



National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

iRODS-based system turbocharged next-gen sequencing analysis during pandemic and beyond



RIVM: the Dutch National Institute for Public Health and the Environment

- Living in a safe, clean, healthy environment
- Preventing and controlling infectious diseases
- Good healthcare and a healthy lifestyle





Bioinformatics and Computational Services

*Bioinformatics, as related to genetics and genomics, is a scientific subdiscipline that involves using computer technology to **collect**, **store**, **analyze** and **disseminate** biological data and information, such as DNA and amino acid sequences or annotations about those sequences.*

Research environment for scientific employees/researchers and the support of the environment.

Data management system for unstructured data.



Content

- History, why did we start?
- Challenges, bioinformatics environment.
- (some) Products: Intorods, Jobengine, BioRODS, Archiving SURF.
- 2021: Amount of analyses of the previous year in half a week.
- Success stories.
- Future plans.



History, why did we start?

Bioinformatics is used in more and more places, also at the National Institute for Public Health and the Environment (RIVM).

Examples:

- the spread of pathogens
- (genetic) population research
- effects of lifestyle choices
- the origin of chronic diseases
- consequences of changes in the living environment



Challenges, bioinformatics platform

Challenges:

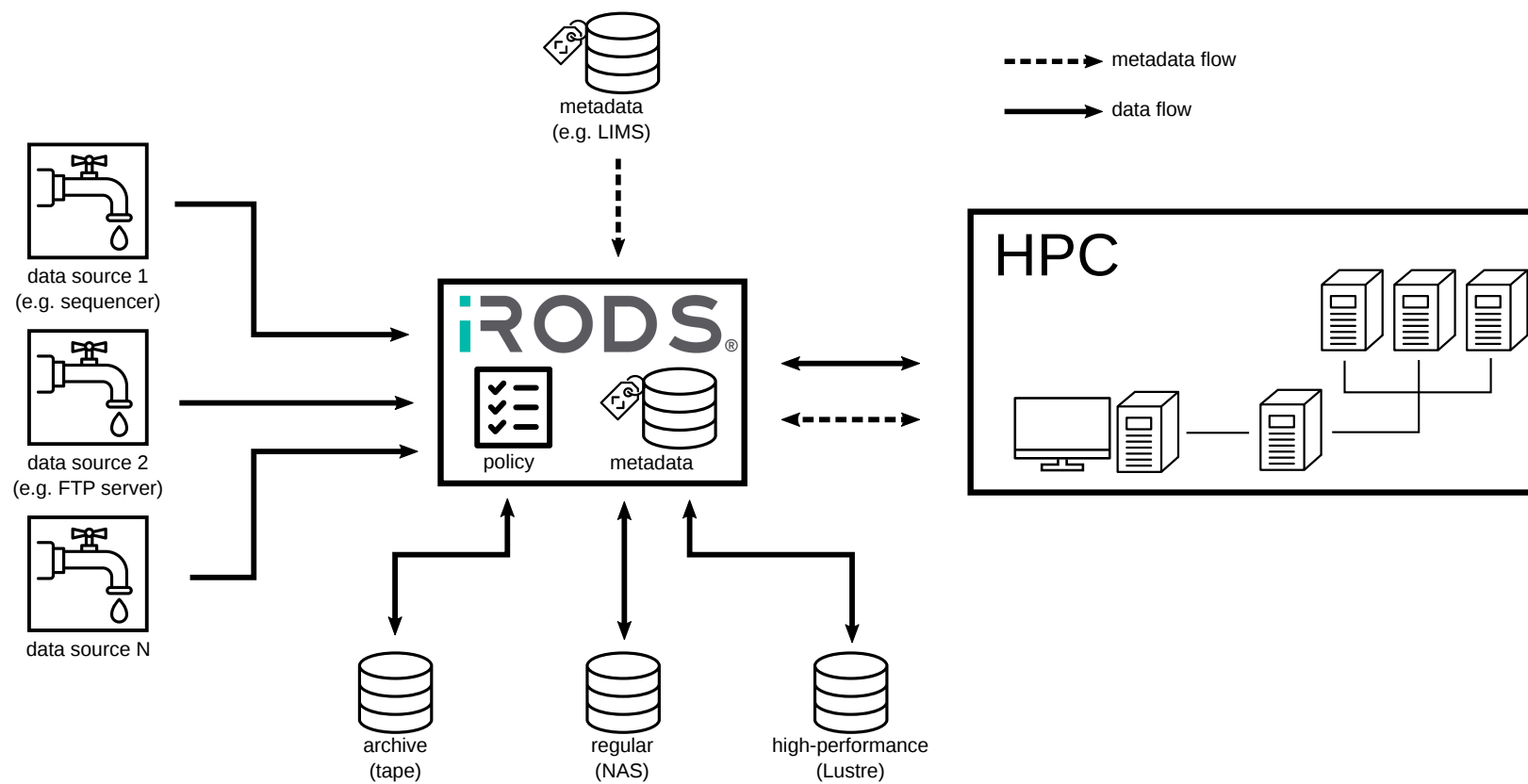
- Large amount of data, data should be FAIR.
- Lots of computational resources necessary to analyze the data.
- Reproducible data analyses.
- A platform where people can develop and share pipelines.

