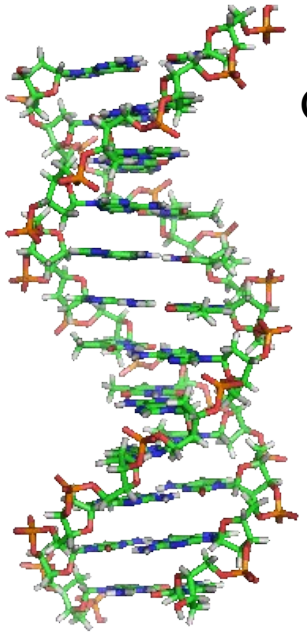
Setup and usage of an e-BioScience infrastructure for high throughput DNA sequence analysis

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

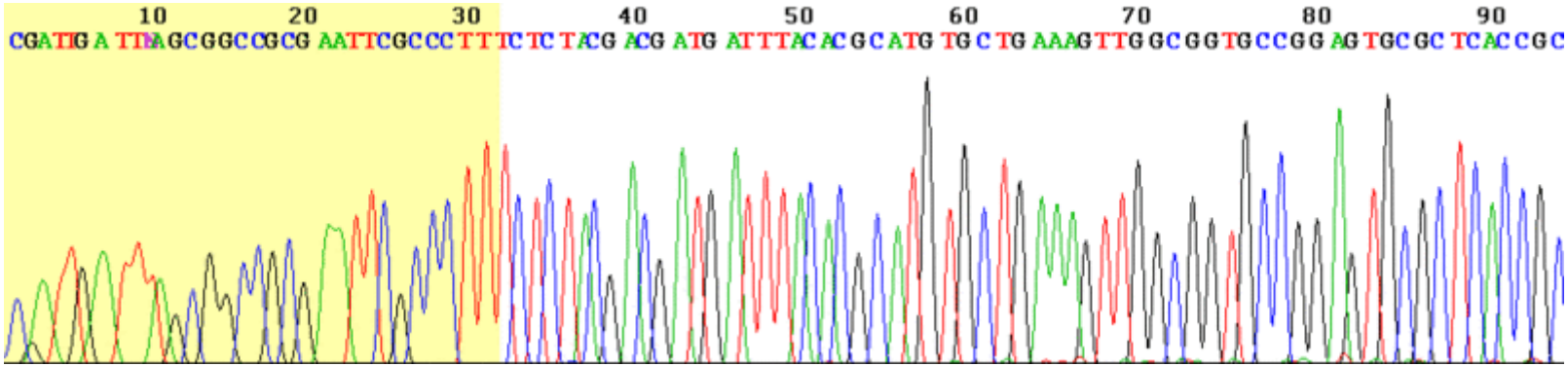
Since 1963

Gel or capillary sequencing

One sequence at the time

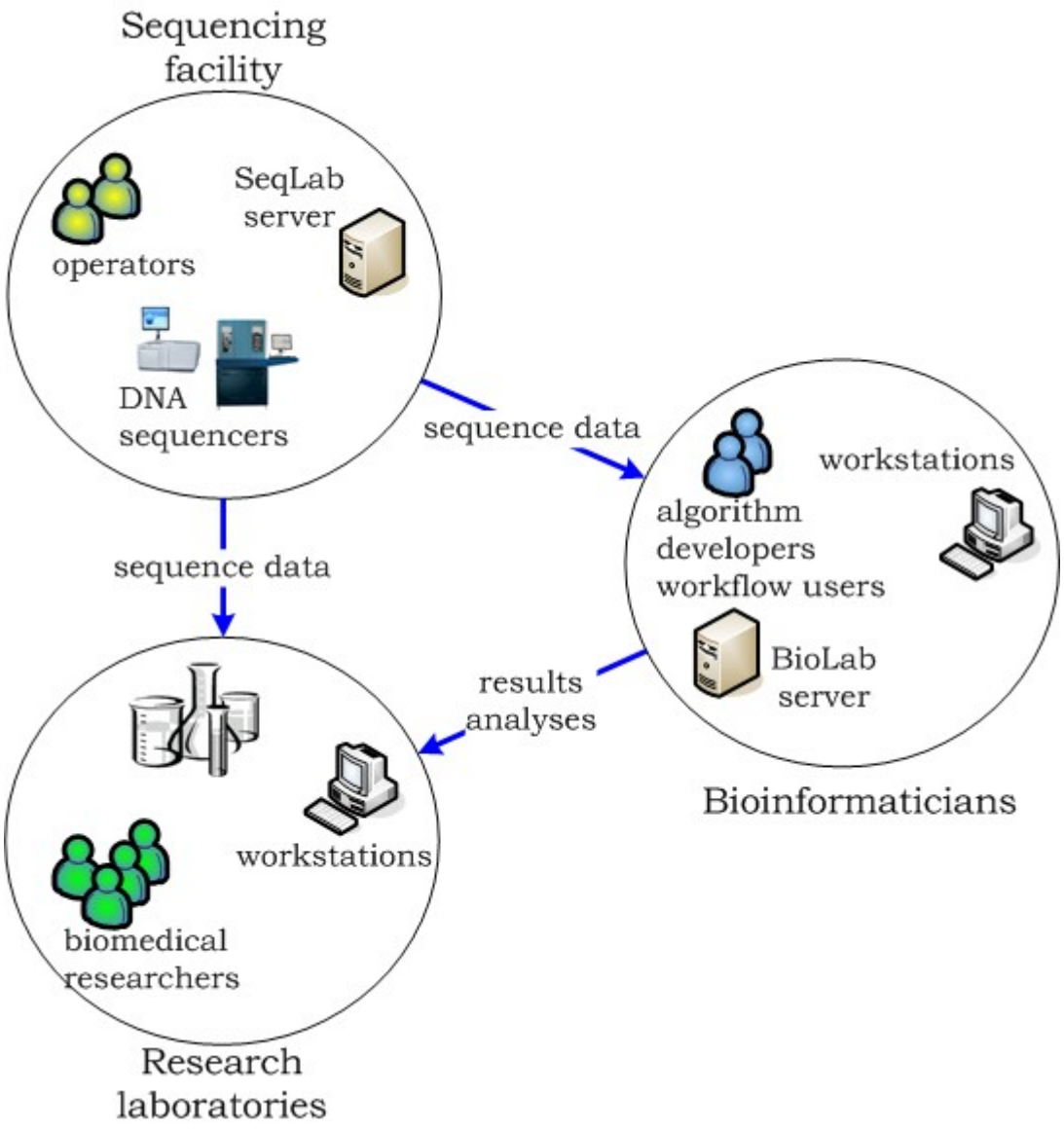
Robots: up to 384 samples in one run

http://whitehead.mit.edu/



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

Resources


Sequencing facility
SeqLab server
operators
DNA sequencers
sequence data
sequence data
workstations
algorithm developers
workflow users
BioLab server
Bioinformaticians
results analyses
workstations
biomedical researchers
Research laboratories

High throughput DNA sequencing

Aka: Second generation, next generation or massively parallel sequencing

Since 2005

Characteristics:

