Towards visual analytics of bio-workflows

George Byelas

University Medical Centre Groningen, the Netherlands

NBIC-2013, Lunteren, Apr. 16th, 2013

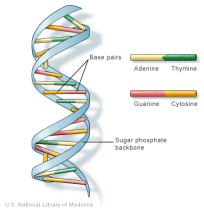
umcg

molgenis

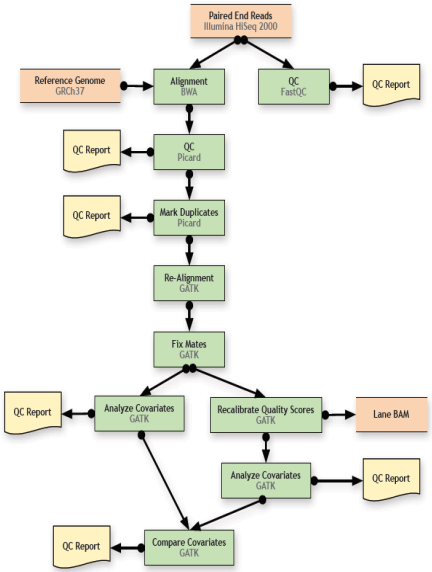
Content

- Bio-workflows
- State of the art workflow visualizations
- How workflow visualization can be improved

Example: NGS alignment workflow



HiSeq



Per Project:
1. Aligned reads
2. QC-reports
3. SNP lists

Raw data → ***alignment workflow*** → Result data

10-100 samples

20 – 200 days

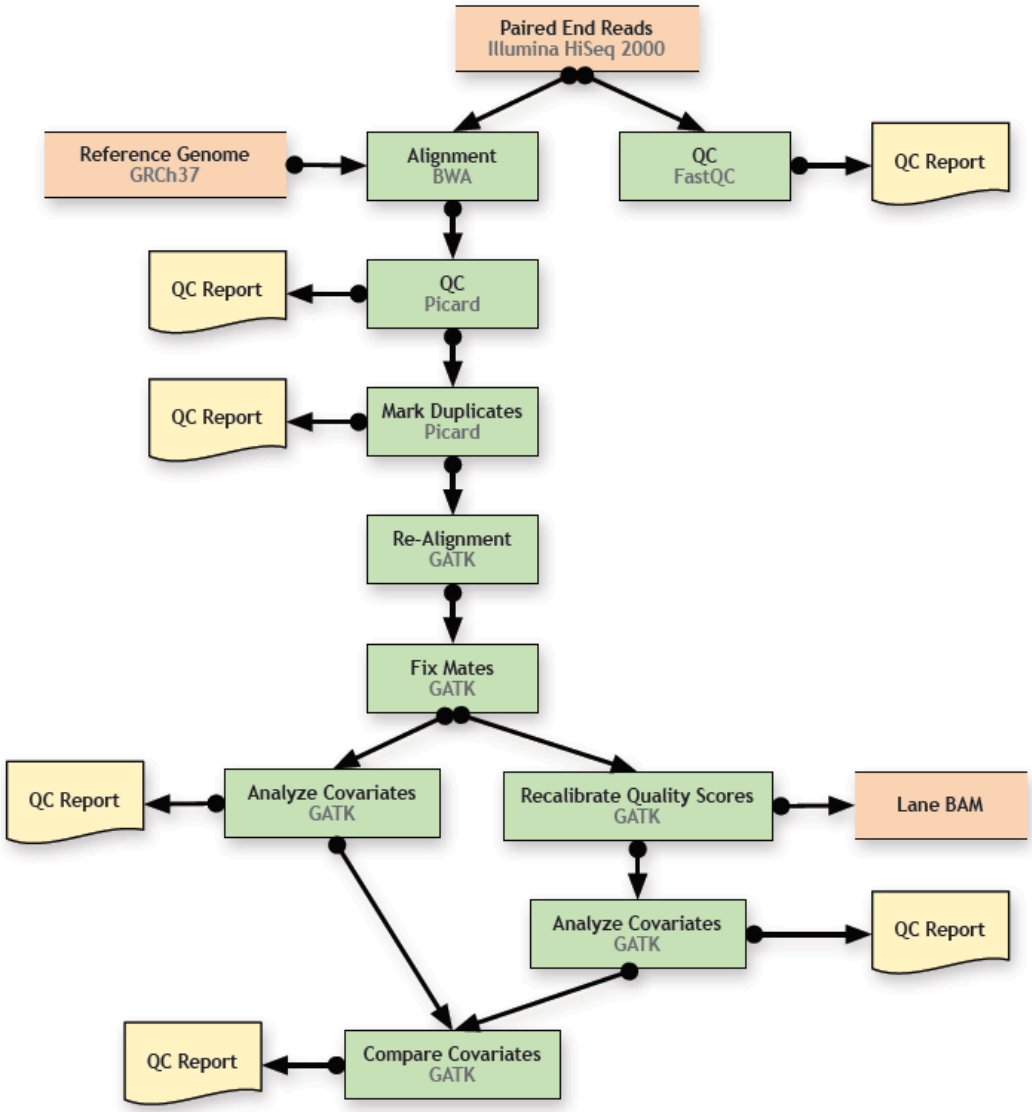
80 – 800 GB

umcg

molgenis

Alignment & SNP calling workflow

31 steps, ≥ 2 days per sample



- Input
  - Analysis protocols
  - Sample DNA data
  - Reference DNA data
- Analysis
  - Scripts are generated and executed
- Output
  - Aligned DNA and QC reports

umcg

molgenis

An analysis job (script) generated from a protocol

backend specific

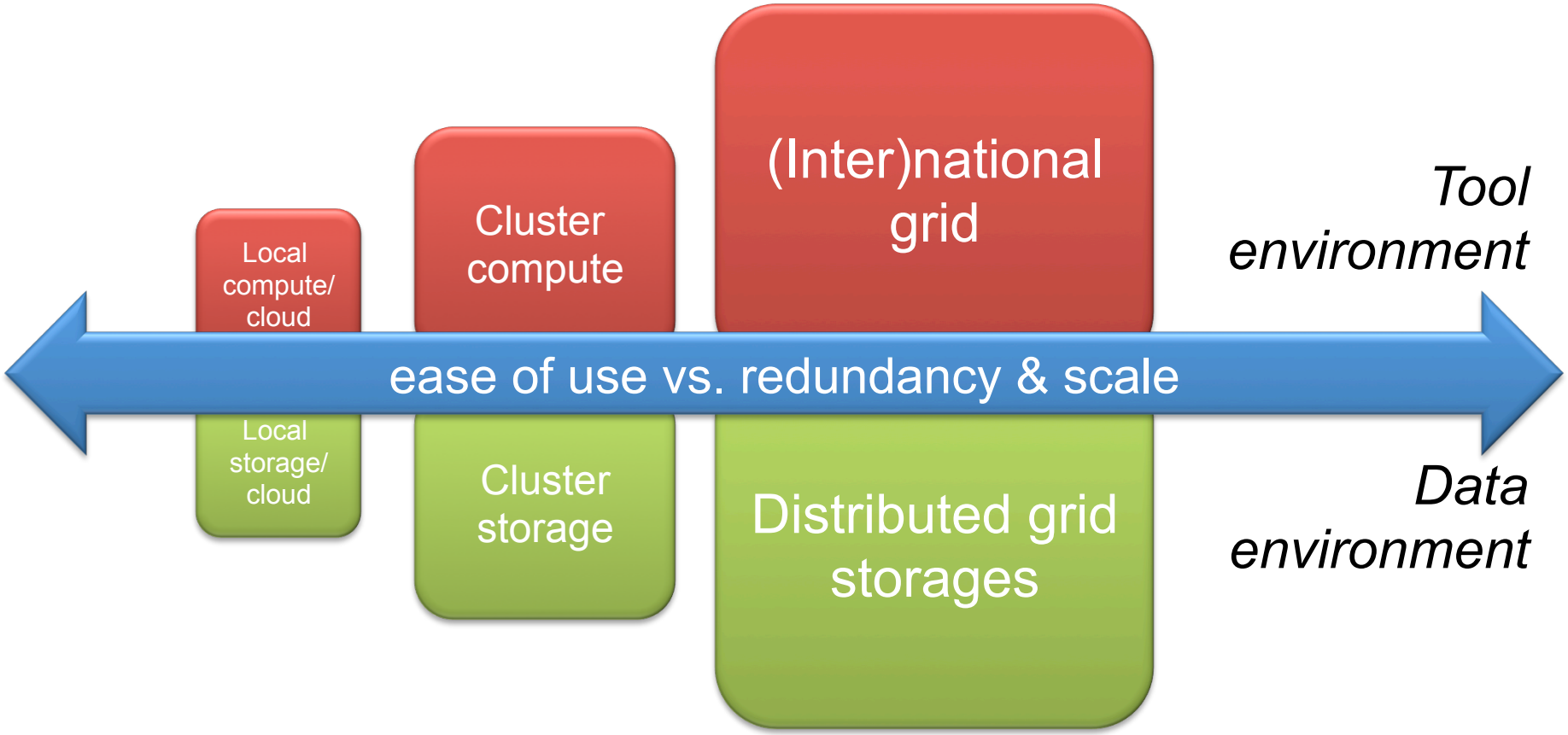
```
#!/bin/bash
#PBS -q test
#PBS -l nodes=1:ppn=4
#PBS -l walltime=08:00:00
#PBS -l mem=6gb
#PBS -e $GCC/test_compute/projects/batch4/intermediate/test1/err/err_test1_BwaElement1A102a_FC81D90ABXX_L7.err
#PBS -o $GCC/test_compute/projects/batch4/intermediate/test1/out/out_test1_BwaElement1A102a_FC81D90ABXX_L7.out

mkdir -p $GCC/test_compute/projects/batch4/intermediate/test1/err
mkdir -p $GCC/test_compute/projects/batch4/intermediate/test1/out
printf "test1_BwaElement1A102a_FC81D90ABXX_L7_started " >>$GCC/test_compute/projects/batch4/intermediate/test1/log_test1.txt
date "+DATE: %m/%d/%y%tTIME: %H:%M:%S" >>$GCC/test_compute/projects/batch4/intermediate/test1/log_test1.txt
date "+start time: %m/%d/%y%t %H:%M:%S" >>$GCC/test_compute/projects/batch4/intermediate/test1/
test1_BwaElement1A102a_FC81D90ABXX_L7.txt
echo running on node: `hostname` >>$GCC/test_compute/projects/batch4/intermediate/test1/
test1_BwaElement1A102a_FC81D90ABXX_L7.txt
```

analysis specific

```
/target/gpfs2/gcc/tools//bwa-0.5.8c_patched/bwa aln \
/target/gpfs2/gcc/resources/hg19/indices/human_g1k_v37.fa \
$GCC/test_compute/projects/batch4/rawdata/110121_I288_FC81D90ABXX_L7_HUMrutRGADIAAPE_1.fq.gz \
-t 4 \
-f $GCC/test_compute/projects/batch4/intermediate/A102a_110121_I288_FC81D90ABXX_L7_HUMrutRGADIAAPE_1.fq.gz.sai
```

