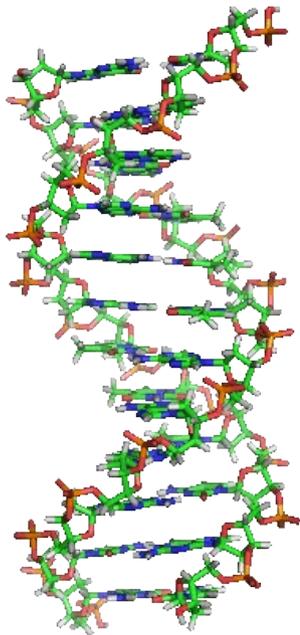


Setup and usage of an e-BioScience infrastructure for high throughput DNA sequence analysis

Authors: Barbera van Schaik, Angela Luyf, Silvia Olabarriaga
Co-authors: Michel de Vries, Katja Ritz, Frank Baas, Antoine van Kampen



Bioinformatics Laboratory,
Clinical Epidemiology, Biostatistics and Bioinformatics
Academic Medical Center
Amsterdam, the Netherlands

b.d.vanschaik@amc.uva.nl



Traditional sequencing

Since 1963

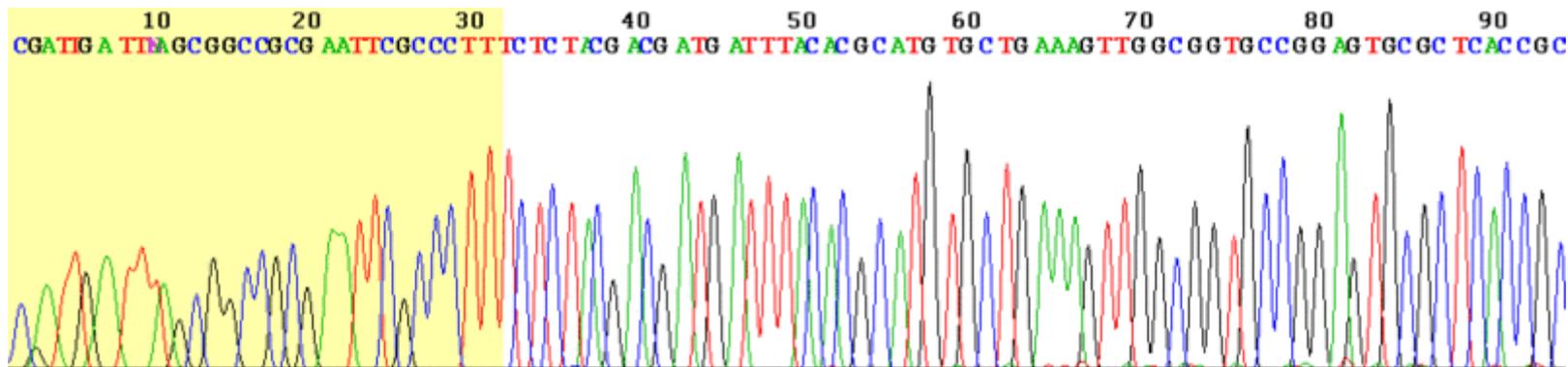
Gel or capillary sequencing

One sequence at the time

Robots: up to 384 samples in one run

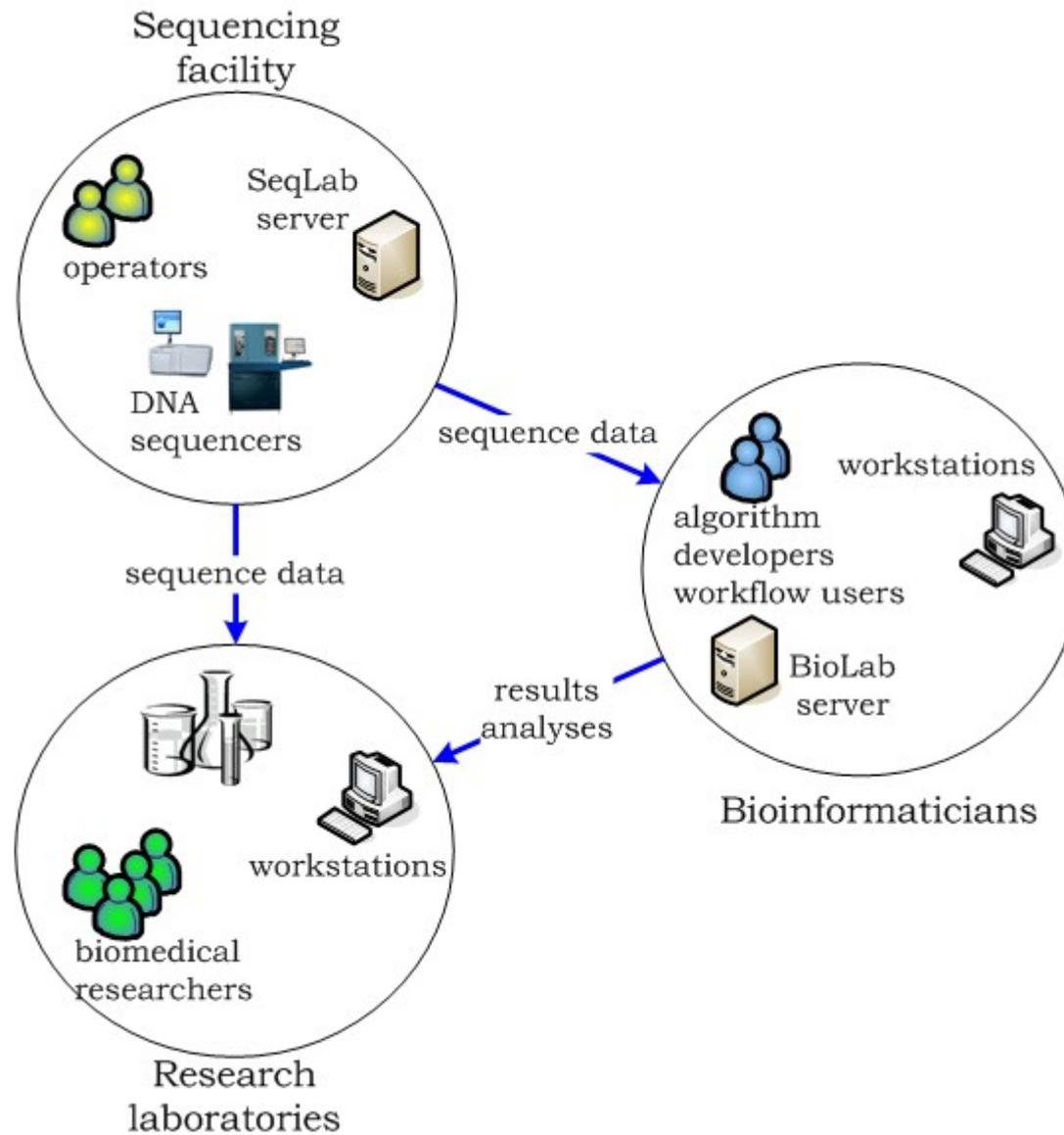


<http://whitehead.mit.edu/>



http://en.wikipedia.org/wiki/DNA_sequencing

Resources



High throughput DNA sequencing

Aka: Second generation, next generation or massively parallel sequencing

Since 2005

Characteristics:

Immobilize DNA fragments on a bead or plate

Multiply DNA fragments in parallel

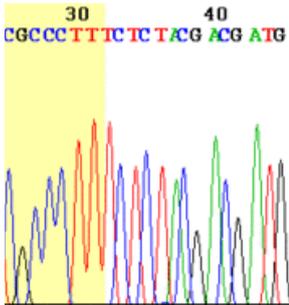
Flow of labelled nucleotides over plate
(and make pictures)

Process images to determine nucleotide sequence



```
>GCJ743F01C2VYF length=227 xy=1145_3765 region=1 run=R_2010_02_19_06_54_29_
TGCCTTTTTACTTTCTTCAGGGATAACATTTAACCTGTAGTCTGCTGTTGGGGTAATCTG
AGTCTGATGTGGCAAGTTCTGCTGCGTTTGCATCAATGGCATGTTTGTGCTATCATAGTT
GTCTGTGTGTAACGGAGGTATGATTTACCCCGAGTCACAGGGGTTTGGAAACCACAGGTA