Immobilize DNA fragments on a bead or plate

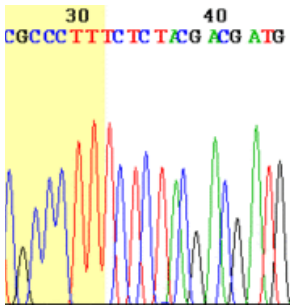
Multiply DNA fragments in parallel

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_
TGCGTTTTTACTTTCTTCAGGGATACCATTTAACCTGTAGTCTGCTGTTGGGGTAATCTG
AGTCTGATGTGGCAAGTTCTGCTGCGTTTGCATCAATGGCATGTTTGTGCTATCATAGTT
GTCTGTGTGTAAACGGAGGTATGATTTACCCCGAGTCACAGGGGTTTGGAACCACAGGTA
ACGCGTTTAGACCAGGGGTTCCATGCTAATCTCTCTATCTTGGACCA
>GCJ743F01C3XSO length=219 xy=1157_3654 region=1 run=R_2010_02_19_06_54_29_
TGTAGTTTTCTTTCTTCAGGGATACCATTTACCTGTAGTCTGCTGTTGGGGTAATCTGAG
TCTGATGTGGCAAGTTCTGCTGCGTTTGCATCAATGGCATGTTTGTGCTATCATAGTTGT
