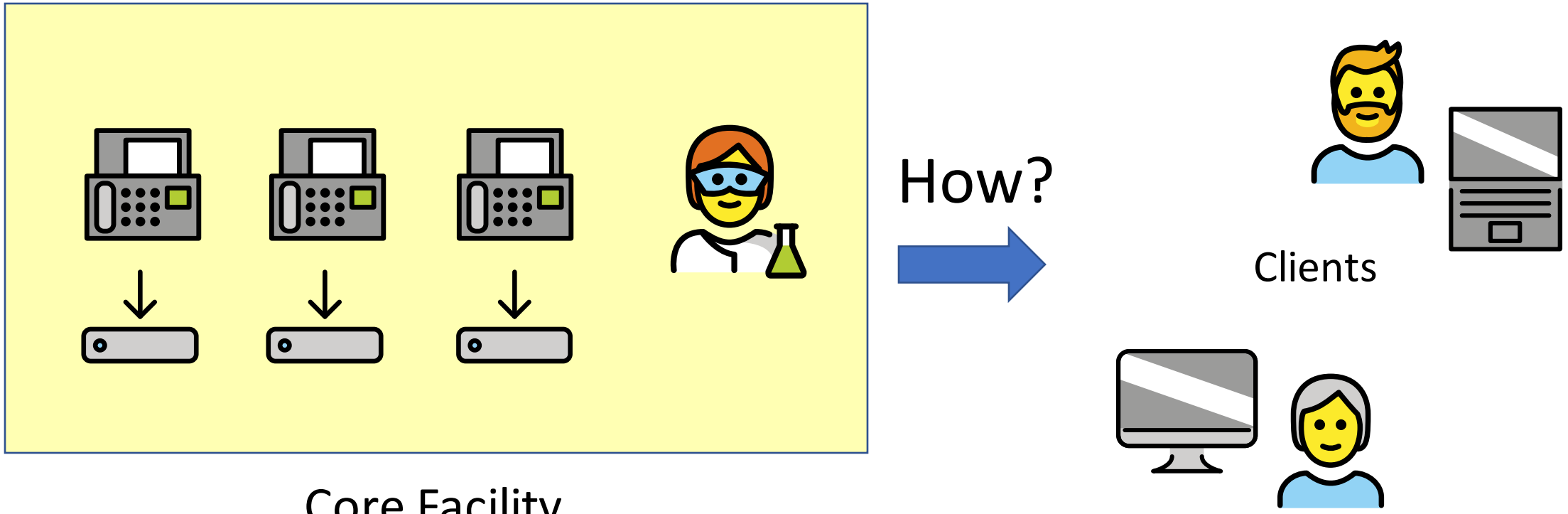


The life cycle of –omics data

Data generation → Data pre-processing → Data analysis → (Data archiving) → Data deletion