ACGCGTTTAGACCAGGGTTCCATGCTAATCTCTCTATCTTGGACCA
>GCJ743F01C3XSO length=219 xy=1157_3654 region=1 run=R_2010_02_19_06_54_29_
TG TAGTTTTCTTTCTTCAGGGATAACATTTACCTGTAGTCTGCTGTTGGGGTAATCTGAG
TCTGATGTGGCAAGTTCTGCTGCGTTTGCATCAATGGCATGTTTGTGCTATCATAGTTGT
CTGTGTGTAACGGAGGTATGATTTCCCGAGTCACAGGGTGGAAACCACAGGTAAGCGT
```

Traditional versus high throughput DNA sequencing



<http://en.wikipedia.org/>



<http://www.454.com/>



<http://solid.appliedbiosystems.com/>

Sanger, one run:

? hours (human genome took 15 years)

1-384 sequences

300-1000 nt per sequence

1 KB - 384 KB data

= 1,000-384,000 bases

Year: 1963 – now

Roche 454, one run:

7.5 hours

1,000,000 sequences

500 nt per sequence

35 GB data (including images)

500 MB data (excluding images)

= 500,000,000 bases

Year: 2005 – now

ABI SoLiD, one run:

3-5 days

150,000,000 sequences

50-100 nt per sequence

2-4 TB data (raw data)

= 15,000,000,000 bases

Year: 2005 – now

Bacterial genomes

T and B cell variation

Applications

Gene expression

Re-sequencing

Alternative splicing

Virus discovery

Use case: virus discovery

Laboratory

Blood sample: extract viral DNA and RNA

Viruses in small concentrations

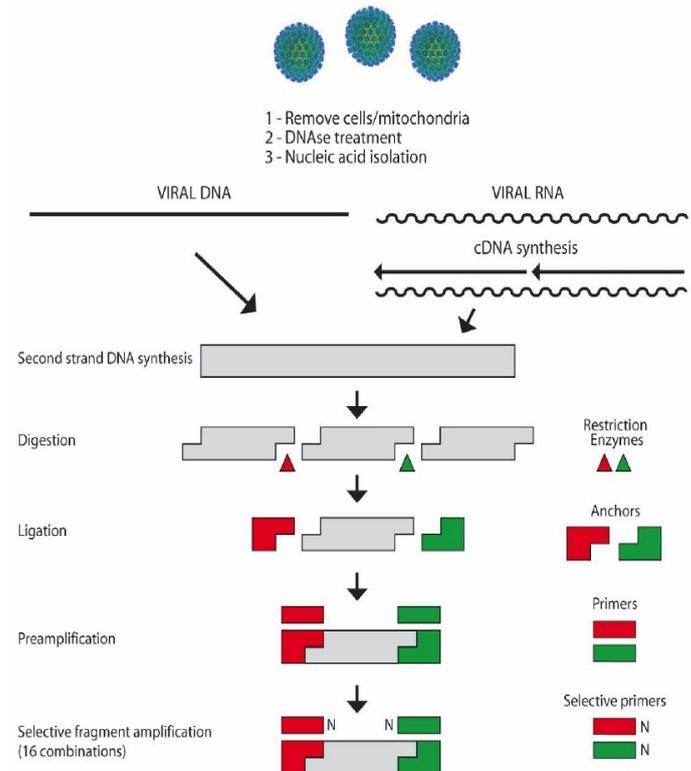
Sample also contains human material

Bioinformatics

Identify all sequences (sequence alignment)