CTGTGTGTAAACGGAGGTATGATTTCCCCGAGTCACAGGGTTGGAACCACAGGTAAGCGT
```

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

Applications

T and B cell variation

Bacterial genomes

Gene expression

Re-sequencing

Alternative splicing

Virus discovery

Use case: virus discovery

Laboratory

Blood sample: extract viral DNA and RNA

Viruses in small concentrations

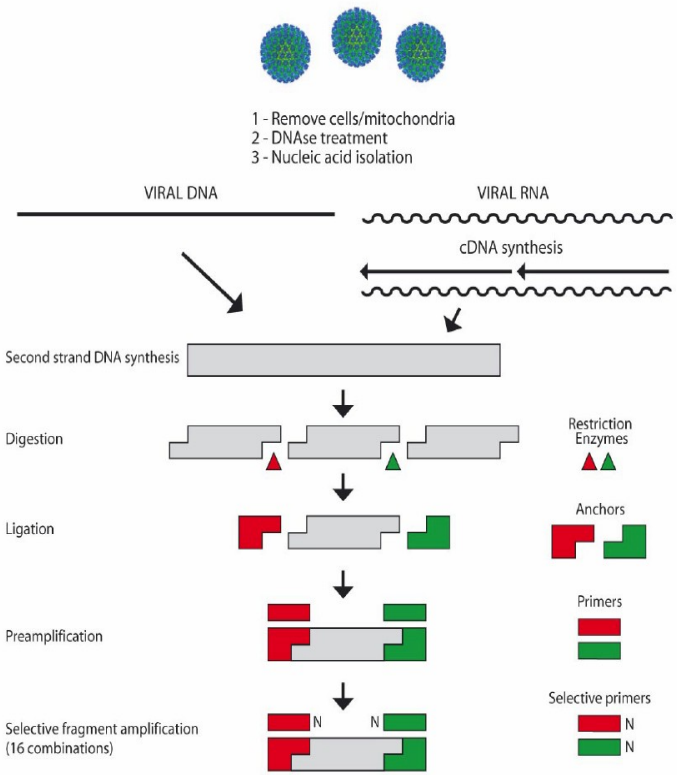
Sample also contains human material

Bioinformatics

Identify all sequences (sequence alignment)

Filter out human material

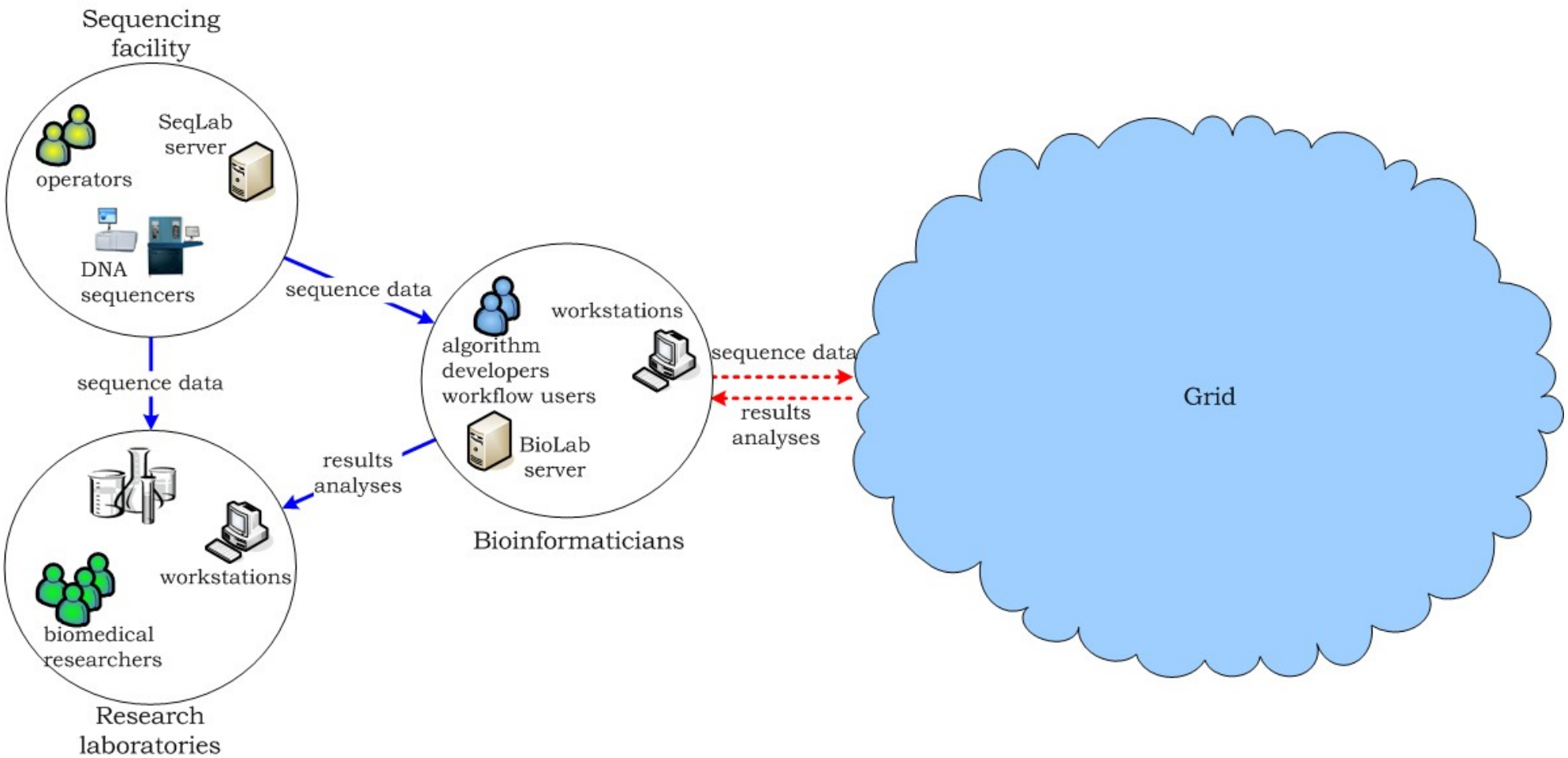
Report viral and bacterial DNA/RNA in sample