Bioinformatics platform

1. Shared Linux environment, cooperation for creating bioinformatics pipelines.
2. High performance computing (HPC) cluster, fast Lustre storage.
3. Data management system (DMS) for organizing the data flow.
4. Process automation tool to automatically start pipelines.
5. Courses (git, Linux, HPC, Snakemake) for using the environment.

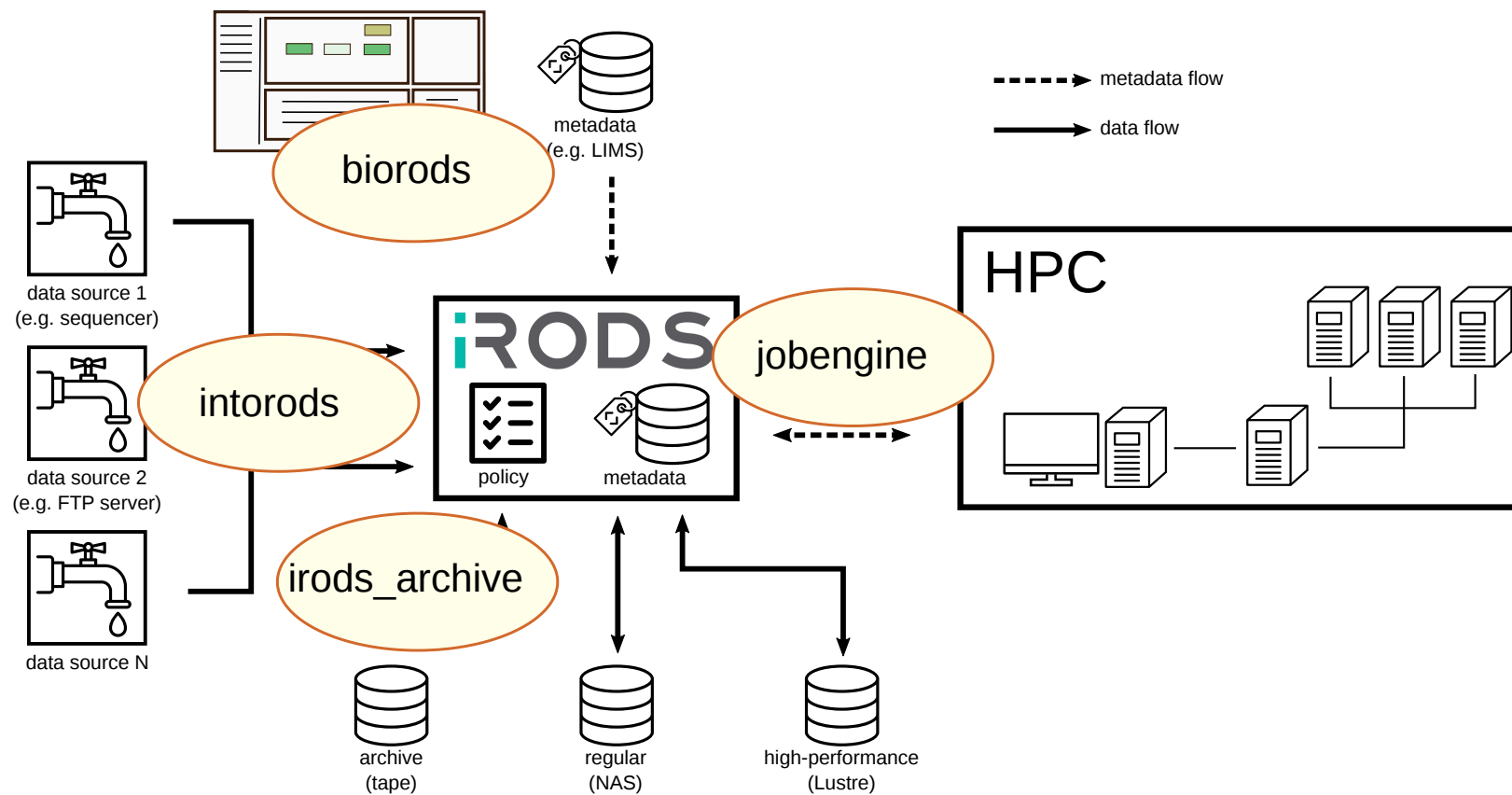


RIVM iRODS architecture





In-house development





Intorods: iRODS data import tool

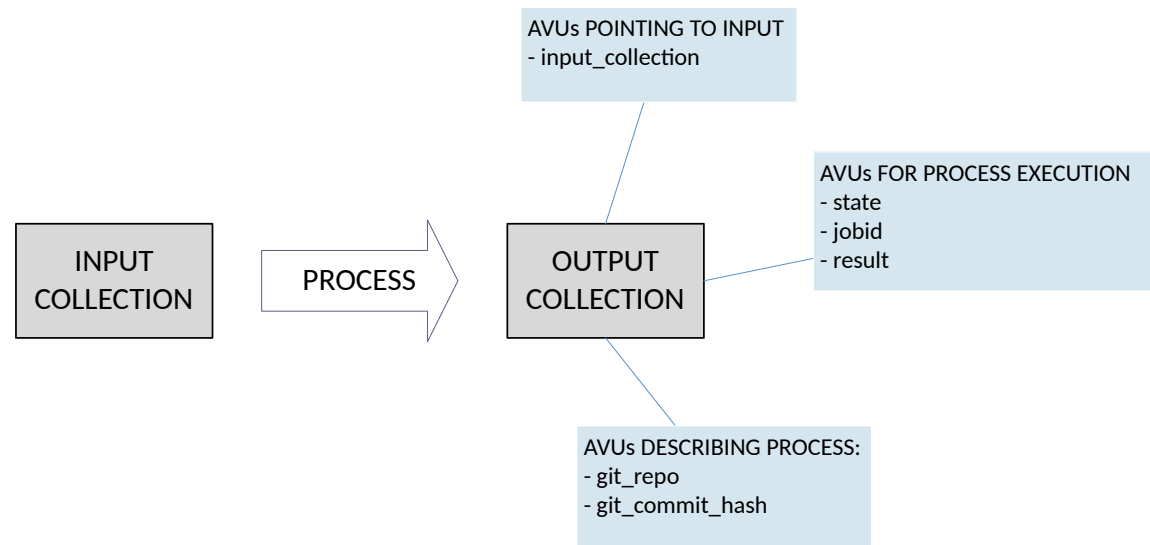
- multiprocess parallel data import
- reliable through use of checksum verification
- multiple import sources:
 - local file system
 - other irods instance
 - SMB
 - scp
 - (s)ftp