Filter out human material

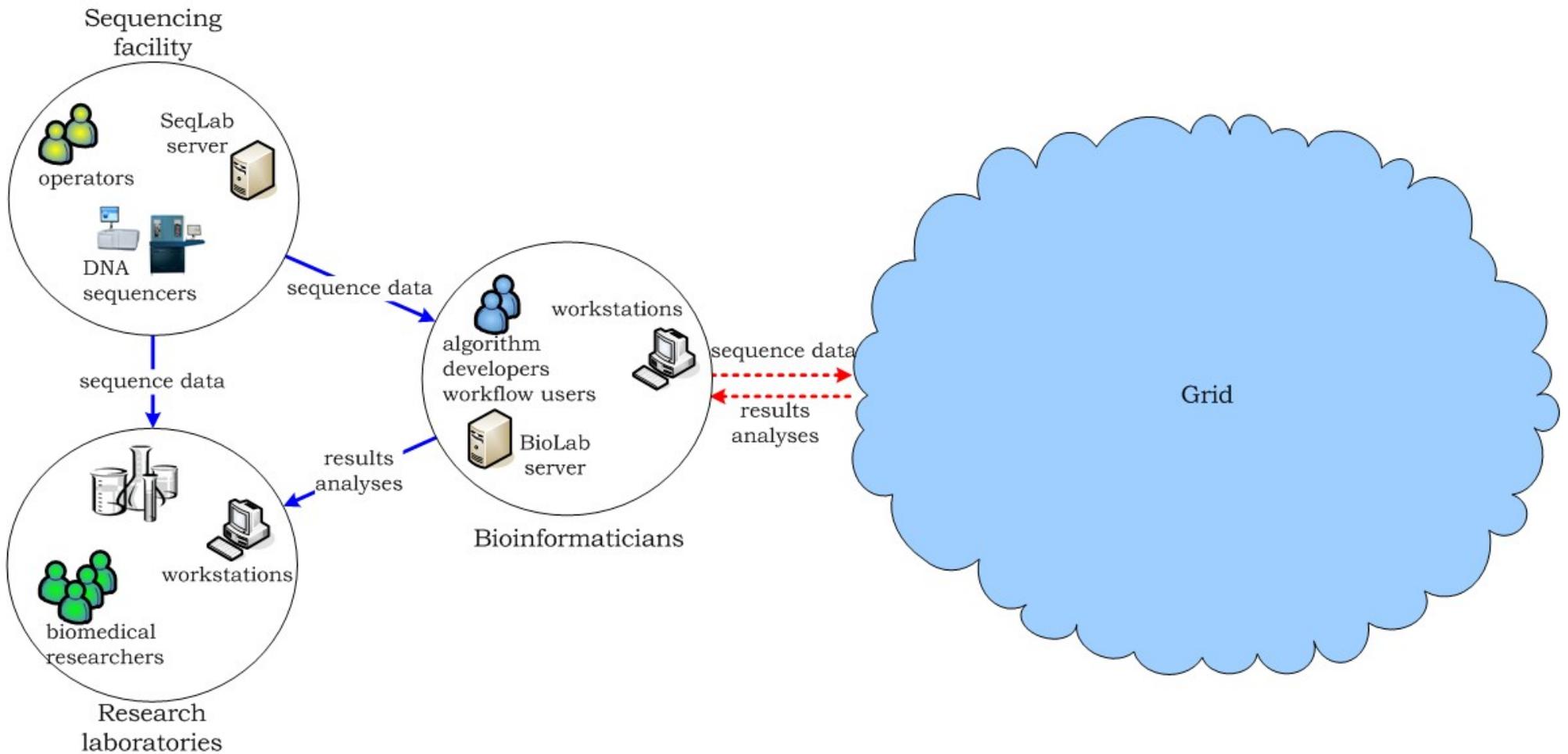
Report viral and bacterial DNA/RNA in sample