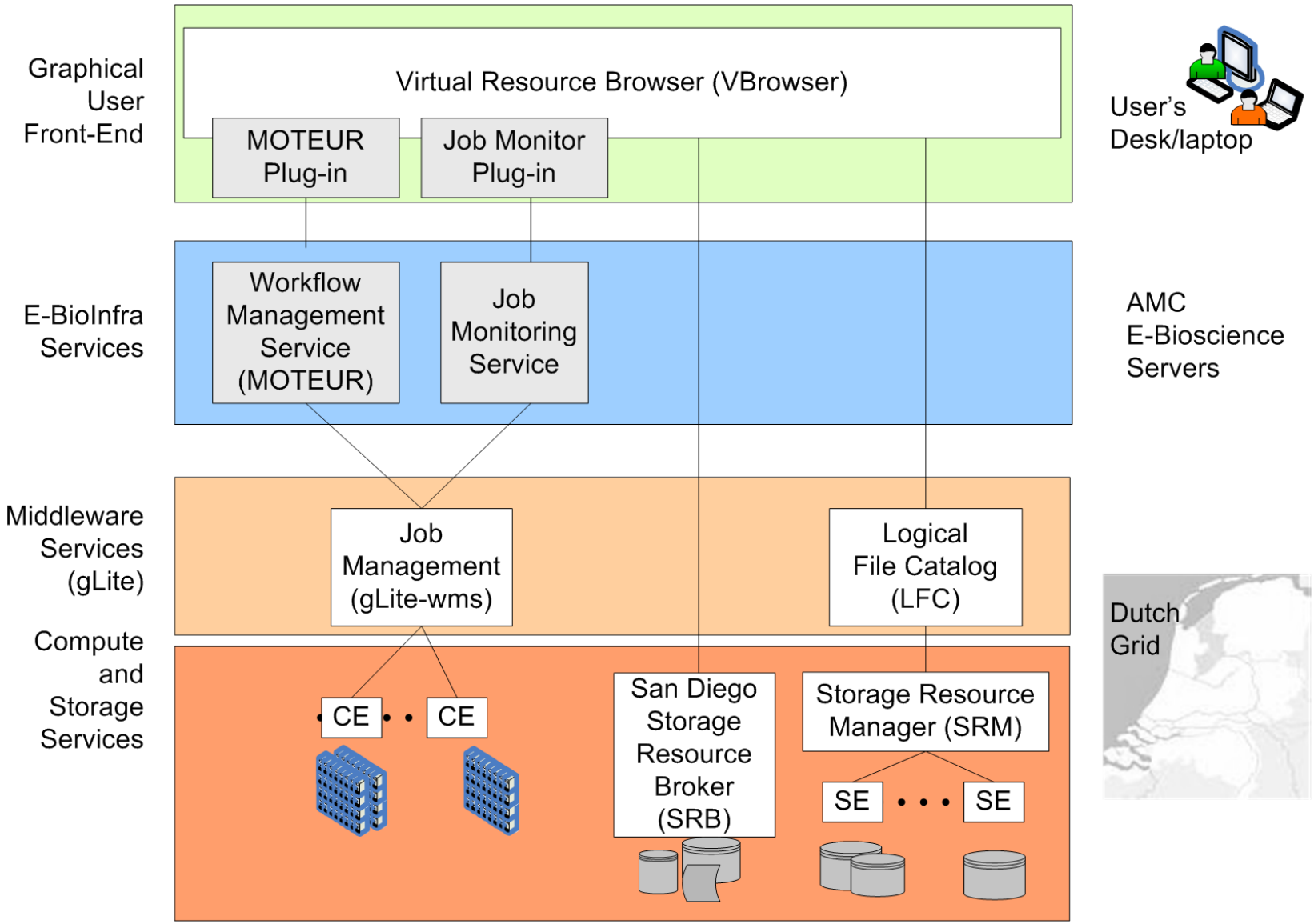
Michel de Vries

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

Resources


Sequencing facility
SeqLab server
operators
DNA sequencers
sequence data
sequence data
workstations
algorithm developers
workflow users
BioLab server
Bioinformaticians
sequence data
results analyses
Grid
results analyses
workstations
biomedical researchers
Research laboratories

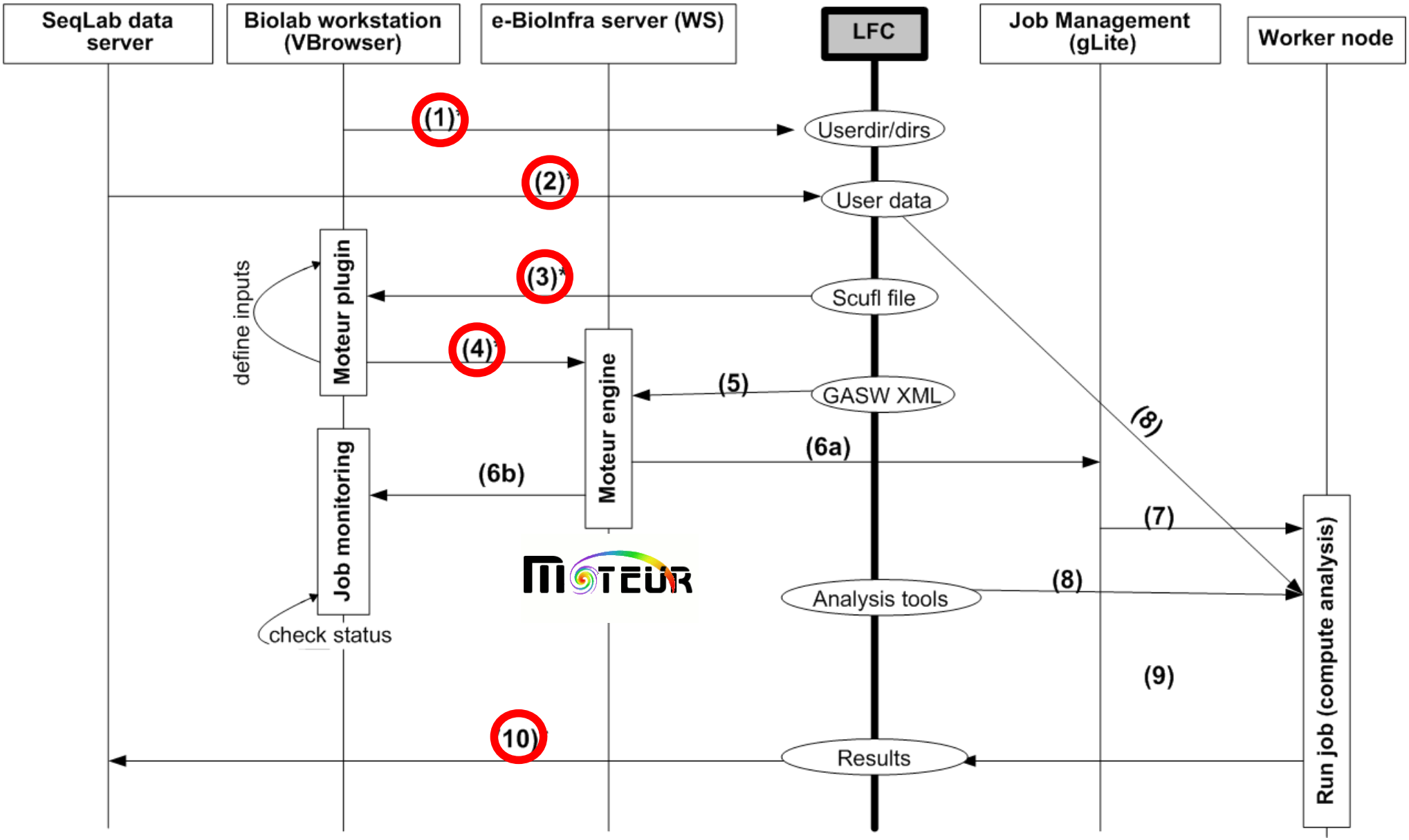
eBioScience infrastructure



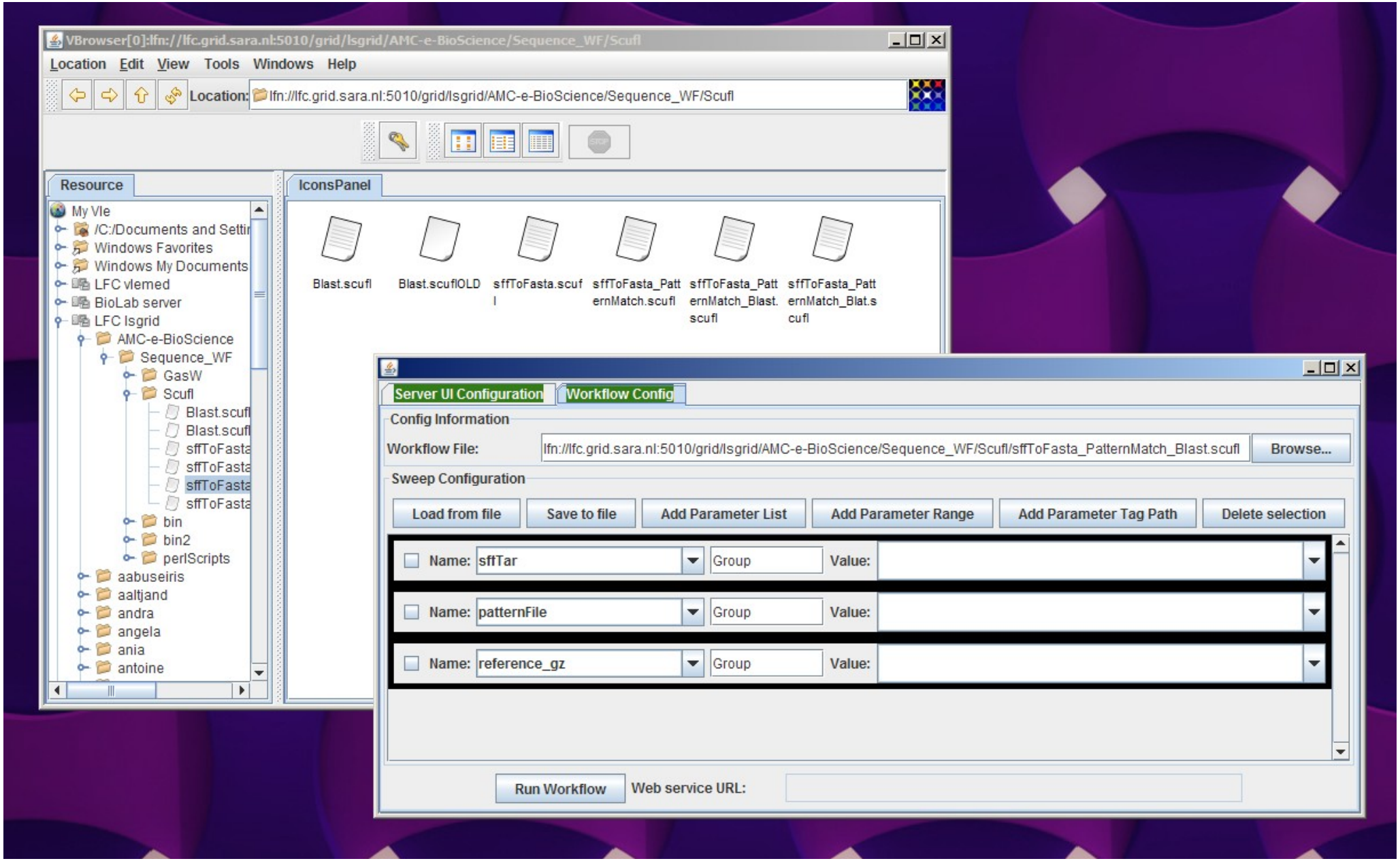
Olabarriaga SD, Glatard T, de Boer PT: A Virtual Laboratory for Medical Image Analysis. IEEE Transactions on Information Technology In Biomedicine, in press