Demultiplexing and QC

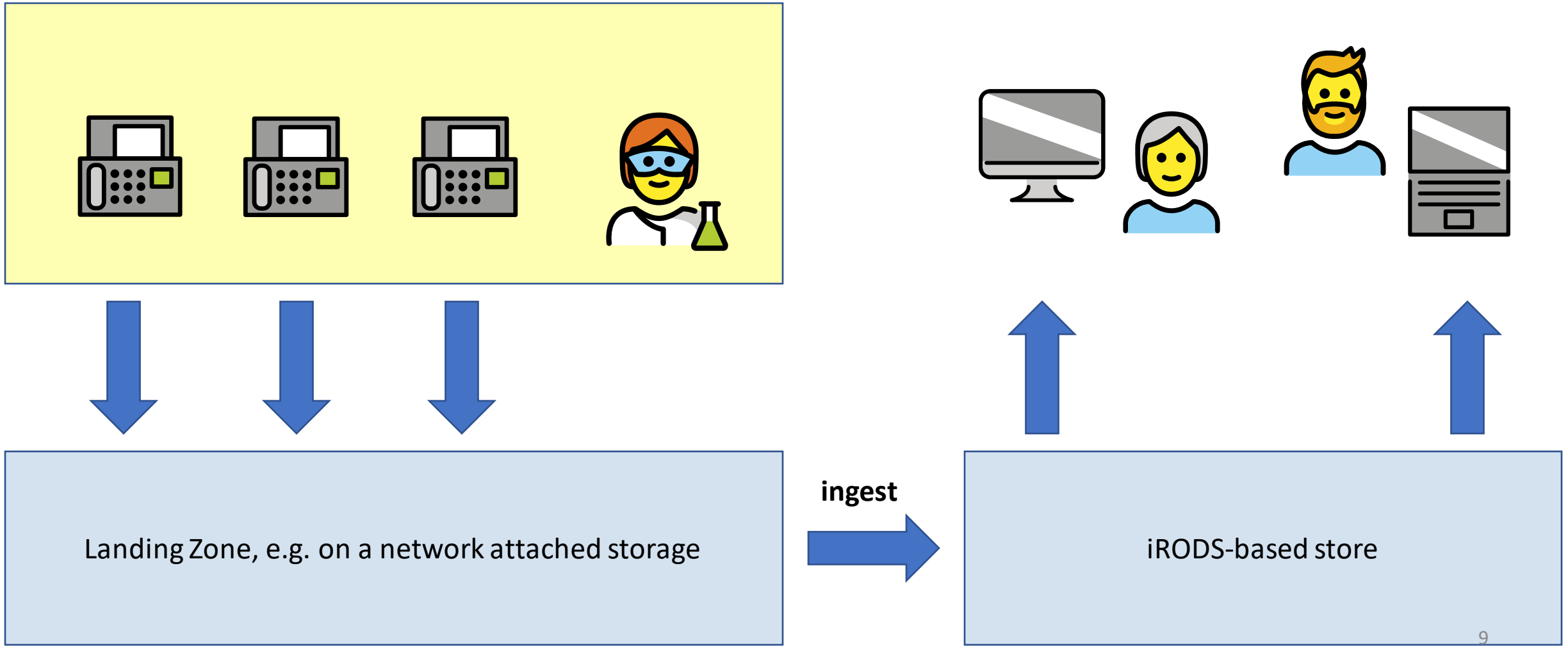
Data needs to reach the client's computer



An iRODS-based solution unlocks powerful tools for data management

- Automated ingest of newly generated data
  - Annotate collections with extracted metadata (per experiment)
- Secure collaboration
  - Data privacy considerations: human genomic data!
  - Logging access and auditing
  - Cross-institutional authentication possible
- Rule engine for archiving and deletion
- Access to ecosystem: WebDAV, Metalnx, icommands, S3, APIs, ...