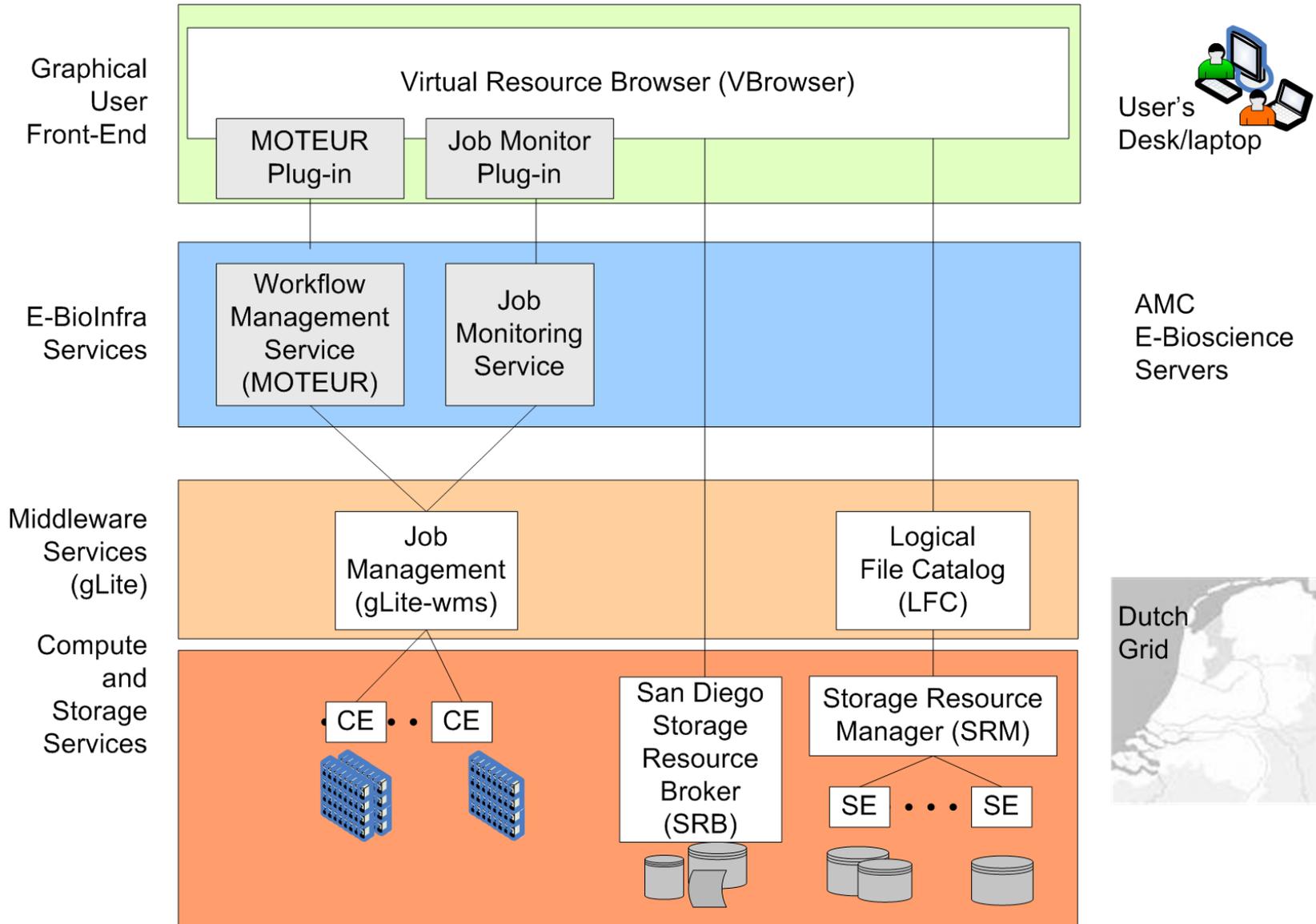
Michel de Vries

11 experiments
6,000,000 sequences
Repeat after database updates 7

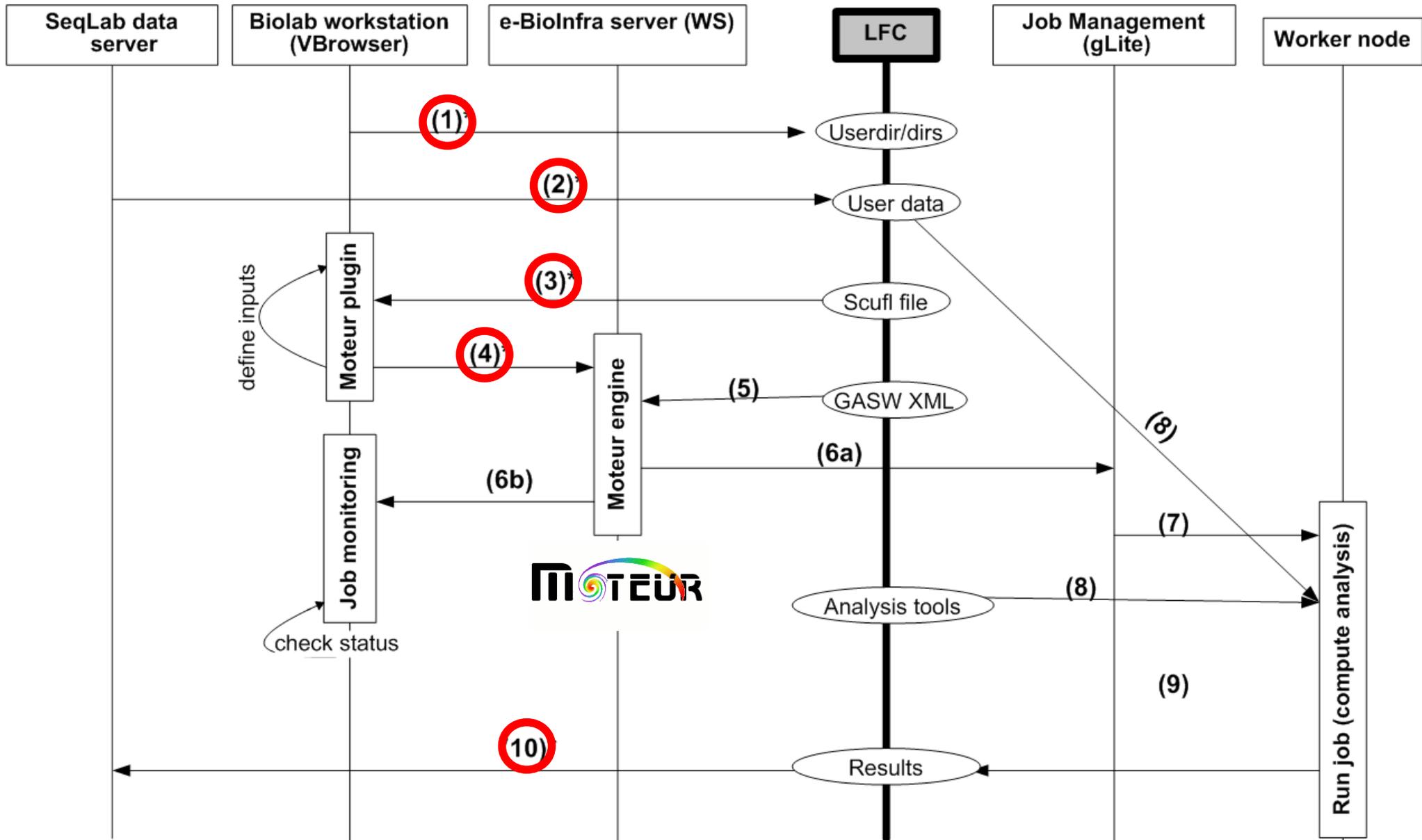
Resources



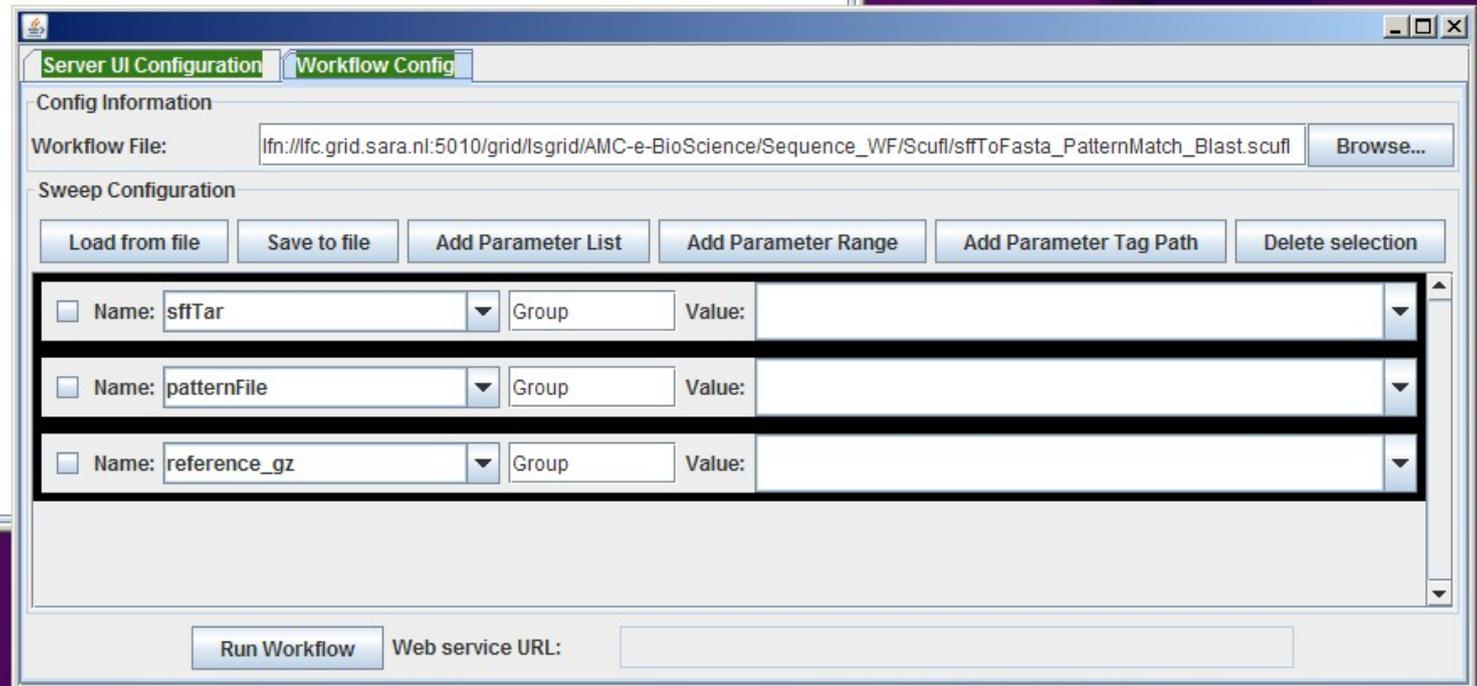
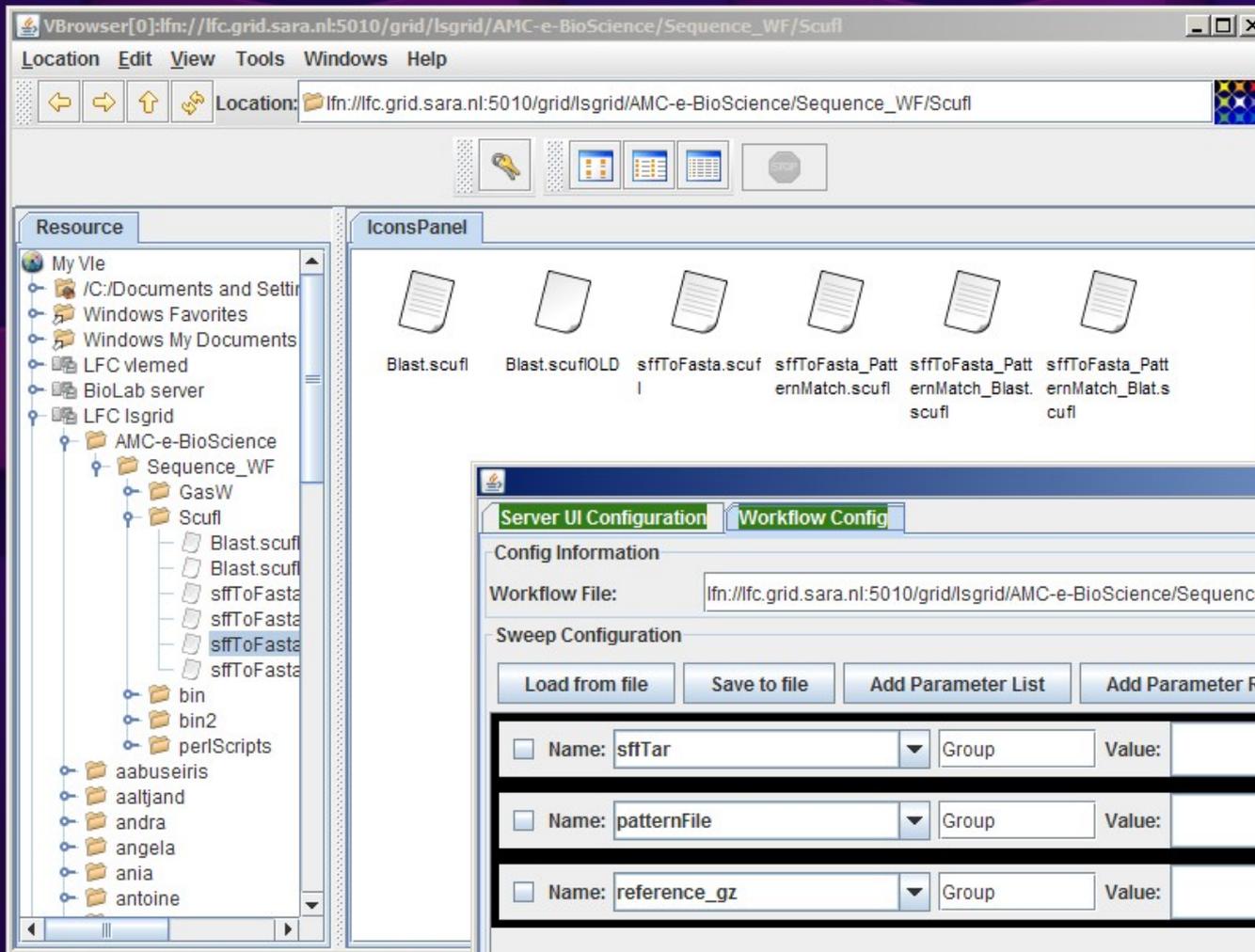
eBioScience infrastructure



User actions and automated actions



In practice – workflow submission



In practice – workflow monitoring

The screenshot displays a web-based workflow monitoring interface. On the left, a 'Resource window' shows a tree view of available resources, including local hosts and various servers. The main area features a 'CobraViewer' showing a workflow graph with nodes and data objects. A 'JOB STATUS: workflow-JGWg3t' window is open on the right, displaying a table of job actions.