http://www.vl-e.nl/vlemed
http://www.bioinformaticslaboratory.nl/

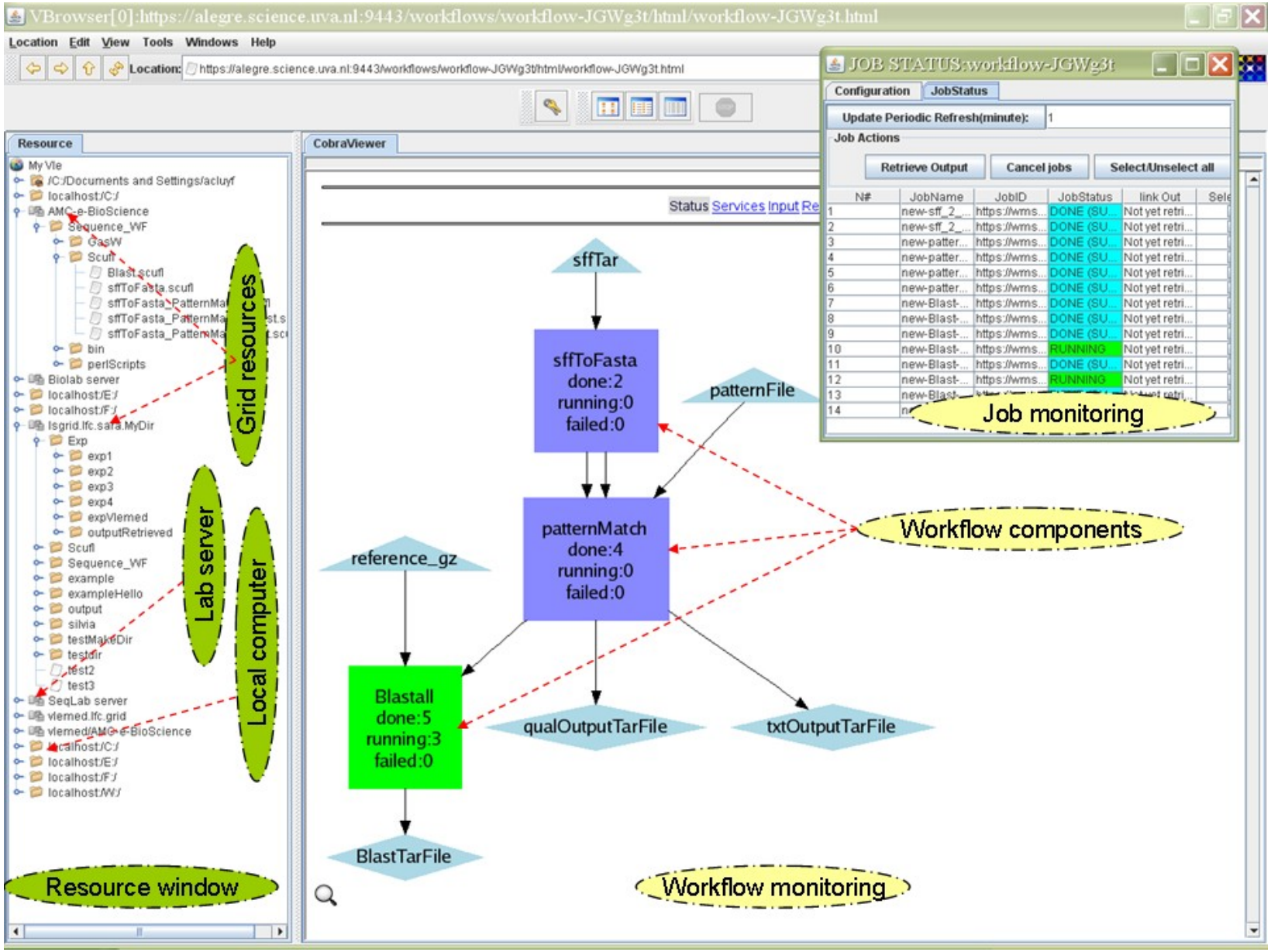
User actions and automated actions



In practice – workflow submission



In practice – workflow monitoring



Collaborators monitor progress via web browser

Bioinformatics Laboratory

Jump | Search

Sequencing

Hello Barbera Van Schaik
Log Out

Create personal sidebar

- **Sequencing Web**
- Create New Topic
- Index
- Search
- Changes
- Preferences

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

Edit | Attach

Streptococcus pneumoniae

Contact: JurgenPiet