RODEOS - Raw Omics Data accEss and Organization System

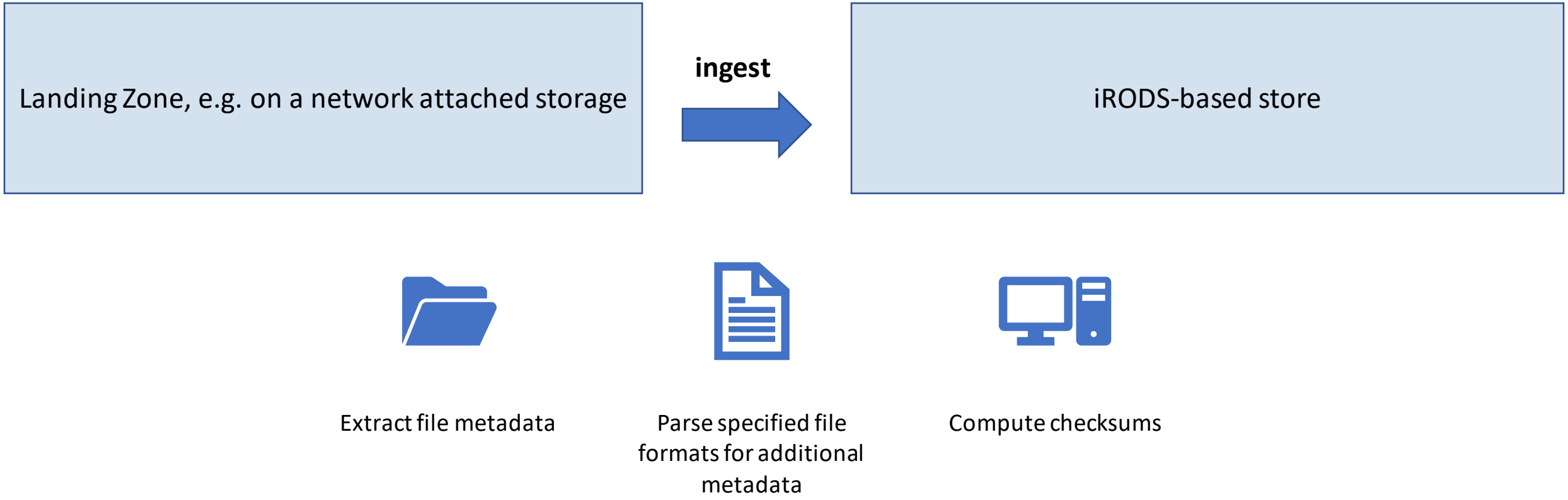


Minimal setup

- iRODS "cluster"
- Landing zone
  - Network-attached storage
- Host running the ingest service
  - Read access to landing zone
  - Write access to iRODS
- Network!

https://github.com/bihealth/rodeos-docker-compose

iRODS ingest in general



https://github.com/irods/irods_capability_automated_ingest

rodeos-ingest is an extension to irods-capabilities-automated-ingest

rodeos_ingest.genomics.illumina.bcl

- Ingest a directory that contains raw Illumina flowcell data into irods collection
  - Parse Illumina metadata files and annotate collection with these attributes

rodeos_ingest.genomics.illumina.fastq

- Ingest a directory of .fastq(.gz) files (after demultiplexing) into an irods collection
- Not implemented yet:
  - Parse accompanying QC report(s) and sequencing statistics

https://github.com/bihealth/rodeos-ingest

Illumina sequencing metadata

RunInfo.xml

RunParameters.xml

```
<FlowcellRFIDTag>
  <SerialNumber>000000000-K6R67</SerialNumber>
  <PartNumber>15028382</PartNumber>
  <ExpirationDate>2022-10-25T00:00:00</ExpirationDate>
  <LotNumber>20592330</LotNumber>
</FlowcellRFIDTag>
<PR2BottleRFIDTag>
  <SerialNumber>MS3156853-00PR2</SerialNumber>
  <PartNumber>15041807</PartNumber>
  <ExpirationDate>2022-11-10T00:00:00</ExpirationDate>
  <LotNumber>20596808</LotNumber>
</PR2BottleRFIDTag>
<ReagentKitRFIDTag>
  <SerialNumber>MS3154893-600V3</SerialNumber>
  <PartNumber>15043962</PartNumber>
  <ExpirationDate>2022-10-18T00:00:00</ExpirationDate>
  <LotNumber>20598840</LotNumber>
</ReagentKitRFIDTag>
<Resumable>true</Resumable>
<ManifestFiles />
<AfterRunWashMethod>Post-Run Wash</AfterRunWashMethod>
<Setup>
  <SupportMultipleSurfacesInUI>true</SupportMultipleSurfacesInUI>
  <ApplicationVersion>4.0.0.1769</ApplicationVersion>
  <NumTilesPerSwath>19</NumTilesPerSwath>
  <NumSwaths>1</NumSwaths>
  <NumLanes>1</NumLanes>
  <ApplicationName>MiSeq Control Software</ApplicationName>
</Setup>
<RunID>220211_M06205_0040_000000000-K6R67</RunID>
<ScannerID>M06205</ScannerID>
<RunNumber>40</RunNumber>
<FPGAVersion>9.5.12</FPGAVersion>
<MCSVersion>4.0.0.1769</MCSVersion>
<RTAVersion>1.18.54.4</RTAVersion>
<Barcode>000000000-K6R67</Barcode>
```