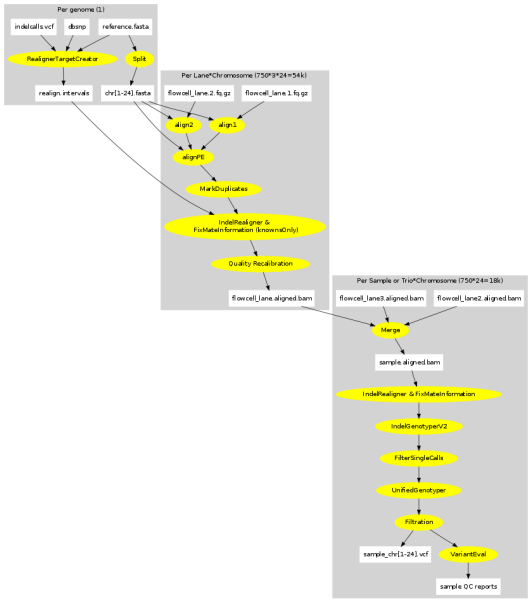
```
printf "test1_BwaElement1A102a_FC81D90ABXX_L7_finished " >>$GCC/test_compute/projects/batch4/intermediate/test1/log_test1.txt
date "+finish time: %m/%d/%y%t %H:%M:%S" >>$GCC/test_compute/projects/batch4/intermediate/test1/
test1_BwaElement1A102a_FC81D90ABXX_L7.txt
date "+DATE: %m/%d/%y%tTIME: %H:%M:%S" >>$GCC/test_compute/projects/batch4/intermediate/test1/log_test1.txt
```

Computational environments



Bio-workflow complexity

- Many analysis steps
  - Many analysis jobs
  - Different analysis tools and their dependencies
- Large various data involved
- Heterogeneous resources



umcg

molgenis

Number of analysis jobs to show

- Imputation step in GoNL

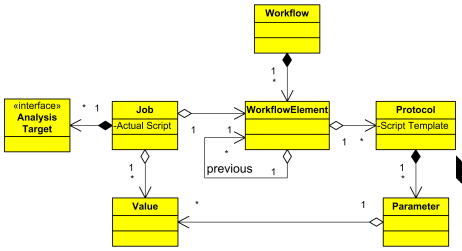
$$number\_jobs = \sum_{chr=1..22} \frac{Lenght_{chr}}{5\_megabase} * \frac{number\_of\_samples}{500}$$

$$10000 = 500 * \frac{10000}{500}$$

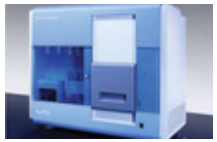
State of the art workflow visualizations

MOLGENIS software toolkit

Model (xml)



Workflow



NextGenSeq

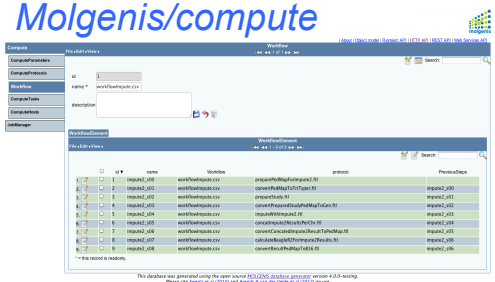
Model organisms

Generator (java)

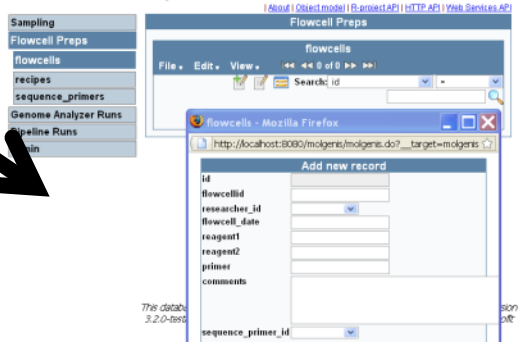
molgenis

Use (web)