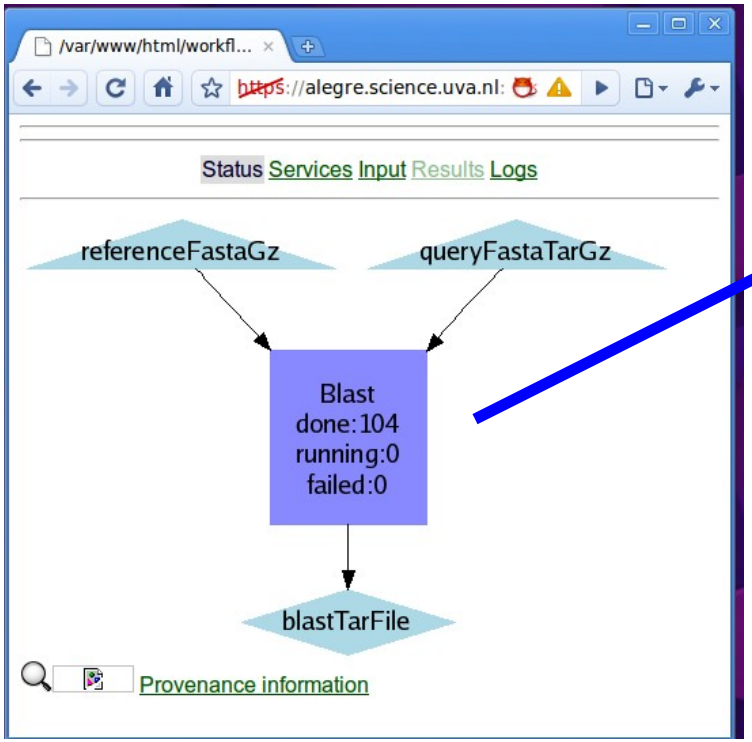
Contact biolab: BarberaVanSchaik

- Streptococcus pneumoniae
  - Description
  - Sequence runs
  - Methods

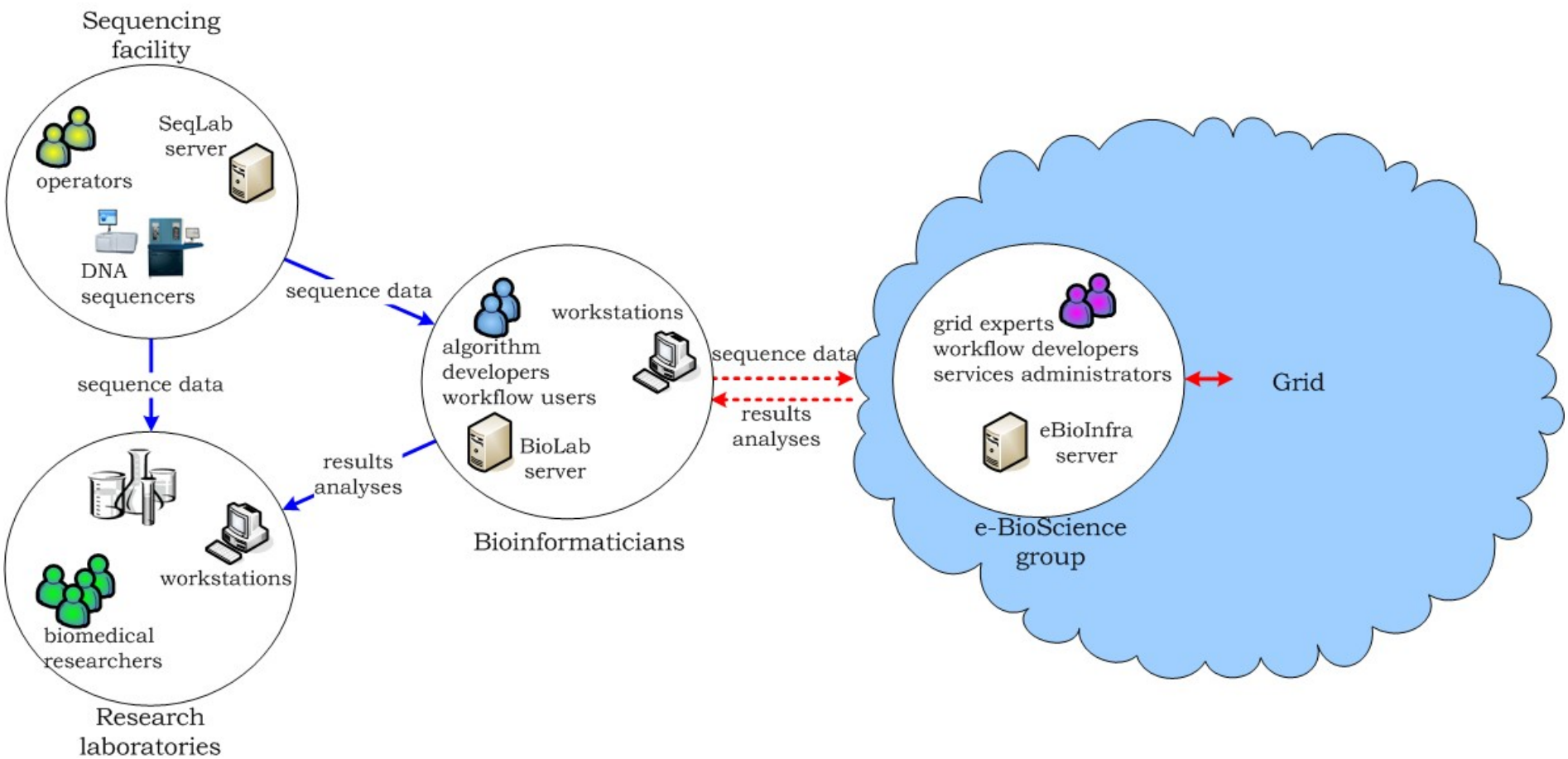
April 2010

- 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



Resources


Sequencing facility
SeqLab server
operators
DNA sequencers
sequence data
sequence data
workstations
algorithm developers
workflow users
BioLab server
Bioinformaticians
sequence data
results analyses
results analyses
grid experts
workflow developers
services administrators
eBioInfra server
e-BioScience group
Grid
workstations
biomedical researchers