<https://github.com/rivm-sys0/intorods>



Jobengine: Automation of analysis processes

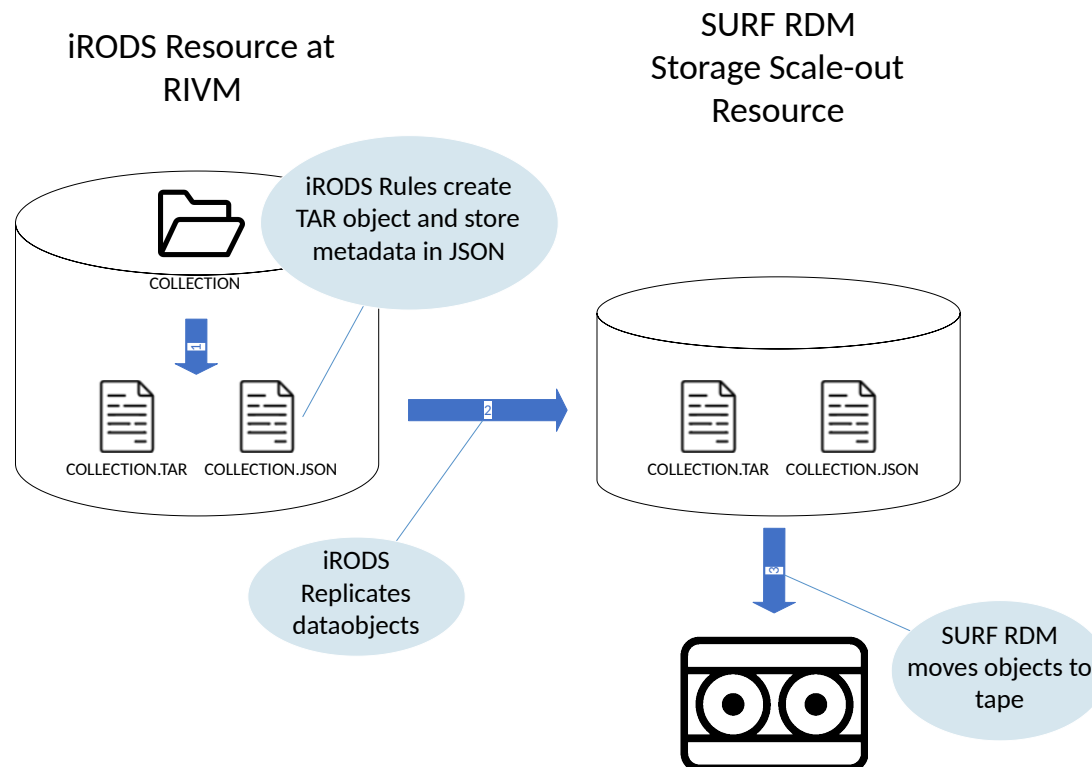
- Used to run *pipeline code* on datasets stored in iRODS
- Controlled by *metadata* on the output collection of the pipeline process
- Starts and monitors tasks on the compute cluster
- Can handle a single pipeline run, or a group of pipelines called a *process group*





Archiving: SURF RDM Storage Scale-out

- **Research Data Management Storage Scale-out** using SURF iRODS Consumer.
- iRODS rules to archive collections





BioRODS: Web interface for users and admins

'Traditional' collection tree

Data provenance graph

User actions

Main menu

Collection contents

Collection and object metadata

The screenshot displays the BioRODS web interface for a specific collection. On the left is a dark sidebar with a 'Main menu' containing links like 'Home', 'About', 'Help', 'Search', 'Reports', 'ONT sample sheet', 'Processes', 'Projects', 'Dataset Uploads', 'Collection browser', and 'Job list'. The main content area is titled 'Collection: /hivmZone/projects/salm/230217_NB502001_0403_AHC7N7AFX5_0007'. It features a 'Collection tree' on the left, a 'Data graph' in the center showing a network of data objects and their relationships, and a table of 'Collection contents' on the right. The table lists objects with columns for Name, Size, Create Date, and Owner. Below the table is a 'Showing 1 to 10 of 51 rows' indicator. To the right of the table is a 'Collection and object metadata' section with tabs for 'Collection metadata', 'Object metadata', and 'Formatted'. The 'Object metadata' tab is active, showing a table of attributes and values for a specific object.