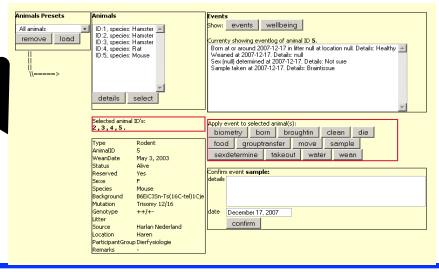
Molgenis/compute



Solexa Sequencer LIMS



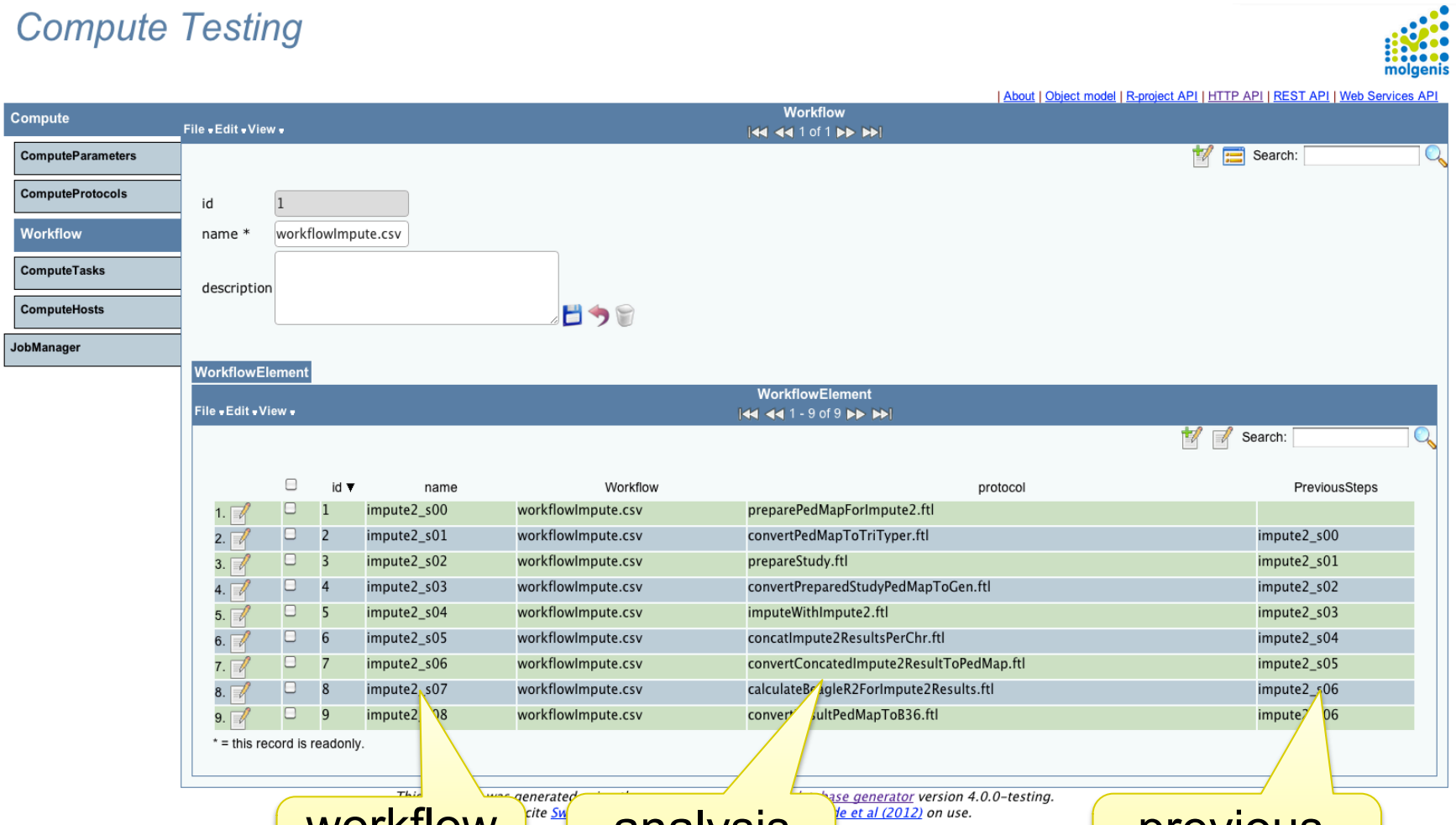
Animal Observatory



umcg

molgenis

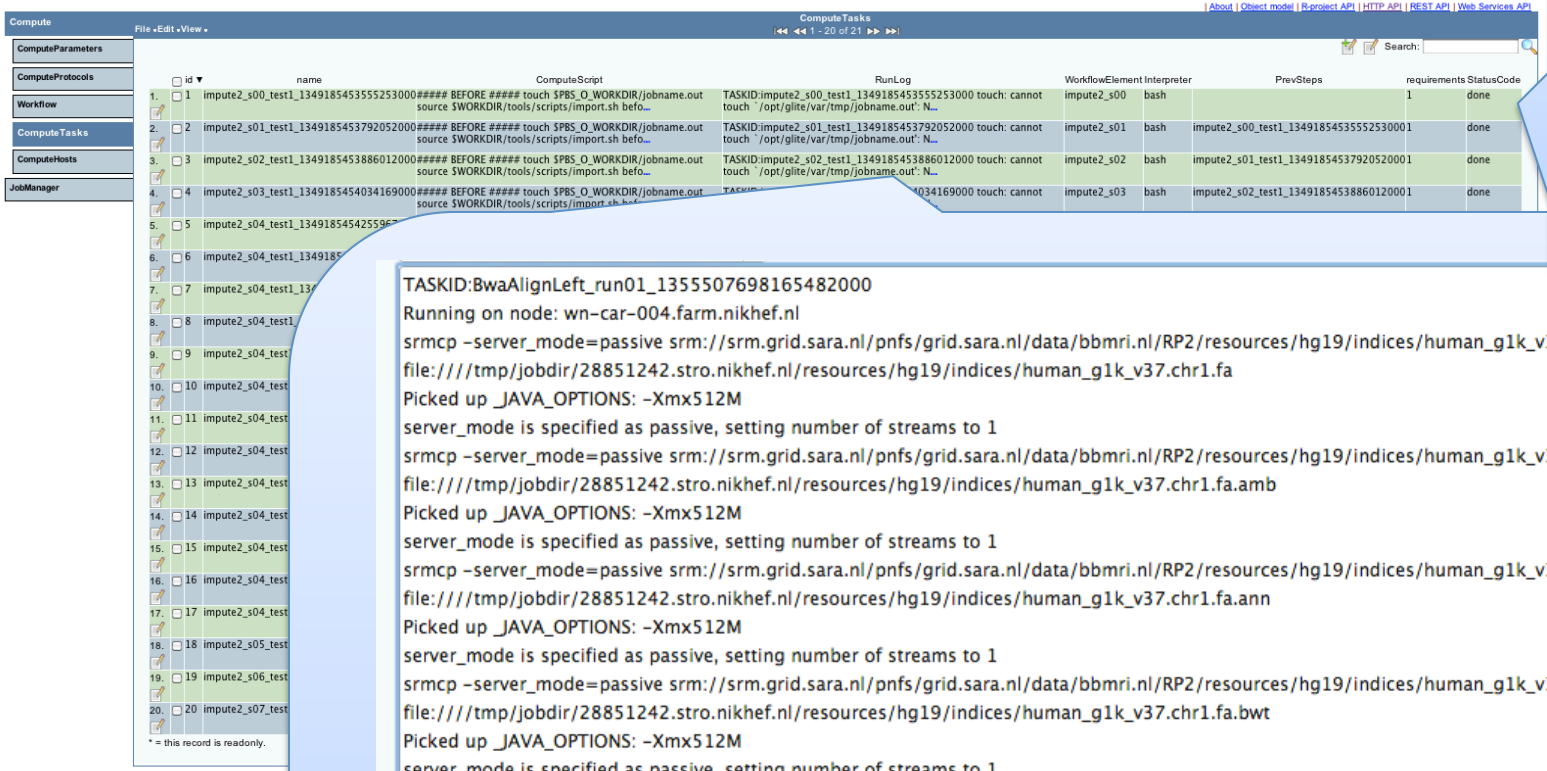
Workflow design view in the generated Molgenis web-UI

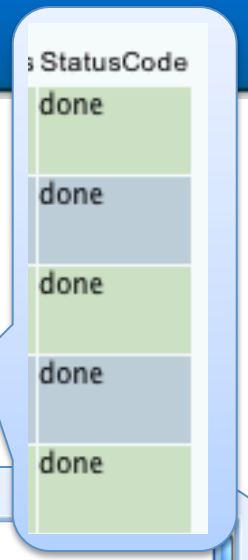


umcg

molgenis

Workflow run-time view (analysis jobs)

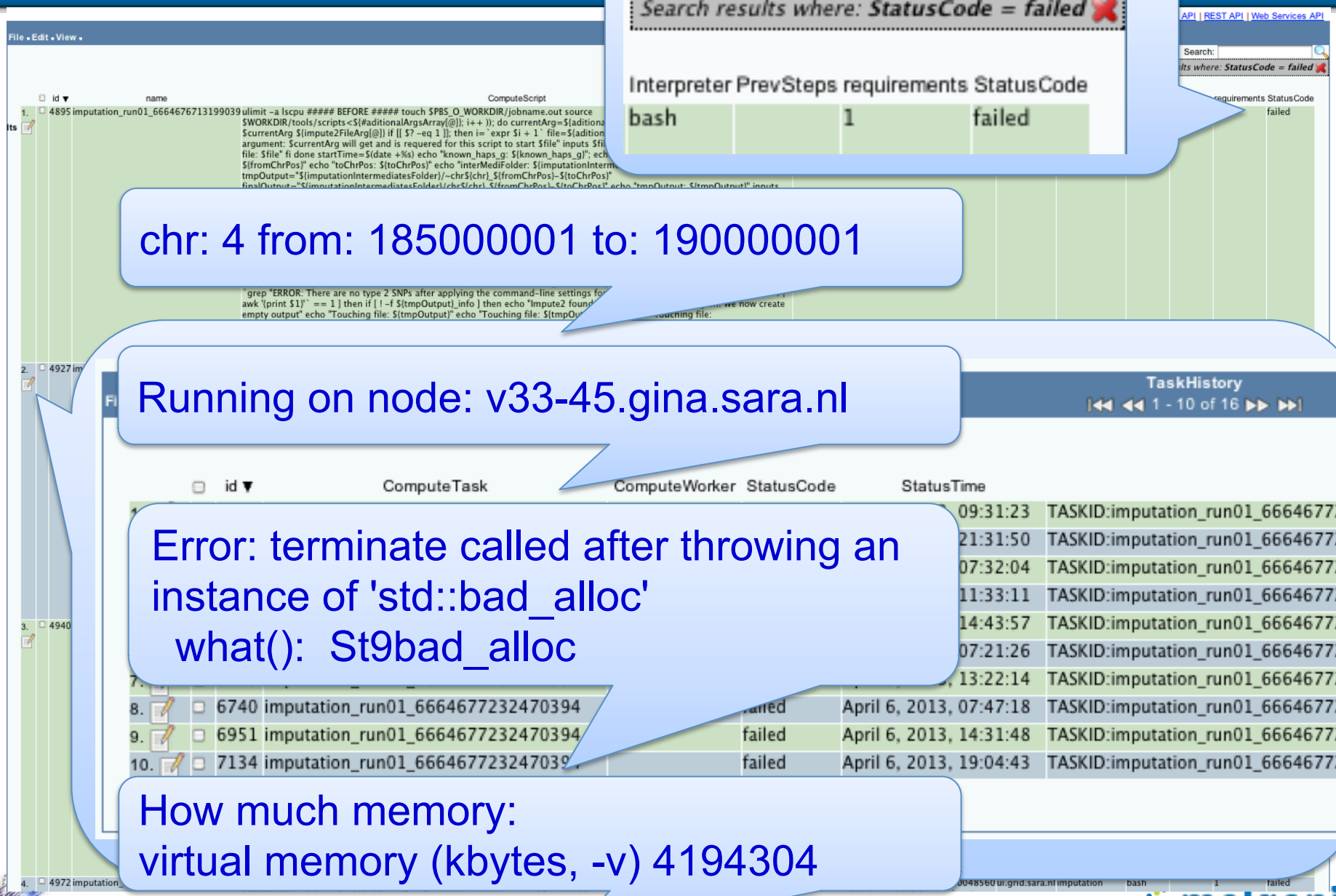




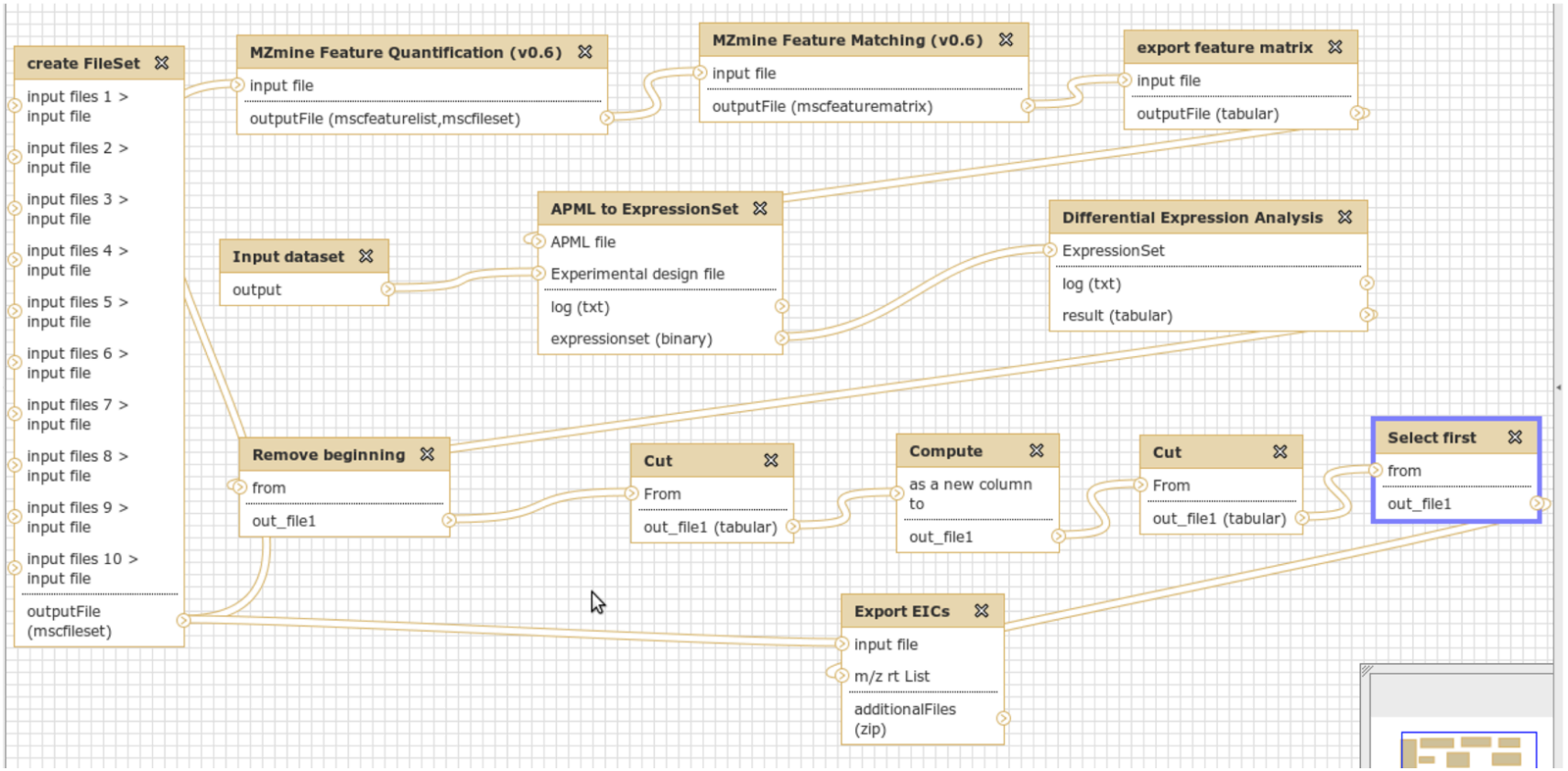
TASKID:BwaAlignLeft_run01_1355507698165482000
Running on node: wn-car-004.farm.nikhef.nl
srmcp -server_mode=passive srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2/resources/hg19/indices/human_g1k_v37.chr1.fa file:////tmp/jobdir/28851242.stro.nikhef.nl/resources/hg19/indices/human_g1k_v37.chr1.fa
Picked up _JAVA_OPTIONS: -Xmx512M
server_mode is specified as passive, setting number of streams to 1
srmcp -server_mode=passive srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2/resources/hg19/indices/human_g1k_v37.chr1.fa.amb file:////tmp/jobdir/28851242.stro.nikhef.nl/resources/hg19/indices/human_g1k_v37.chr1.fa.amb
Picked up _JAVA_OPTIONS: -Xmx512M
server_mode is specified as passive, setting number of streams to 1
srmcp -server_mode=passive srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2/resources/hg19/indices/human_g1k_v37.chr1.fa.ann file:////tmp/jobdir/28851242.stro.nikhef.nl/resources/hg19/indices/human_g1k_v37.chr1.fa.ann
Picked up _JAVA_OPTIONS: -Xmx512M
server_mode is specified as passive, setting number of streams to 1
srmcp -server_mode=passive srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2/resources/hg19/indices/human_g1k_v37.chr1.fa.bwt file:////tmp/jobdir/28851242.stro.nikhef.nl/resources/hg19/indices/human_g1k_v37.chr1.fa.bwt
Picked up _JAVA_OPTIONS: -Xmx512M
server_mode is specified as passive, setting number of streams to 1
srmcp -server_mode=passive srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2/resources/hg19/indices/human_g1k_v37.chr1.fa.fai file:////tmp/jobdir/28851242.stro.nikhef.nl/resources/hg19/indices/human_g1k_v37.chr1.fa.fai
Picked up _JAVA_OPTIONS: -Xmx512M
server_mode is specified as passive, setting number of streams to 1
srmcp -server_mode=passive srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2/resources/hg19/indices/human_g1k_v37.chr1.fa.pac file:////tmp/jobdir/28851242.stro.nikhef.nl/resources/hg19/indices/human_g1k_v37.chr1.fa.pac
Picked up _JAVA_OPTIONS: -Xmx512M
server_mode is specified as passive, setting number of streams to 1
srmcp -server_mode=passive srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2/resources/hg19/indices/human_g1k_v37.chr1.fa.rbwt file:////tmp/jobdir/28851242.stro.nikhef.nl/resources/hg19/indices/human_g1k_v37.chr1.fa.rbwt
Picked up _JAVA_OPTIONS: -Xmx512M
server_mode is specified as passive, setting number of streams to 1