Illumina sequencing metadata: annotated iRODS collections

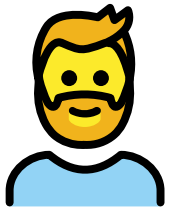
```
$ imeta ls -C /tempZone/home/rods/RINGEST_TEST/2022/220512_M06205_0046_000000000-DFVLL
AVUs defined for collection /tempZone/home/rods/RINGEST_TEST/2022/220512_M06205_0046_00000
attribute: rodeos::ingest::run_info::date
value: 220512
units:
----
attribute: rodeos::ingest::run_info::flowcell
value: 000000000-DFVLL
units:
----
attribute: rodeos::ingest::run_info::instrument
value: M06205
units:
----
attribute: rodeos::ingest::run_info::run_id
value: 220512_M06205_0046_000000000-DFVLL
units:
----
attribute: rodeos::ingest::run_info::run_number
value: 46
units:
----
attribute: rodeos::ingest::run_parameters::application_name
value: MiSeq Control Software
units:
----
attribute: rodeos::ingest::run_parameters::application_version
value: 4.0.0.1769
units:
----
attribute: rodeos::ingest::run_parameters::barcode
value: 000000000-DFVLL
units:
----
attribute: rodeos::ingest::run_parameters::flowcell_rfid_tag::expiration_date
value: 2022-11-16T00:00:00
units:
----
```

Interfaces to access data in RODEOS

MetaInx / ZMT

- Access for technical staff
- Manage client access based on users/groups per irods collection

- icommands
  - Command line tools for power users
  - irsync to download data to near-compute storage



- WebDAV
  - Accessible overview for everyone

rods

Resources
Rules
Users
Groups
Collections
Search
Templates
Shared Links
Favorites
Public
Trash

» tempZone » home » rods » RINGEST_TEST » 2022

Action

220512_M06205_0046_000000000-DFVLL

text/directory

Details | Metadata | Permissions

Metadata

CSV | +Metadata

Delete selected

Search...

Showing 1 to 29 of 29 entries

	Attribute	Value		Actions
☐	rodeos::ingest::run_info::date	220512	View	Edit Delete
☐	rodeos::ingest::run_info::flowc...	000000000-DFVLL	View	Edit Delete
☐	rodeos::ingest::run_info::instr...	M06205	View	Edit Delete
☐	rodeos::ingest::run_info::run_id	220512_M06205_0046_0000...	View	Edit Delete