Research laboratories

Troubleshooting

What if jobs fail?

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

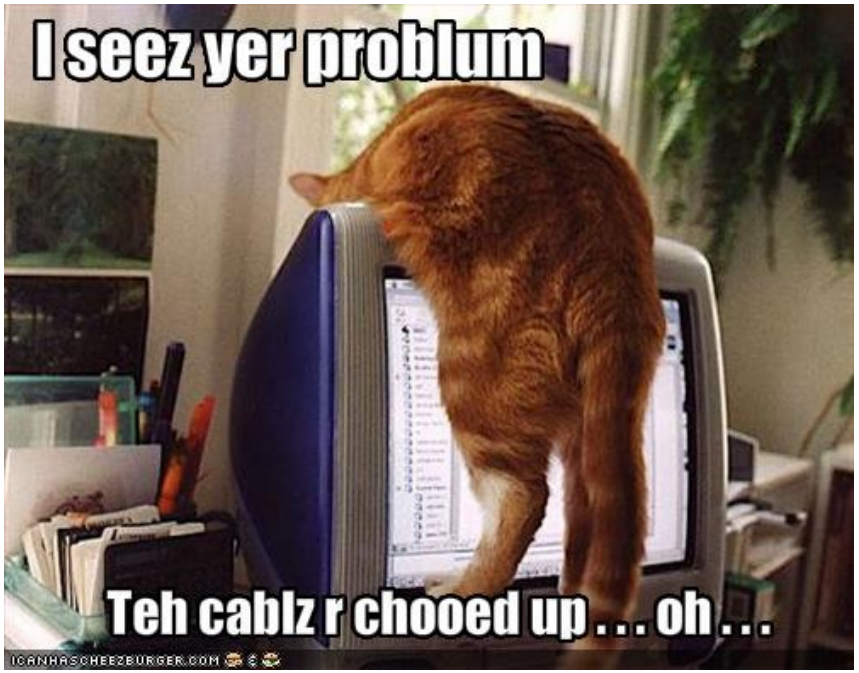
Retrieval of std.out and std.err (job monitoring plug-in)

Check health of servers and services

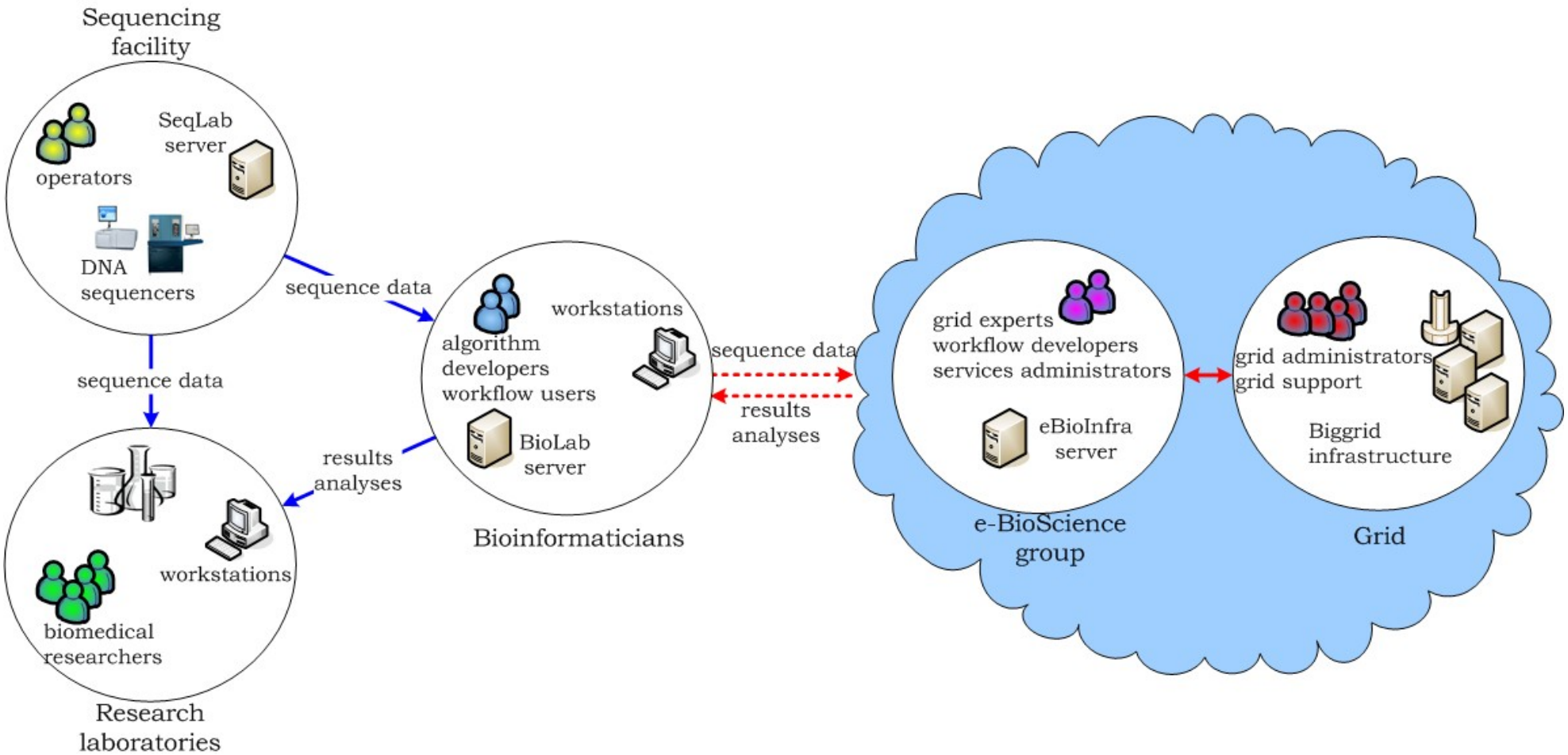
Check e-mail notifications
e.g. scheduled downtimes

Check if data and programs are replicated

Ask the expert



Resources


Sequencing facility
SeqLab server
operators
DNA sequencers
sequence data
sequence data
workstations
algorithm developers
workflow users
BioLab server
Bioinformaticians
results analyses
sequence data
results analyses