umcg

Failed jobs overview



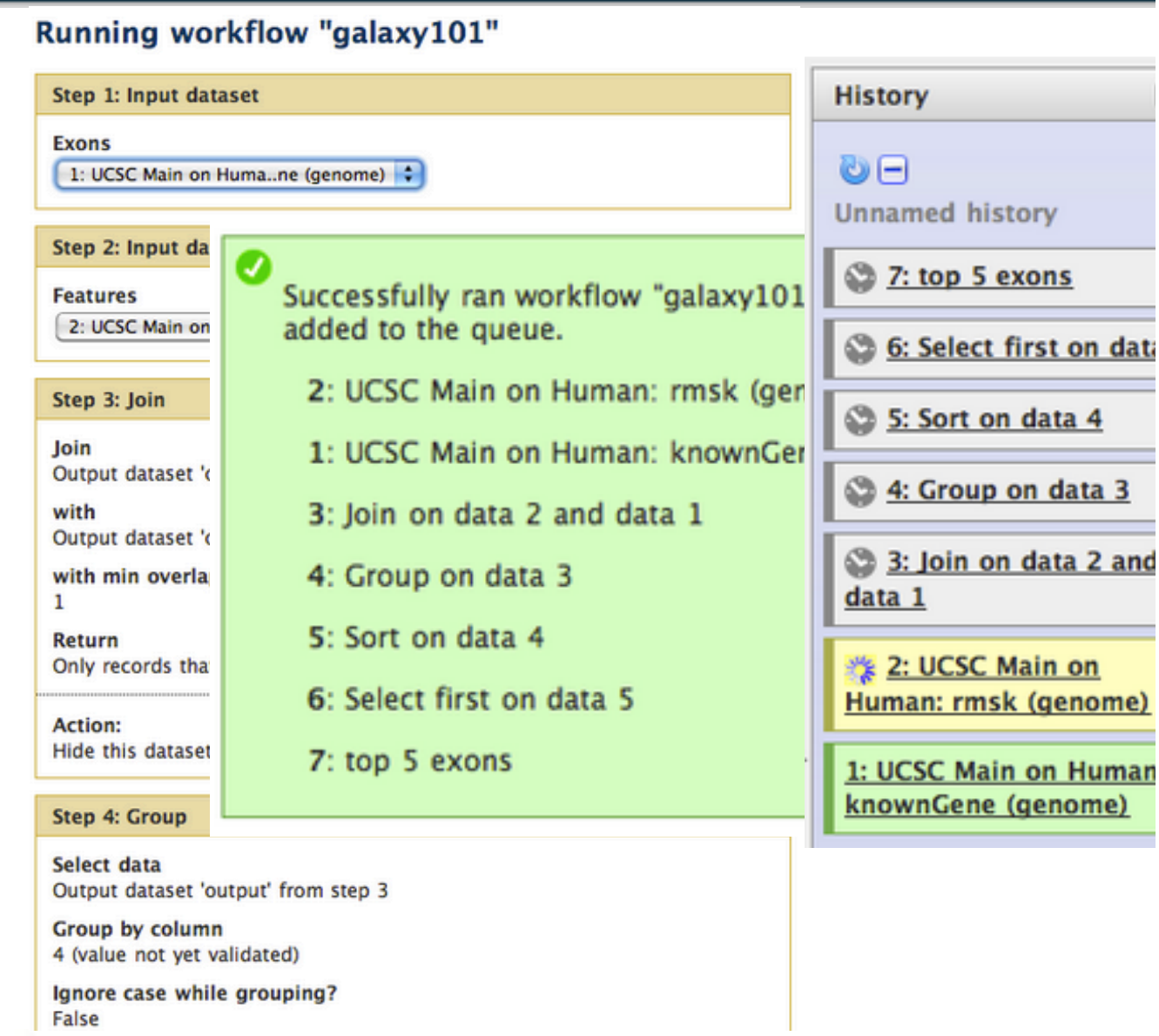
Galaxy design view

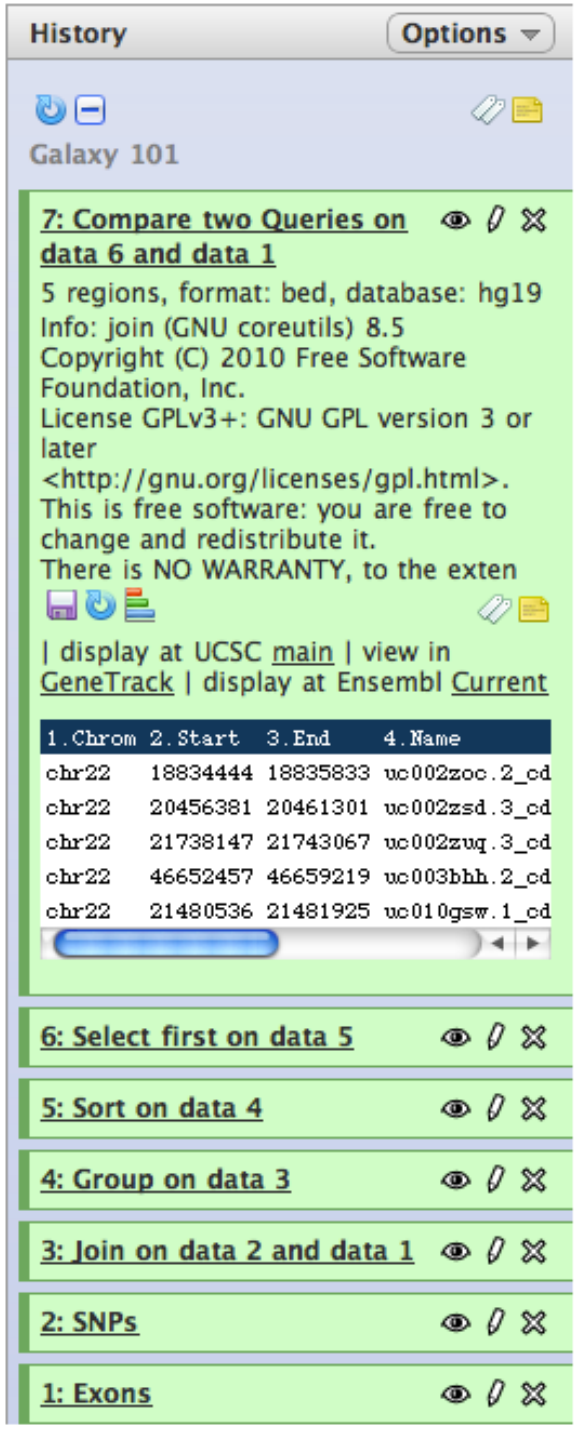


umcg

molgenis

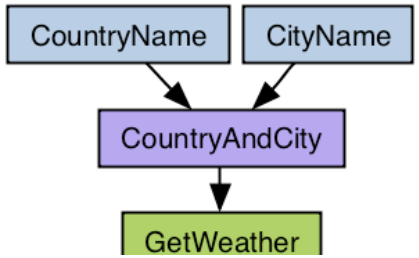
Galaxy run-time view

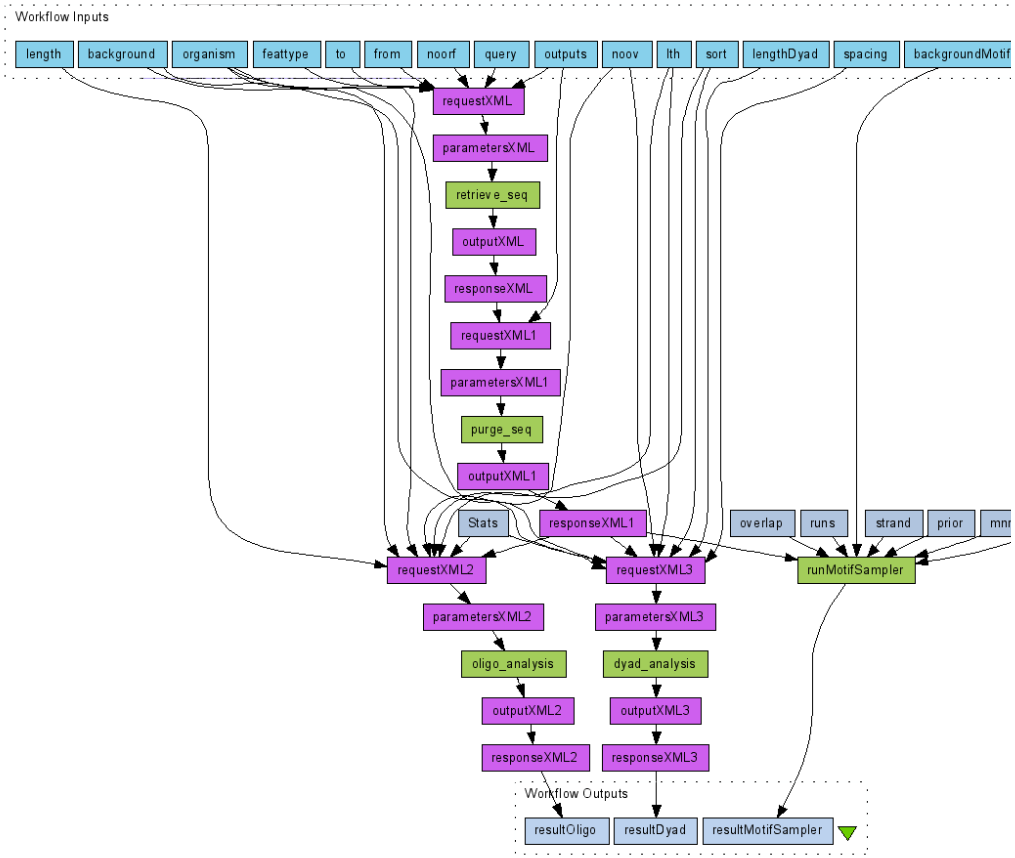