Attribute	Value
complete	true
projectID	salm
sequencing.brand	Bluewin
sequencing.pool	151.224.198.13
sequencing.platform	nextseq
sequencing.run_id	230217_NB502001_0403_AHC7N7
sequencing.run_number	403
sequencing.serial	NB502001
sync.archive.deleted_state	01000
sync.archive.enabled	true
sync.archive.backupcheck	18-02-2023 21:08:14
sync.archive.backuprun	18-02-2023 20:58:52
sync.archive.stage	false
sync.archive.state	01000
sync.collection_size	4781 GB
sync.collection_size_time	18-02-2023 20:58:15



Collection browser (1/3)

NGSweb

Robert Verhagen
rivmZone

Logout

Job list

Collection browser

Dataset Uploads

Projects

Processes

ONT sample sheet

Admin

Reports

Search (New)

Search (Old)

Home

Collection: /rivmZone/projects/yers/230214_NB552493_0139_AHC7MFAFX5_0040

Collection tree

- ▼ rivmZone
 - ▶ data
 - ▶ home
 - ▼ projects
 - ▶ adhoc
 - ▶ asperg
 - ▶ bacid
 - ▶ baseclear
 - ▶ bsr_amr
 - ▶ bsr_rvp
 - ▶ campy
 - ▶ cbm
 - ▶ demo
 - ▶ dsshq
 - ▶ dv_amr
 - ▶ ee_vcov
 - ▶ gasadhoc
 - ▶ gzb
 - ▶ iiv
 - ▶ koolj
 - ▶ miasmas
 - ▶ mlu_hev
 - ▶ mlu_omo
 - ▶ myco
 - ▶ myco_kncv
 - ▶ neg_ctrls
 - ▶ ngslab
 - ▶ nonacris
 - ▶ non-amr
 - ▶ nrs
 - ▶ nrs_ontinf
 - ▶ ovbbd
 - ▶ ovzeno
 - ▶ pienter_up
 - ▶ qc
 - ▶ r-cpe
 - ▶ r-cppa
 - ▶ refsamp
 - ▶ r-mesa

Data graph

Simplify

Graph Levels

3

Name	Size	Create Date	Owner
audit_trail	DIR	15-02-2023 23:36:58	rods
clean_fastq	DIR	15-02-2023 23:37:24	rods
de_novo_assembly	DIR	15-02-2023 23:37:18	rods
de_novo_assembly_filtered	DIR	15-02-2023 23:37:17	rods
identify_species	DIR	15-02-2023 23:36:52	rods
log	DIR	15-02-2023 23:04:32	rods
multiqc	DIR	15-02-2023 23:36:59	rods
qc_clean_fastq	DIR	15-02-2023 23:37:23	rods
qc_de_novo_assembly	DIR	15-02-2023 23:37:02	rods
qc raw fastq	DIR	15-02-2023 23:37:01	rods

Archive

Storage

Pipeline

Validity

ONLINE

Keep online for:

Choose...

bring online

Collection metadata

Object metadata

Formatted

Attribute	Value
import_foldername	/mnt/irodstier1/pipelines/p
import_timestamp	15-02-2023 23:37:53
projectID	yers
sequencing::brand	Illumina
sequencing::host	131.224.198.35
sequencing::platform	nextseq
sequencing::run_id	230214_NB552493_0139
sequencing::run_number	139
sequencing::serial	NB552493
sys::archive::desired_state	01000
sys::archive::enable	true
sys::archive::lastcheck	16-02-2023 00:09:16
sys::archive::lastrun	15-02-2023 23:43:23



Coll. browser contents and integrated document viewer (3/3)

The screenshot displays the NGSweb interface. On the left is a sidebar with navigation options: Erwin van Wieringen, nvmZone, Logout, Job list, Collection browser (selected), Dataset Uploads, Projects, Processes, ONT sample sheet, Admin, Reports, Search (New), and Search (Old). The main area shows a 'Collection: /rnmZone/projects/191108_NB502001_HW7TKAFX_000' with a 'Collection tree' on the left listing various subdirectories like data, home, projects, adhoc, asperg, bacid, baseclear, bsr_amr, bsr_rvp, campy, cbm, demo, dsshig, dv_amr, ee_vcov, gasadhoc, gzb, iiv, koolj, miasmas, mlu_hev, mlu_omo, myco, myco_kncv, neg_ctrls, ngslab, nonacris, non-amr, nrs, nrs_ontinf, ovbld, ovzeno, pienter_up, and qc. The 'qc' directory is expanded, showing files like 191108_NB502001_HW7TKAFX_000, 191108_NB502001_HW7TKAFX_000, and 200828_NB502001_0149_A_HGHJJAFOX2_0010. An integrated document viewer is open, displaying 'NGS_QC_report.html'. The viewer has a 'Close' button and a 'Run stats' tab. The report title is 'QC Report In-house sequencing' with a date of '22-08-22 to 23-02-20'. It includes a paragraph about quality control metrics and a 'Run stats' section with a 'Cluster Density of the Run' subsection. The 'Cluster Density of the Run' section explains the importance of cluster density and provides interpretation help. A line graph titled 'Cluster density per run' shows the cluster density per run (Y-axis, 1000 to 1500) across lanes (X-axis). The graph shows a fluctuating red line with a gray shaded area representing the recommended density for balanced libraries.