Resource window: Lists resources such as 'localhost/C:/', 'AMC-a-BioScience', 'SeqLab server', and 'vlermed.lfc.grid'.

Workflow Graph: Shows a sequence of tasks:

- sffToFasta** (blue box): done:2, running:0, failed:0. Inputs: sffTar, patternFile.
- patternMatch** (blue box): done:4, running:0, failed:0. Inputs: sffToFasta, reference_gz.
- Blastall** (green box): done:5, running:3, failed:0. Inputs: patternMatch, reference_gz.

 Outputs include BlastTarFile, qualOutputTarFile, and txtOutputTarFile.

JOB STATUS: workflow-JGWg3t:

N#	JobName	JobID	JobStatus	link Out	Sele
1	new-sff_2_...	https://wms...	DONE (SU)	Not yet retri...	
2	new-sff_2_...	https://wms...	DONE (SU)	Not yet retri...	
3	new-patter...	https://wms...	DONE (SU)	Not yet retri...	
4	new-patter...	https://wms...	DONE (SU)	Not yet retri...	
5	new-patter...	https://wms...	DONE (SU)	Not yet retri...	
6	new-patter...	https://wms...	DONE (SU)	Not yet retri...	
7	new-Blast...	https://wms...	DONE (SU)	Not yet retri...	
8	new-Blast...	https://wms...	DONE (SU)	Not yet retri...	
9	new-Blast...	https://wms...	DONE (SU)	Not yet retri...	
10	new-Blast...	https://wms...	RUNNING	Not yet retri...	
11	new-Blast...	https://wms...	RUNNING	Not yet retri...	
12	new-Blast...	https://wms...	DONE (SU)	Not yet retri...	
13	new-Blast...	https://wms...	DONE (SU)	Not yet retri...	
14	new-Blast...	https://wms...	DONE (SU)	Not yet retri...	

Yellow callouts highlight 'Job monitoring' (pointing to the job status table), 'Workflow components' (pointing to the workflow graph), and 'Workflow monitoring' (pointing to the overall interface).

Collaborators monitor progress via web browser



Sequencing

TWiki > Sequencing Web > SPneumoniae (2010-04-02, BarberaVanSchaik)

[Edit](#) [Attach](#)

Hello Barbera Van Schaik
[Log Out](#)

Streptococcus pneumoniae

[Create personal sidebar](#)

Contact: [JurgenPiet](#)

[Home](#) **Sequencing Web**

Contact biolab: [BarberaVanSchaik](#)

[+](#) Create New Topic

↓ [Streptococcus pneumoniae](#)

[☰](#) Index

↓ [Description](#)

[🔍](#) Search

↓ [Sequence runs](#)

[+](#) Changes

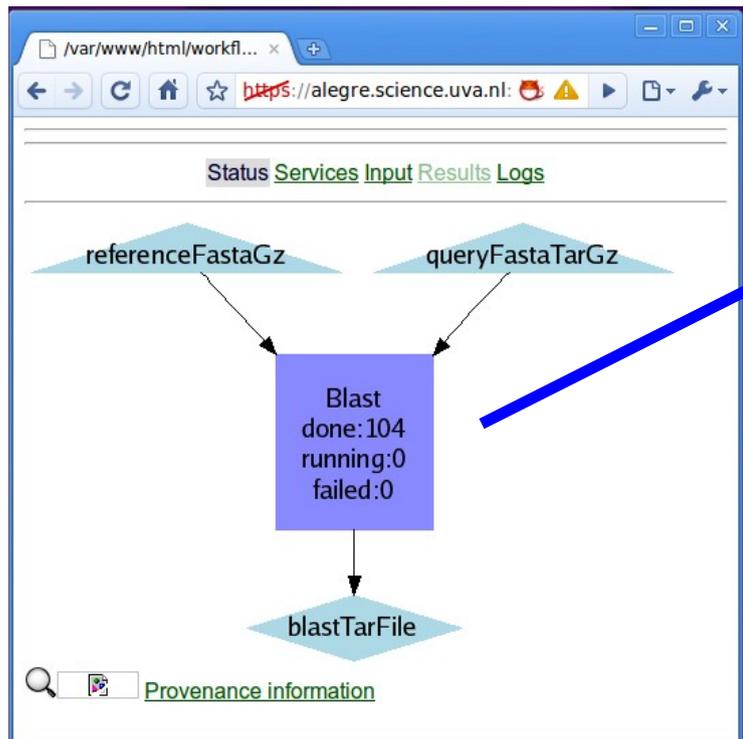
↓ [Methods](#)

[🔧](#) Preferences

April 2010

- 2-4-2010: Submitting grid jobs in batch (bacteria section in chunks)
 - Test: <https://alegre.science.uva.nl:9443/workflows/workflow-B1XLq8/html/workflow-B1XLq8.html>
 - Real: <https://alegre.science.uva.nl:9443/workflows/workflow-MOGk4Y/html/workflow-MOGk4Y.html>