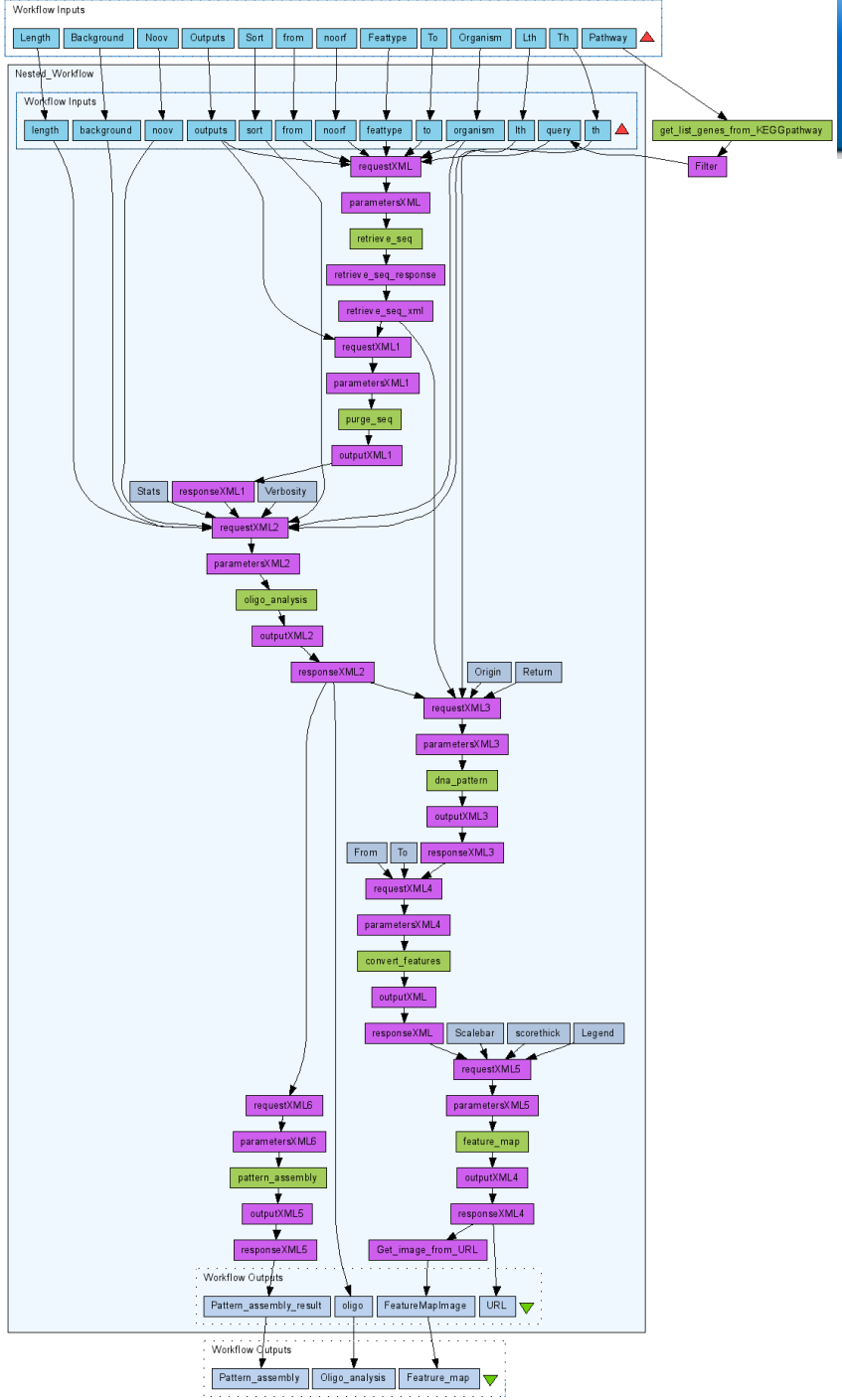


umcg

Taverna



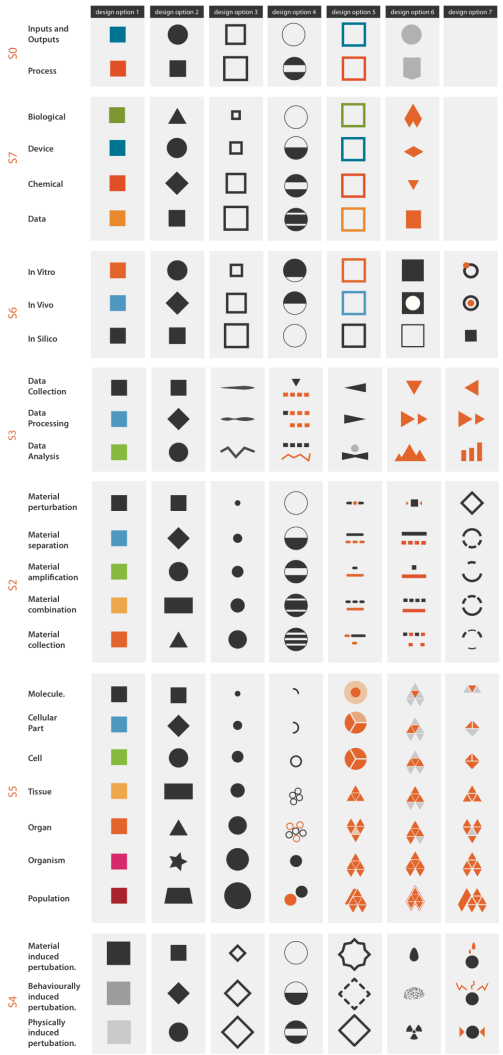




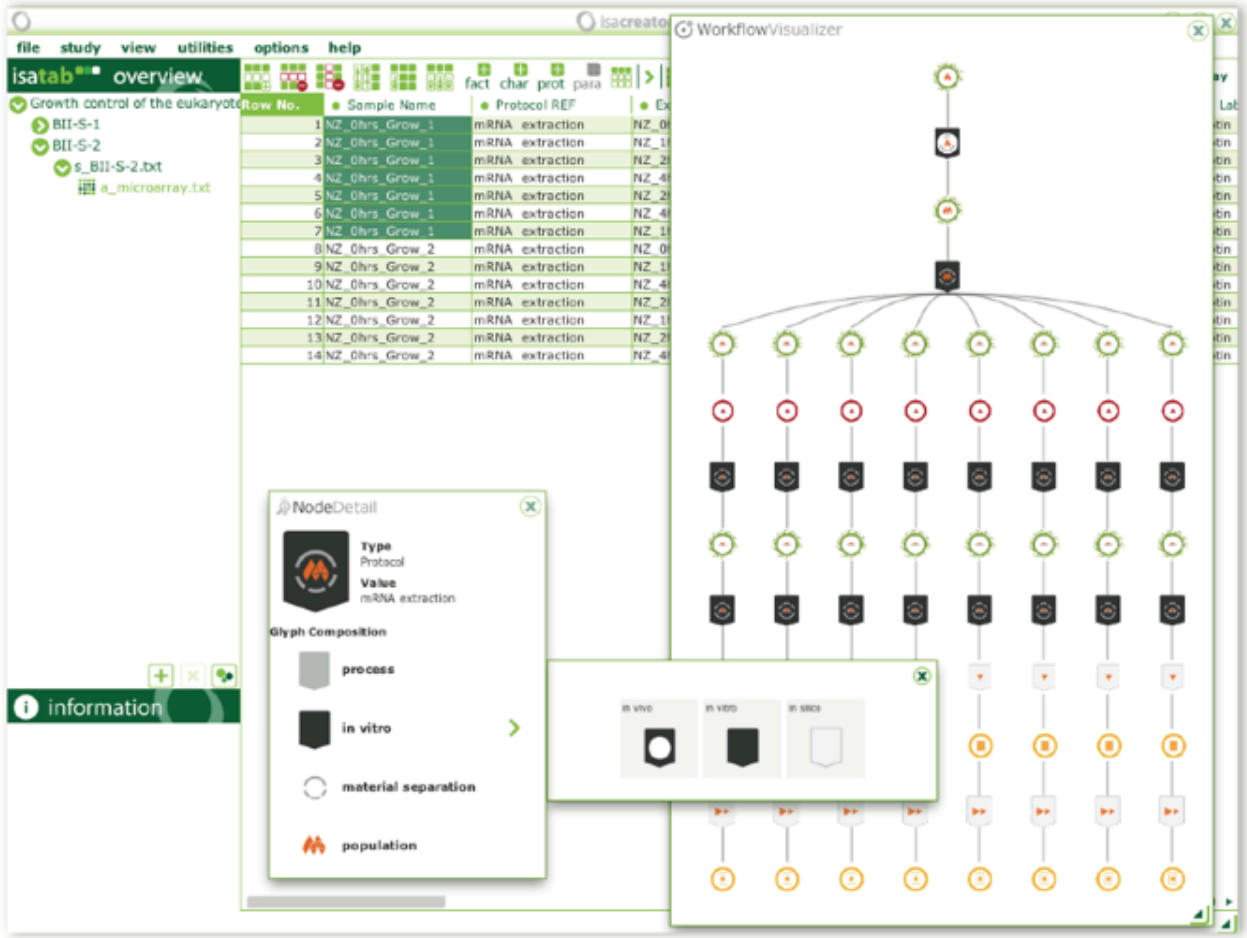
umcg

Glyph-based visualization of bio-workflows (E. Maguire *et al.*)