Action

220512_M06205_0046_000000000-DFVLL

text/directory

Details | Metadata | Permissions

Permissions

+ Permissions

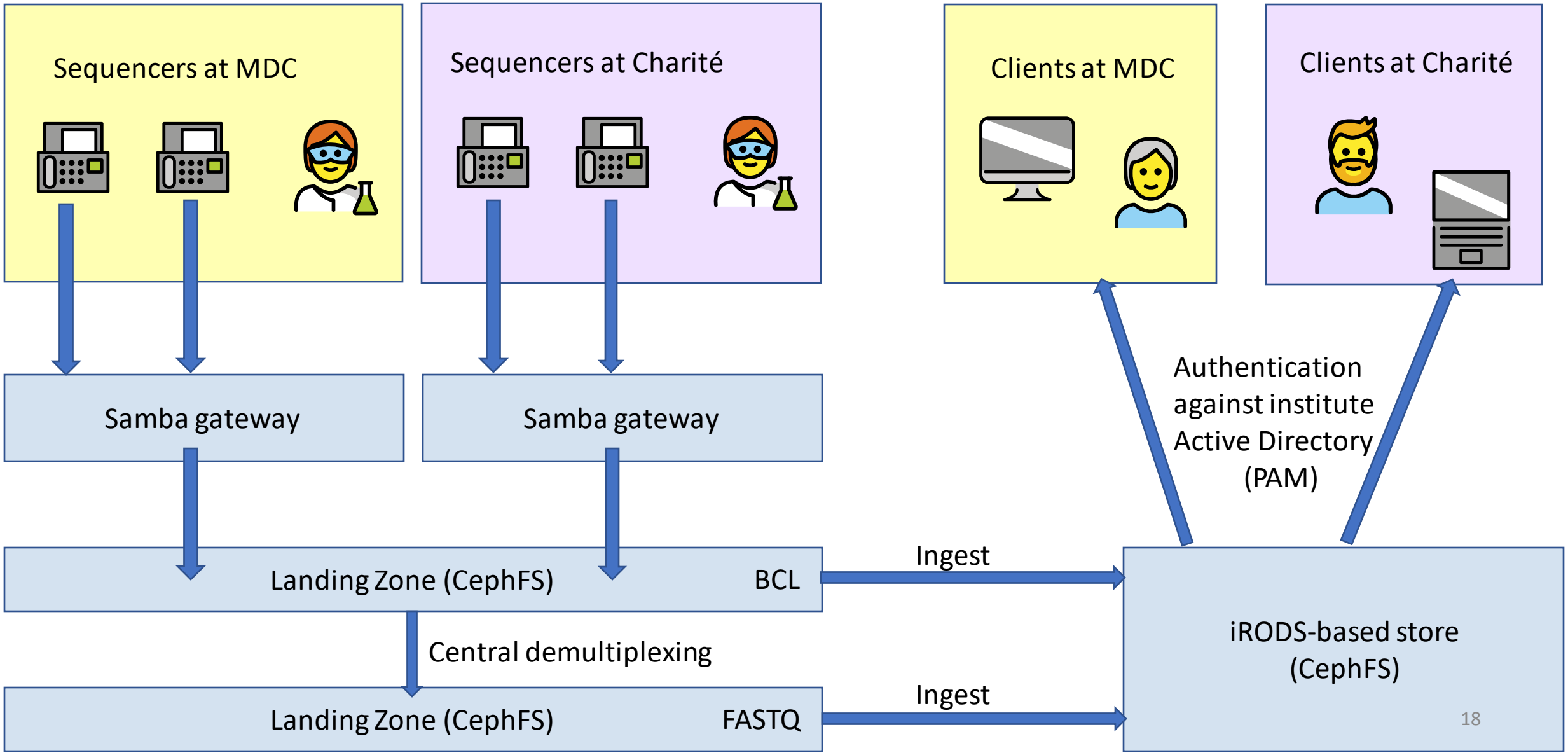
10

Showing 1 to 2 of 2 entries

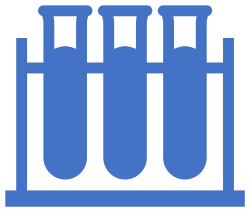
User/Group	iRODS System Role	Permission	Shared Link
raw-data-bih-genomics	group	READ	☐
rods	rodsadmin	OWN	☐

1

Pilot implementation at Berlin Institute of Health



Open questions and future work



Ingest of mass spectronomy data

Needed for metabolomics, proteomics experiments

Metadata formats are not well defined, competing vendors

Tracking sample information for multiplexed runs

Sample sheets could be tracked, but most places have solutions in place already

(LIMS, https://github.com/bihealth/digestiflow-server)

Archival and deletion are not implemented yet

Data life cycle needs to be properly defined

There might be multiple different ones!

- RODEOS is an iRODS-based system to manage and distribute Illumina sequencing and other -omics data
- rodeos-ingest parses metadata from Illumina metadata files

Try it out and let us know what you think:

https://github.com/bihealth/rodeos-docker-compose

https://github.com/bihealth/rodeos-ingest