Collaborators monitor progress via web browser

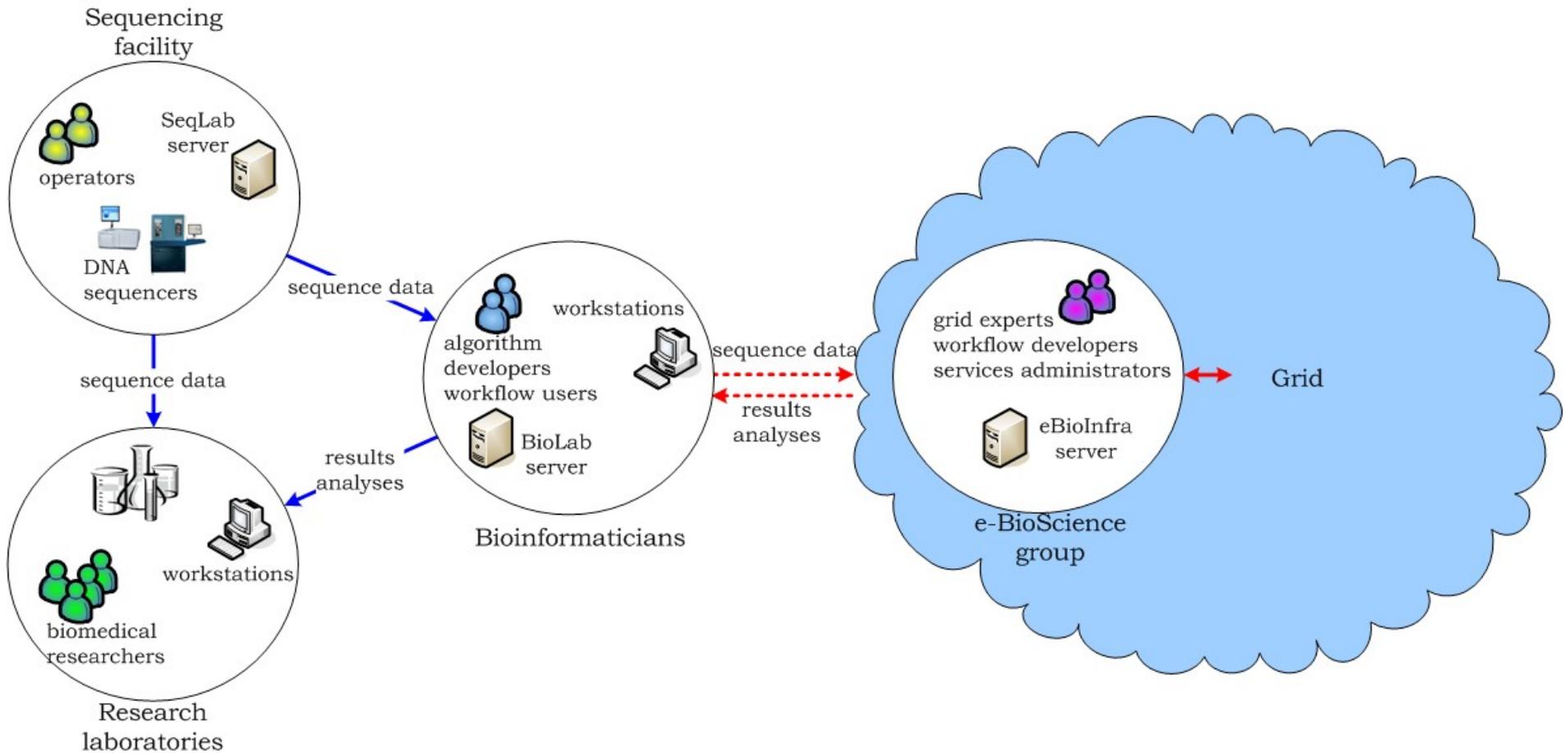


The screenshot shows a web browser window with the URL 'https://alegre.science.uva.nl'. The page title is 'Summary'. Below the title, it displays the job status: 'Waiting: 0 - Running: 0 - Successfully completed: 104 - Failed: 1'. A table follows, showing the distribution of jobs across different states. The 'COMPLETED' row is highlighted in blue. Below the table is a section titled 'Job details' with a table listing individual job statuses, times, and IDs.

State:	Jobs:	
UNKNOWN	0	
WAITING	0	
RUNNING	0	
READY	0	
QUEUED	0	
SUCCESSFULLY_SUBMITTED	0	
TIMEOUT	0	
COMPLETED	104	
ERROR	1	
ABORTED	0	

Status	Time reached	Job id
COMPLETED	02-04-2010 11:15:07	https://wmslb2.grid.sara.nl:9000/a9MsOs1NEYuJ
COMPLETED	02-04-2010	

Resources



Troubleshooting

What if jobs fail?

Examine workflow.out and workflow.err (via vbrowser or web browser)

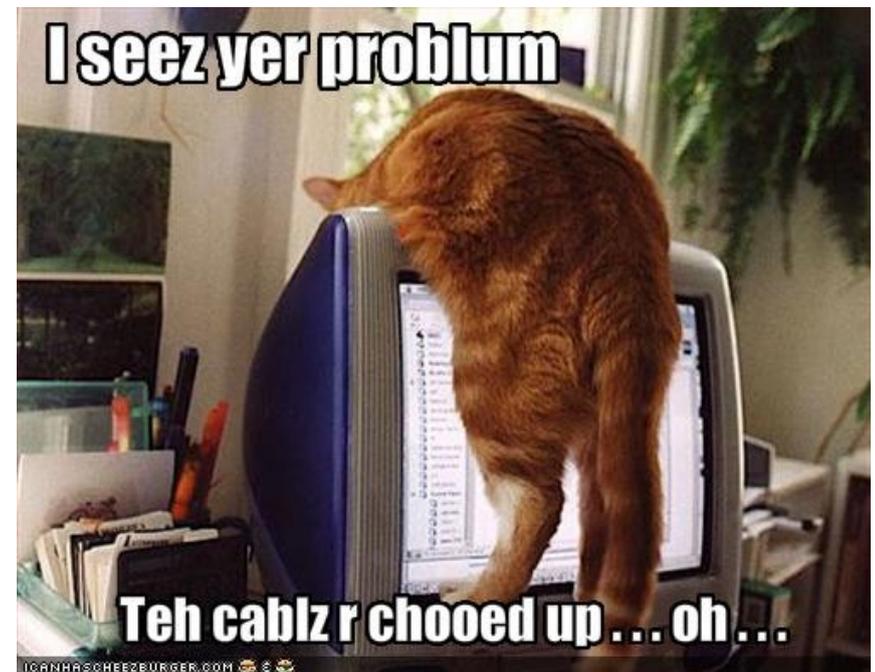
Retrieval of std.out and std.err (job monitoring plug-in)