Classification



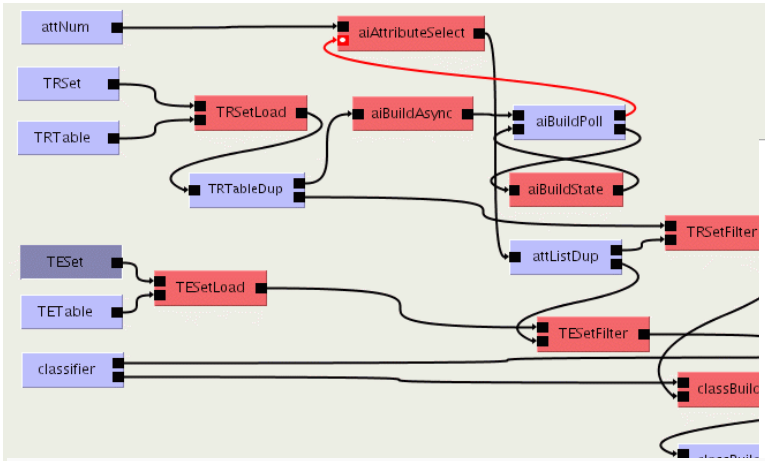
Visualization

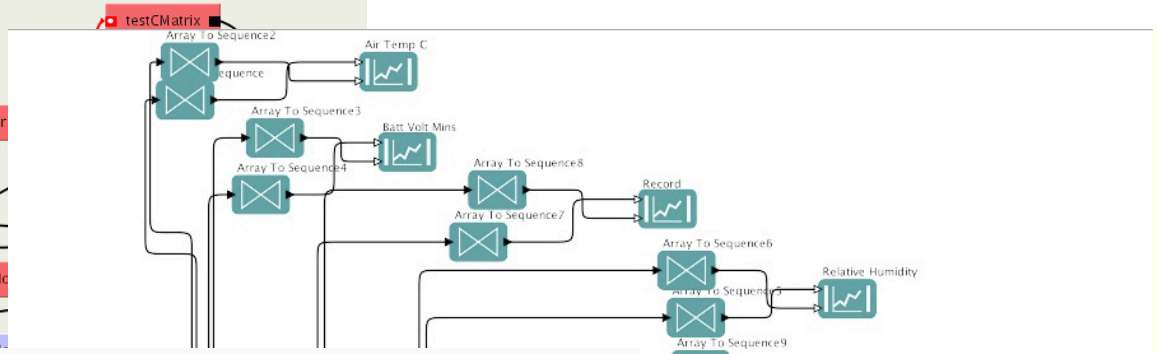


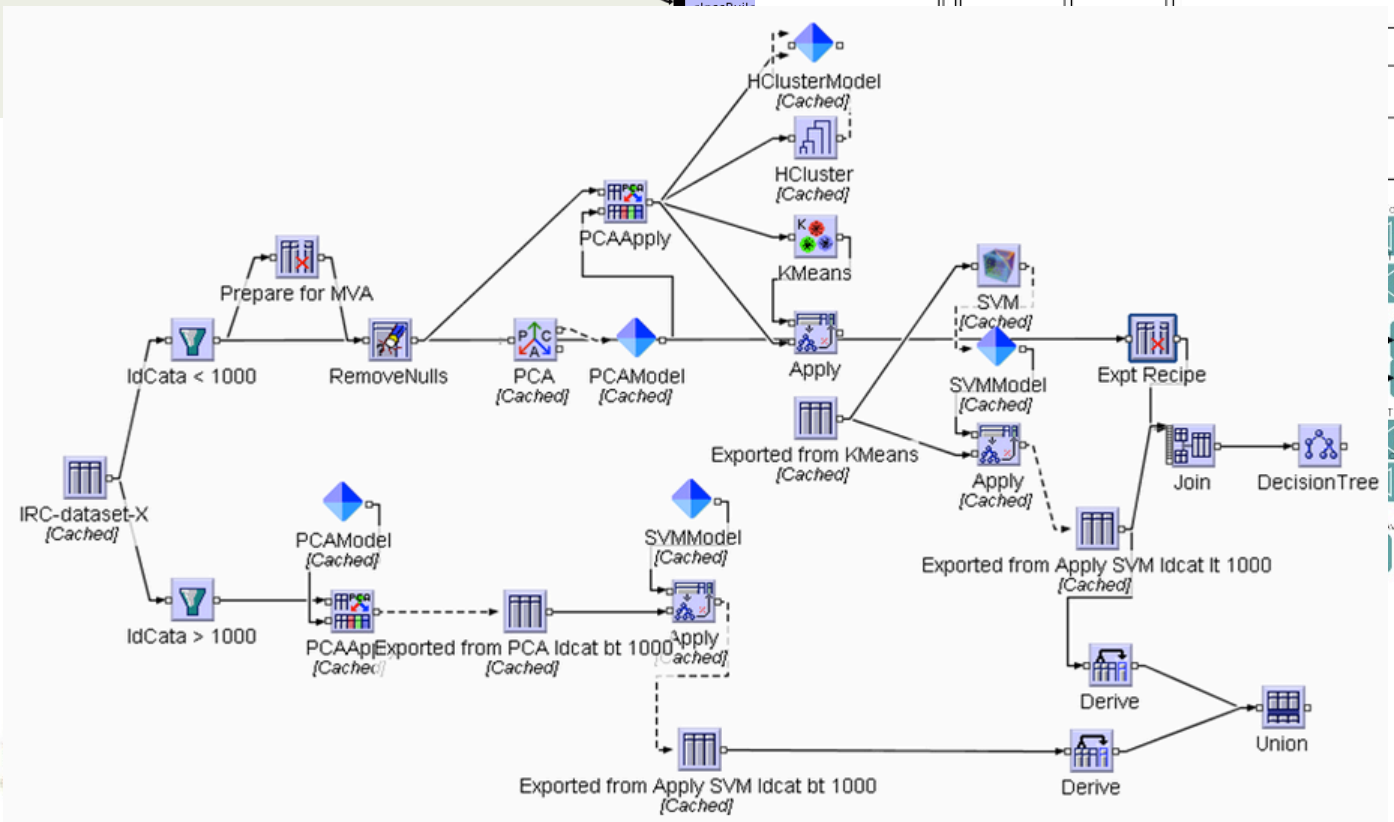
umcg

molgenis

Triana/Kepler/Inforsense



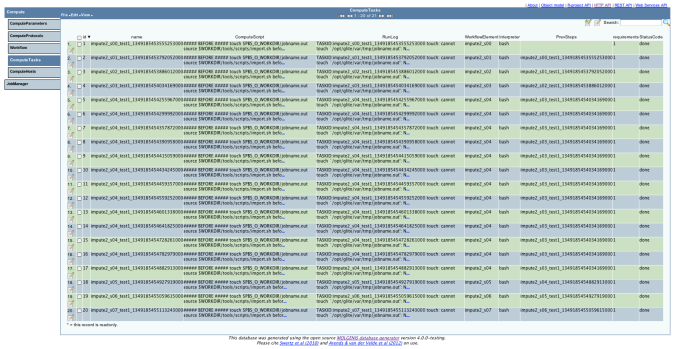




molgenis

How workflow visualization can be improved (techniques taken from software visualization)

Why workflow visualization is complex



- Workflow complexity leads to
  - Large graphs/tables/logs
  - Lots of attributes
  - Workflow changes and refinements

umcg

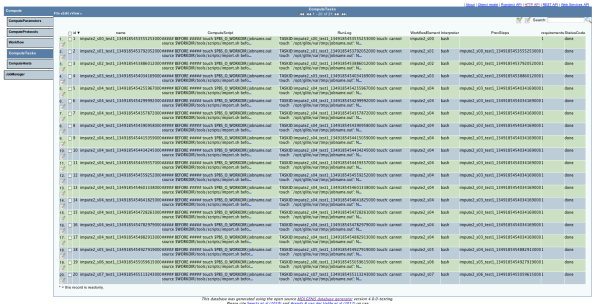
molgenis

Requirements for visual analytics

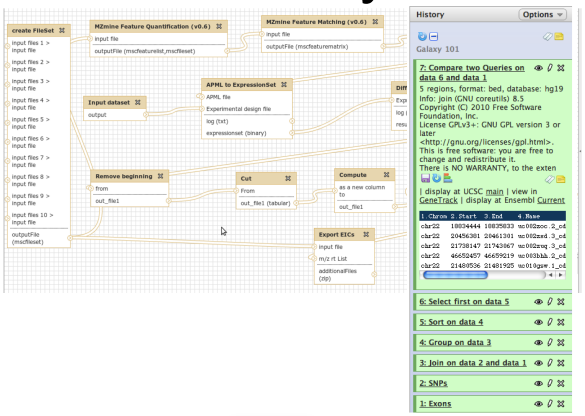
- Different views
  - Design view: structural overview and zoom in
  - Run-time execution view: history and zoom in
  - Behavior view: how parameters influence results data/quality
  - Evolution view: what elements were introduced/removed