grid experts
workflow developers
services administrators
eBioInfra server
e-BioScience group
grid administrators
grid support
Biggrid infrastructure
Grid
workstations
biomedical researchers
Research laboratories

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)

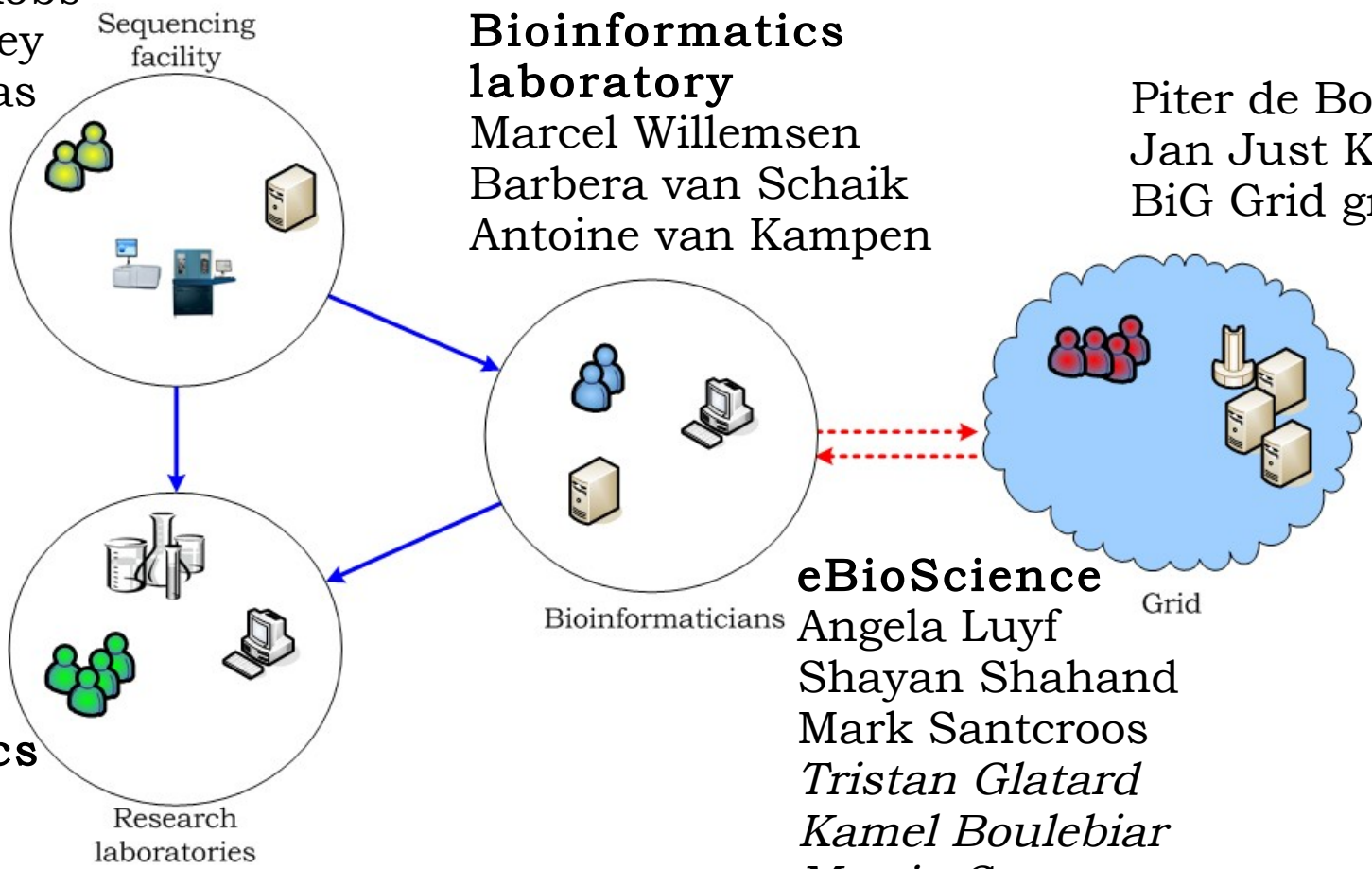
Collaboration: Grid admins, e-bioscientists, bioinformaticians, the lab

Acknowledgements

Sequence facility
Marja Jakobs
Ted Bradley
Frank Baas

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

Neurogenetics
Katja Ritz
Frank Baas

Virus discovery unit
Michel de Vries
Martin Deijs
Lia van der Hoek