Check health of servers and services

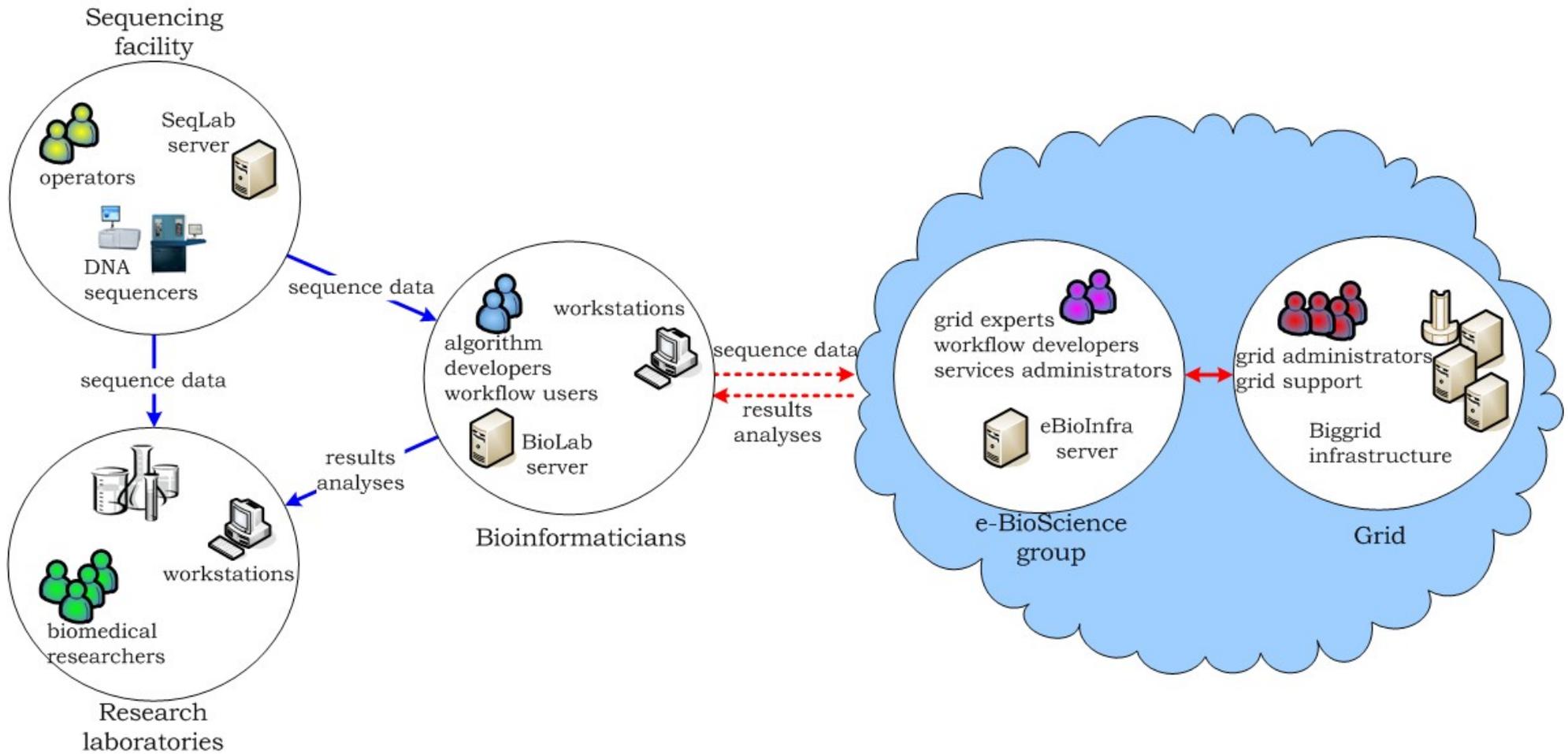
Check e-mail notifications
e.g. scheduled downtimes

Check if data and programs are replicated

Ask the expert



Resources



Important issues

Data security
especially for analysis of patient samples

Maintenance of reference databases/software
e.g. GenBank database updates

User friendliness of interface(s)

Documentation

Support

Positive experiences from the pilot study

Faster: DNA sequence data analyzed more quickly

Easier: Workflows are easier to use than using grid directly

Automation: Parameter sweep (iteration operators)

Collaboration: Grid admins, e-bioscientists, bioinformaticians, the lab

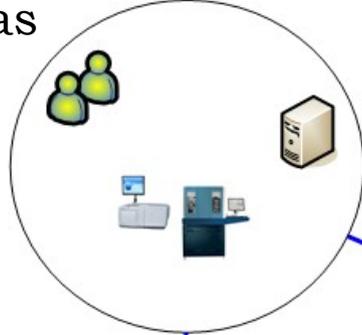


Acknowledgements

Sequence facility

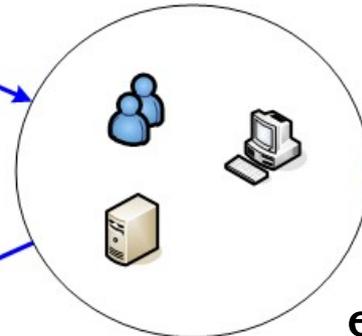
Marja Jakobs
Ted Bradley
Frank Baas

Sequencing
facility

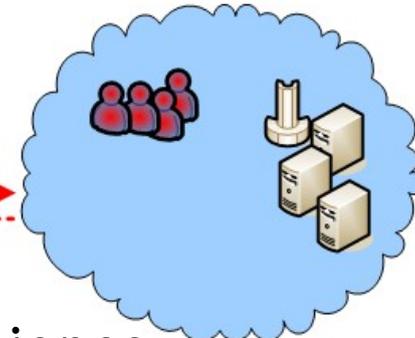


Bioinformatics laboratory

Marcel Willemsen
Barbera van Schaik
Antoine van Kampen



Piter de Boer
Jan Just Keijser
BiG Grid grid-support

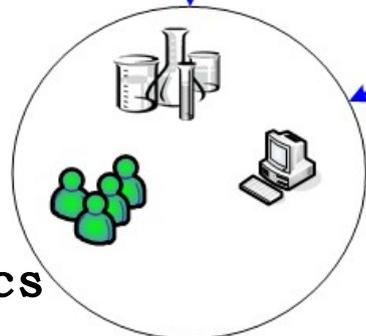


eBioScience

Angela Luyf
Shayan Shahand
Mark Santcroos
Tristan Glatard
Kamel Boulebiar
Martin Stam
Silvia Olabarriaga
Antoine van Kampen

Grid

Bioinformaticians



Research
laboratories

Neurogenetics

Katja Ritz
Frank Baas

Virus discovery unit

Michel de Vries
Martin Deijs
Lia van der Hoek

Research
laboratories