Approach

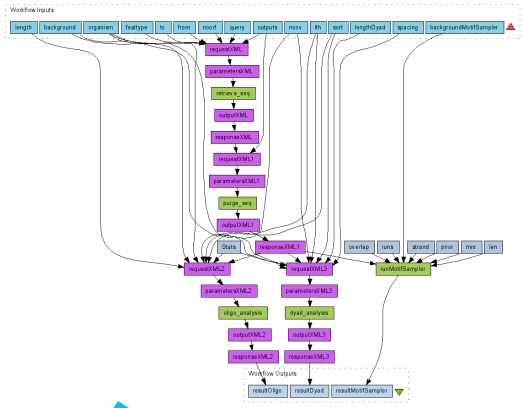
Molgenis

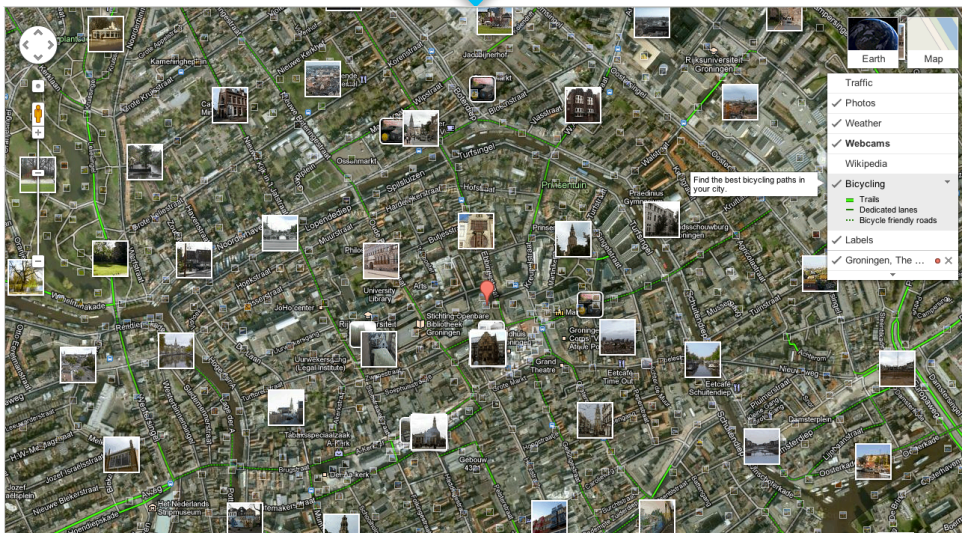


Galaxy



Taverna





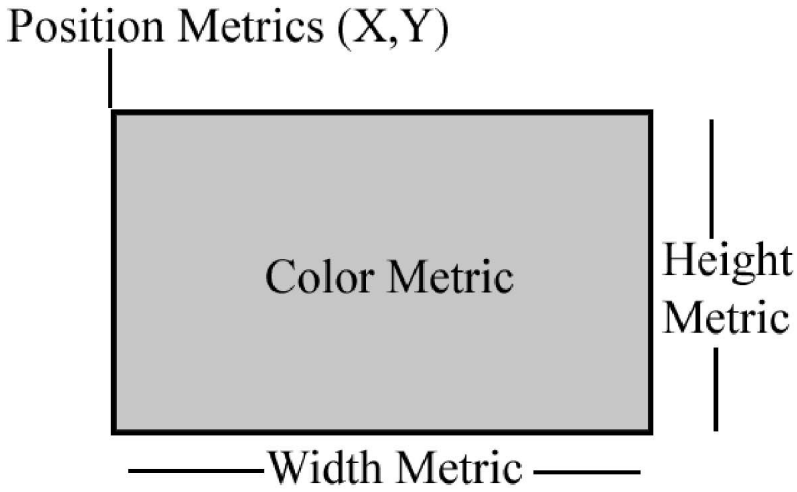
umcg

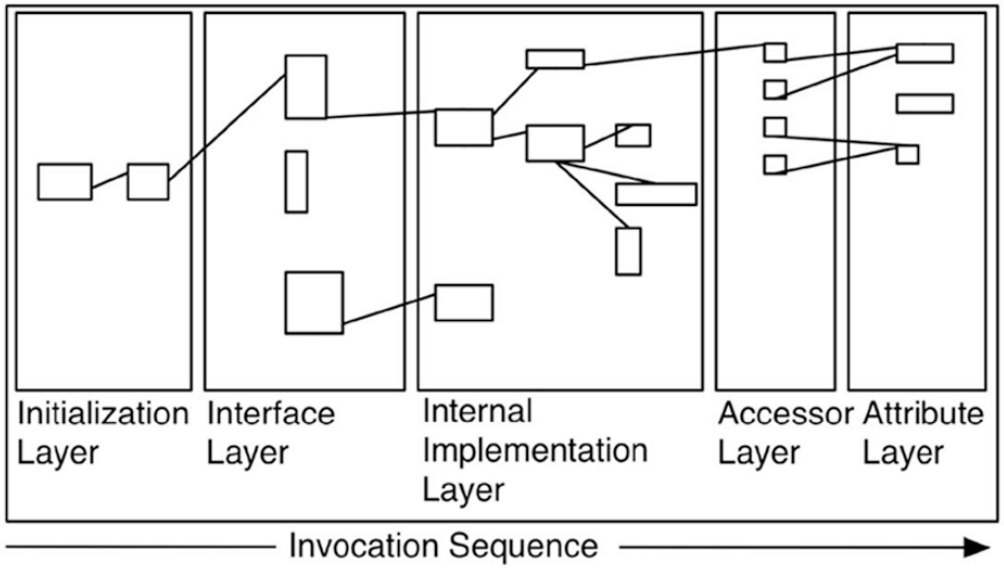
molgenis

Starting point

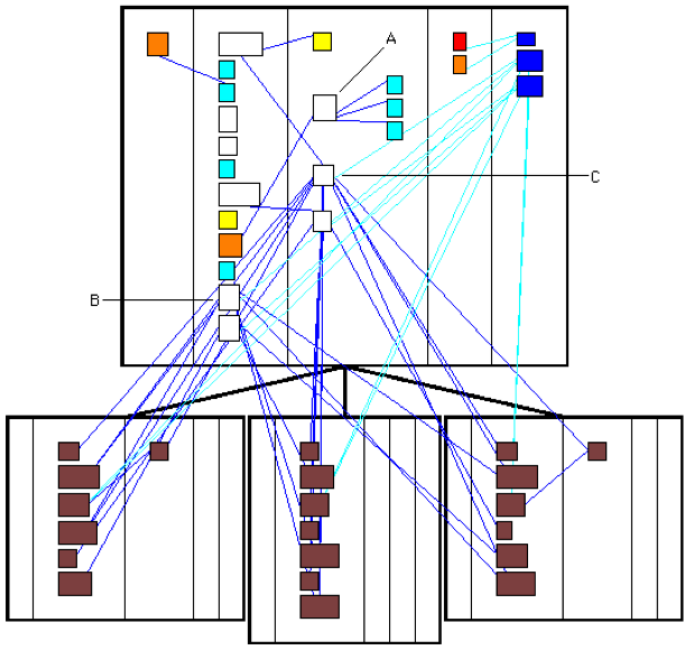
- Adding quality metrics to the workflow structure
  - # of parallel executions
  - sizes of input/output data and quality scores
  - analysis execution time
  - tools and their dependencies
  - resources required (*i.e.* CPU, RAM)

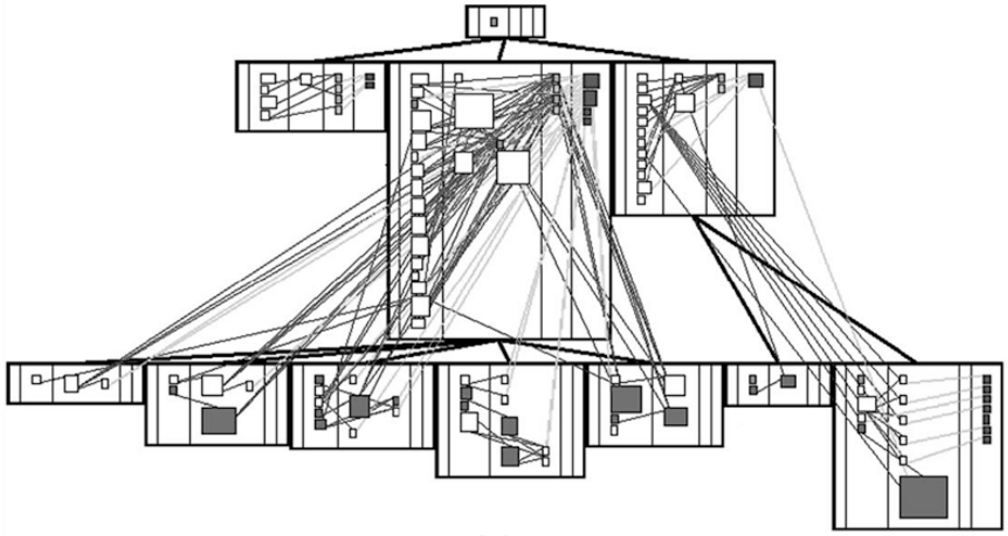
Adding quality metrics to the structural representation





a)

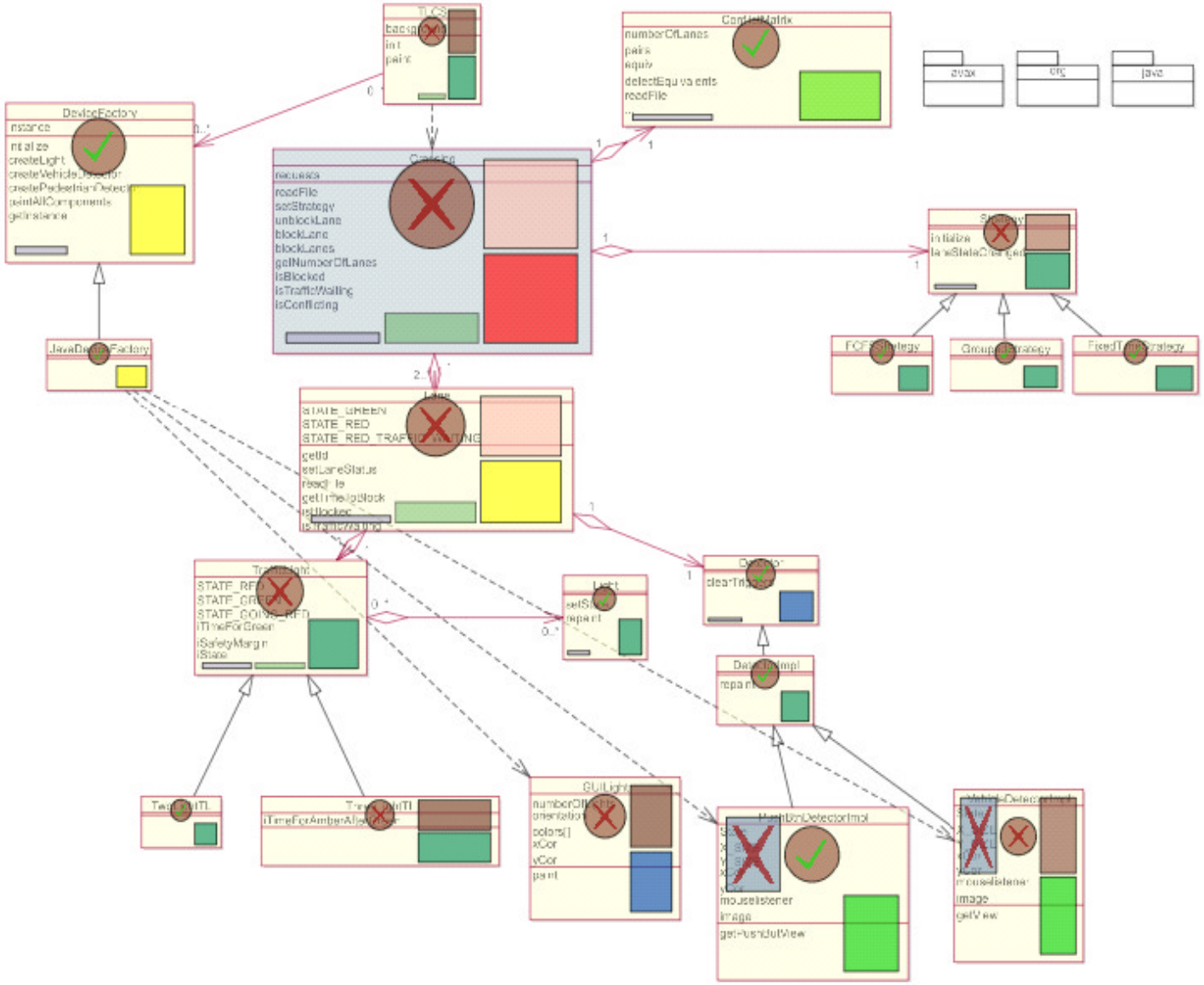




b)

Lanza et al.