NGSweb

Erwin van Wieringen
nvmZone Logout

Job list
Collection browser
Dataset Uploads
Projects
Processes
ONT sample sheet
Admin
Reports
Search (New)
Search (Old)

Collection: /rnmZone/projects/191108_NB502001_HW7TKAFX_000

Collection tree

- ▼ rnmZone
 - ▶ data
 - ▶ home
 - ▶ projects
 - ▶ adhoc
 - ▶ asperg
 - ▶ bacid
 - ▶ baseclear
 - ▶ bsr_amr
 - ▶ bsr_rvp
 - ▶ campy
 - ▶ cbm
 - ▶ demo
 - ▶ dsshig
 - ▶ dv_amr
 - ▶ ee_vcov
 - ▶ gasadhoc
 - ▶ gzb
 - ▶ iiv
 - ▶ koolj
 - ▶ miasmas
 - ▶ mlu_hev
 - ▶ mlu_omo
 - ▶ myco
 - ▶ myco_kncv
 - ▶ neg_ctrls
 - ▶ ngslab
 - ▶ nonacris
 - ▶ non-amr
 - ▶ nrs
 - ▶ nrs_ontinf
 - ▶ ovbld
 - ▶ ovzeno
 - ▶ pienter_up
 - ▼ qc
 - 191108_NB502001_HW7TKAFX_000
 - 191108_NB502001_HW7TKAFX_000
 - 200828_NB502001_0149_A_HGHJJAFOX2_0010

Showing 1 to 8 of 8 rows

NGS_QC_report.html

Close

Run stats

- Cluster Density of the Run
- Run yield and quality
- Control sample stats
- Metrics per sample

QC Report In-house sequencing

Date: 22-08-22 to 23-02-20

The quality control metrics are better interpreted while put into context and when followed up at different time points. This report includes history data from the last six months.

Run stats

Cluster Density of the Run

The cluster density is an important factor to determine the sequencing performance. It influences the quality of the data and the yield of the run. Overclustering may cause lower Q30 scores, lower clusters passing filter, inaccurate demultiplexing and even run failure. Underclustering maintains high data quality, but lowers data output (see: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/cluster-optimization-overview-guide-1000000071511-00.pdf for more information).

Interpretation help:

- Gray shadow: samples that fall within this area are within the recommended density for balanced libraries in a NextSeq run.
- Color lines: represent each of the four lanes of a NextSeq flow cell.

Note: Raw cluster density is not a useful metric for patterned flow cells because the ordered arrangement of nanowells ensures uniform cluster density

Cluster density per run

Value

3e29b003-9917-4557-ae



Job list

NGSweb

Robert Verhagen
rivmZone

Logout

Job list

Collection browser

Dataset Uploads

Projects

Processes

ONT sample sheet

Admin

Reports

Search (New)

Search (Old)

JOB

Processes

Processgroups

Home

Job queue overview

Last refresh:
16-02-2023 12:12:02

Name	State	Description	Group ID	Create time	Start time	End time	Project	Result	Input Collection
a64a22	done	Juno population pipeline	-	16-02-2023 09:25:24	16-02-2023 09:25:36	16-02-2023 10:32:12	rvp_spn	0	..rvp_spn/230213_NB552493_0138_AHC7MFAFX5_0010
52b712	done	cgMLST using chewBBACA	1b7bfd90	15-02-2023 23:06:18	15-02-2023 23:39:17	16-02-2023 00:38:35	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
d1b5d2	done	SeqSphere	1b7bfd90	15-02-2023 23:05:25	15-02-2023 23:39:48	16-02-2023 00:42:50	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
67ce78	done	Juno assembly pipeline	1b7bfd90	15-02-2023 23:04:33	15-02-2023 23:06:34	15-02-2023 23:38:45	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0010
7dc899	done	Typing bacteria (7-locus MLST and serotyper)	1b7bfd90	15-02-2023 23:03:42	15-02-2023 23:40:22	16-02-2023 00:29:54	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
4d7dc5	done	Pipeline for antimicrobial resistance	1b7bfd90	15-02-2023 23:02:52	15-02-2023 23:40:53	16-02-2023 00:20:13	yers	0	..ts/yers/230214_NB552493_0139_AHC7MFAFX5_0040
d62fb1	done	SeqSphere	462fa3ee	15-02-2023 23:00:55	15-02-2023 23:44:17	16-02-2023 00:38:41	svstec	0	..svstec/230214_NB552493_0139_AHC7MFAFX5_0034
4056ff	done	Typing bacteria (7-locus MLST and serotyper)	462fa3ee	15-02-2023 23:00:06	15-02-2023 23:44:51	16-02-2023 00:30:00	svstec	0	..svstec/230214_NB552493_0139_AHC7MFAFX5_0034
d22d25	done	cgMLST using chewBBACA	462fa3ee	15-02-2023 22:59:20	15-02-2023 23:45:26	16-02-2023 00:41:11	svstec	0	..svstec/230214_NB552493_0139_AHC7MFAFX5_0034
e8f40e	done	Juno assembly pipeline	462fa3ee	15-02-2023 22:58:35	15-02-2023 23:07:05	15-02-2023 23:43:19	svstec	0	..svstec/230214_NB552493_0139_AHC7MFAFX5_0009



Processes

NGSweb

Robert Verhagen
rivmZone

Logout

Job list

Collection browser

Dataset Uploads

Projects

Processes

ONT sample sheet

Admin

Reports

Search (New)

Search (Old)

ACTIONS

Add new process

Process details : Juno assembly

Fasta formatter for CLC Bio - AMR

Generate SC2LFV report

Illumina demultiplex and QC

Jovian

Jovian_dev2

Juno AMR

Juno assembly

Juno cgMLST

Juno population

Juno typing

Link RAW data

Miseq demultiplex only

Miseq demultiplex

Settings

Managers

Usage

Description

Juno assembly pipeline

GIT Repository

https://github.com/RIVM-bioinformatics/Juno_pipeline

☒ Ensure data is online before starting process

☒ Download data from iRODS to input directory

☐ Modifies input collection, no output

☐ Restarts on error

☐ Distribution pipeline

Submit



Project processgroups

Project details : salm

Settings Contacts **Processgroups** Collections Users Managers Groups

Process group default New process group

NEW DATA → Juno assembly-1 → Juno AMR-1, SeqSphere on HPC-1, Juno typing-1, Juno cgMLST-1

Selected process: *Juno assembly-1* ([Juno assembly](#))

Add Process	Actions	Dependencies
Process name (optional) <input type="text"/> Assembly with spades Add Process	Select input Delete selected Git Tag v2.0.6 Submit LSF queue bio-prio Submit Set for all processes	Add dependency Delete dependency



Search

NGSweb

Robert Verhagen
rvmZone

Logout

Job list

Collection browser

Dataset Uploads

Projects

Processes

ONT sample sheet

Admin

Reports

Search (New)

Search (Old)

Simple search

Search for: Collection

by: Collection Metadata

00000000-KTTV7

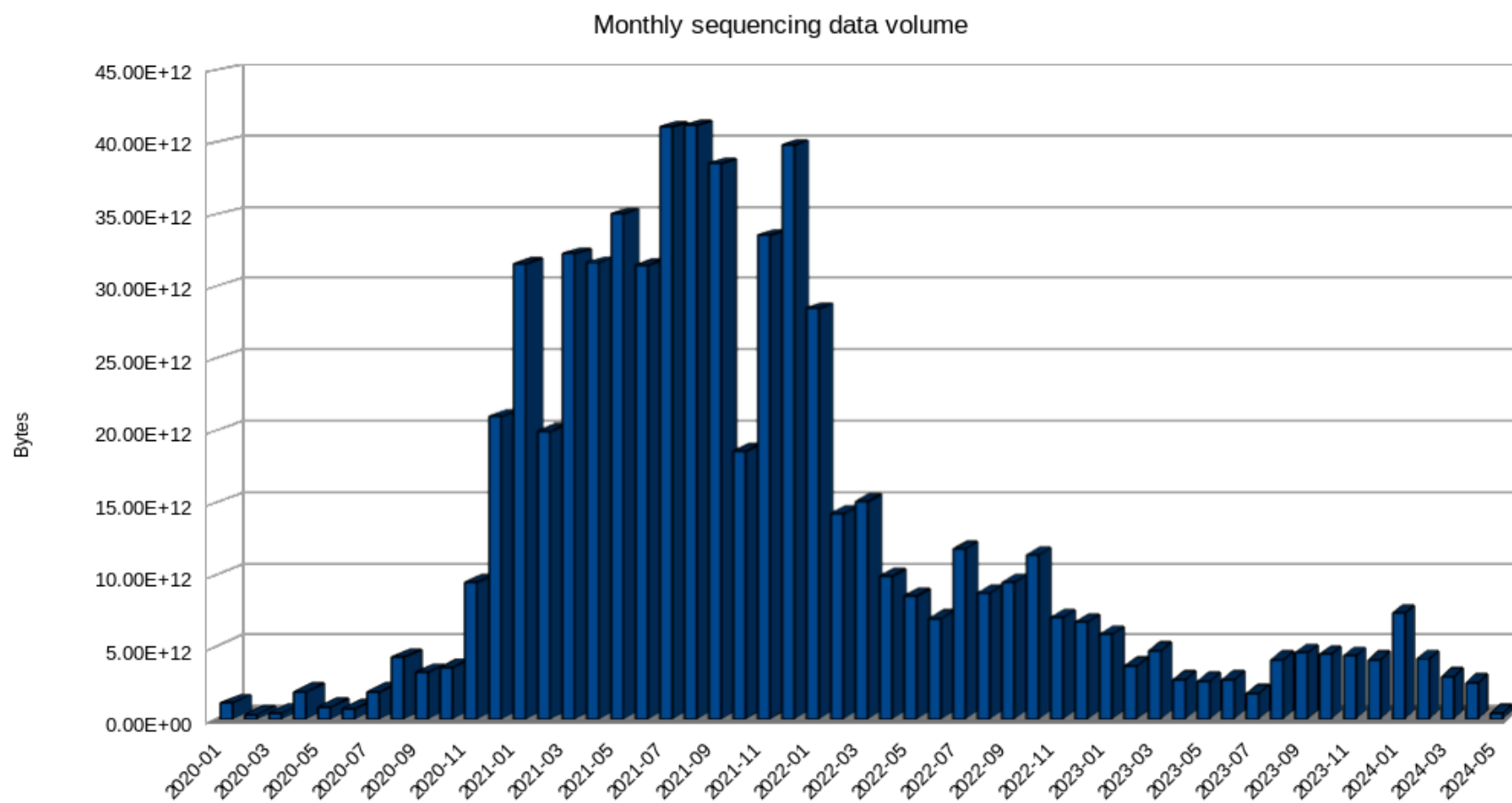
Search results

projectID	collection
ngslab	/rvmZone/projects/ngslab/miseq/230214_M07171_0165_00000000-KTTV7



The pandemic: scaling up

COVID-19 pandemic resulted in sequencing data volumes increasing to 100-fold





Succes stories

- Scale up in a short time.
- Create workplace where people can do research and run production.
- Share pipelines, also externally e.g. <https://github.com/RIVM-bioinformatics/juno-assembly>
- Automating import of sequence data and the starting of pipelines saved a lot of time.
- One place for datasets and archiving to Surf gives a big cost reduction in storage usage.



Future plans

- All (semi) unstructured datasets in iRODS.
- Use SURF and cloud compute for upscaling the HPC.
- Data sharing externally.
- Make our code Open Source.