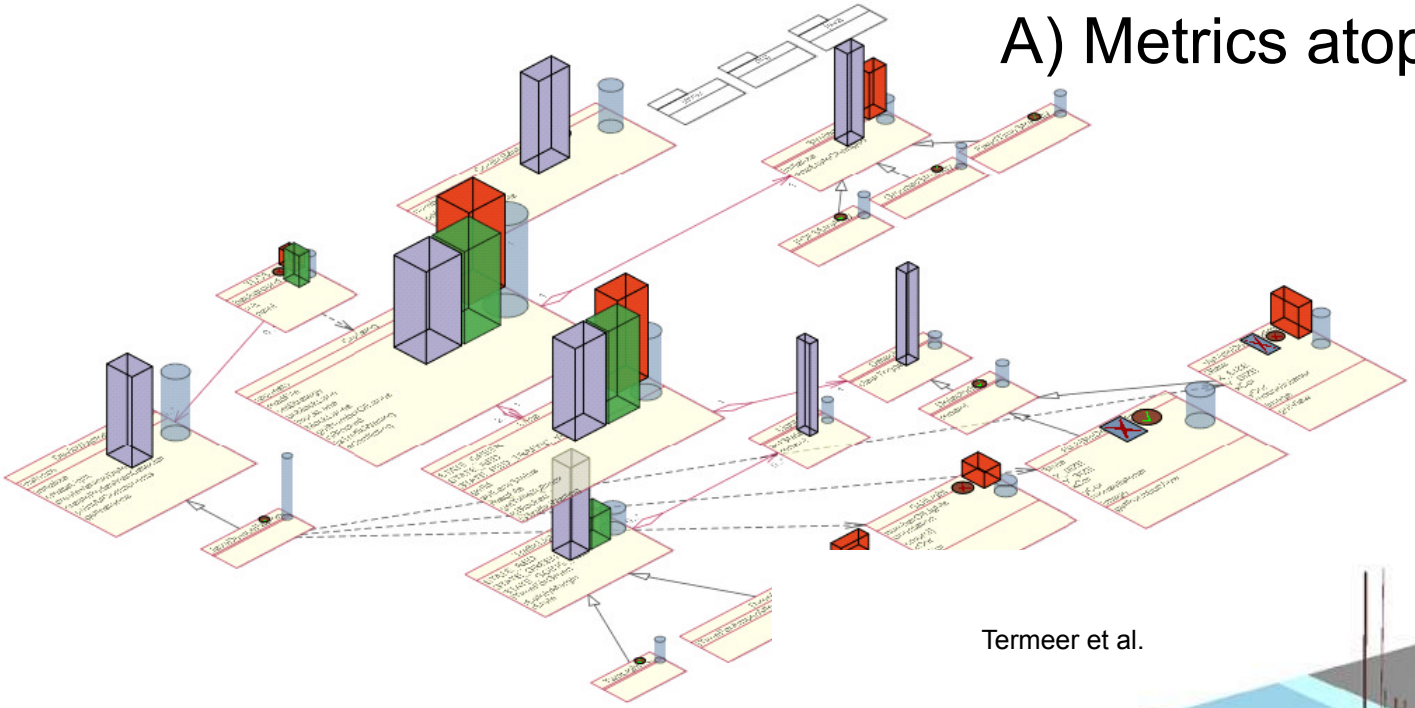
Adding quality metrics atop of a given diagram



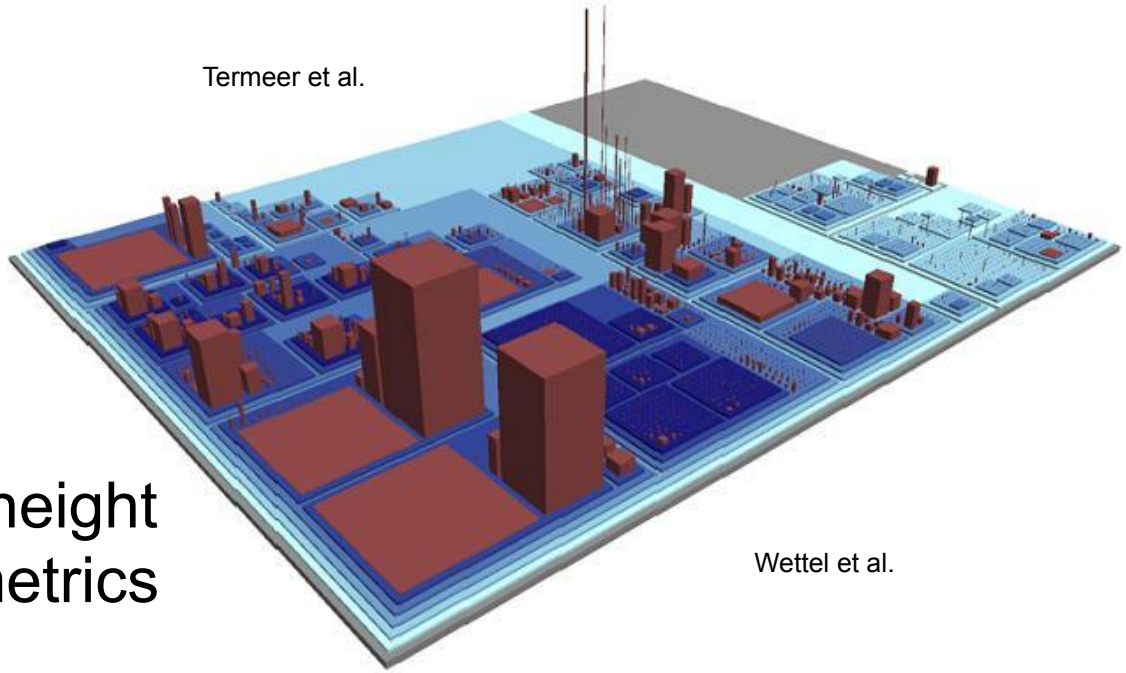
Termeer et al.

Adding metrics in 3D

A) Metrics atop of the diagram



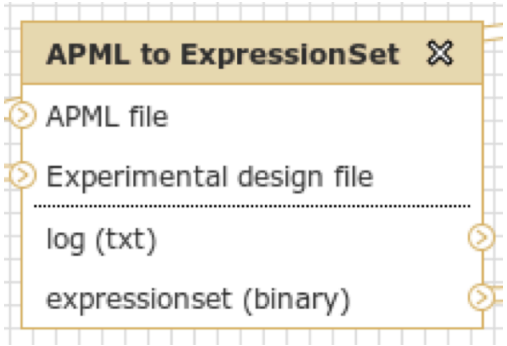
Termeer et al.



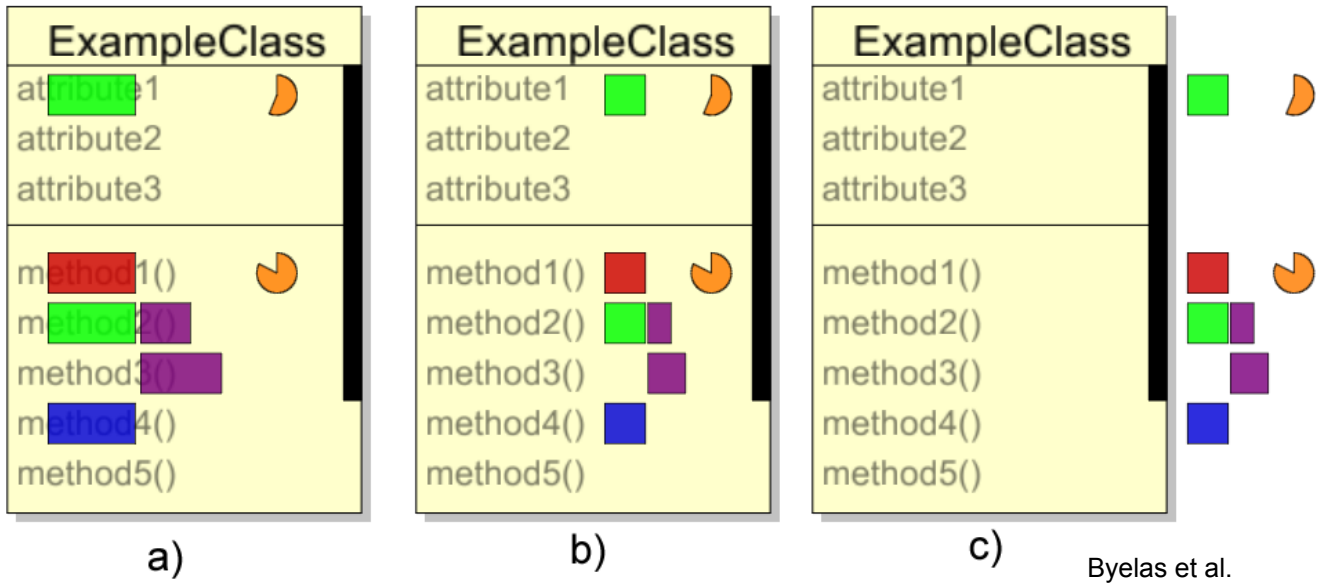
B) Using sizes and height to show metrics

Wettel et al.

Adding metrics on the level of entity attributes



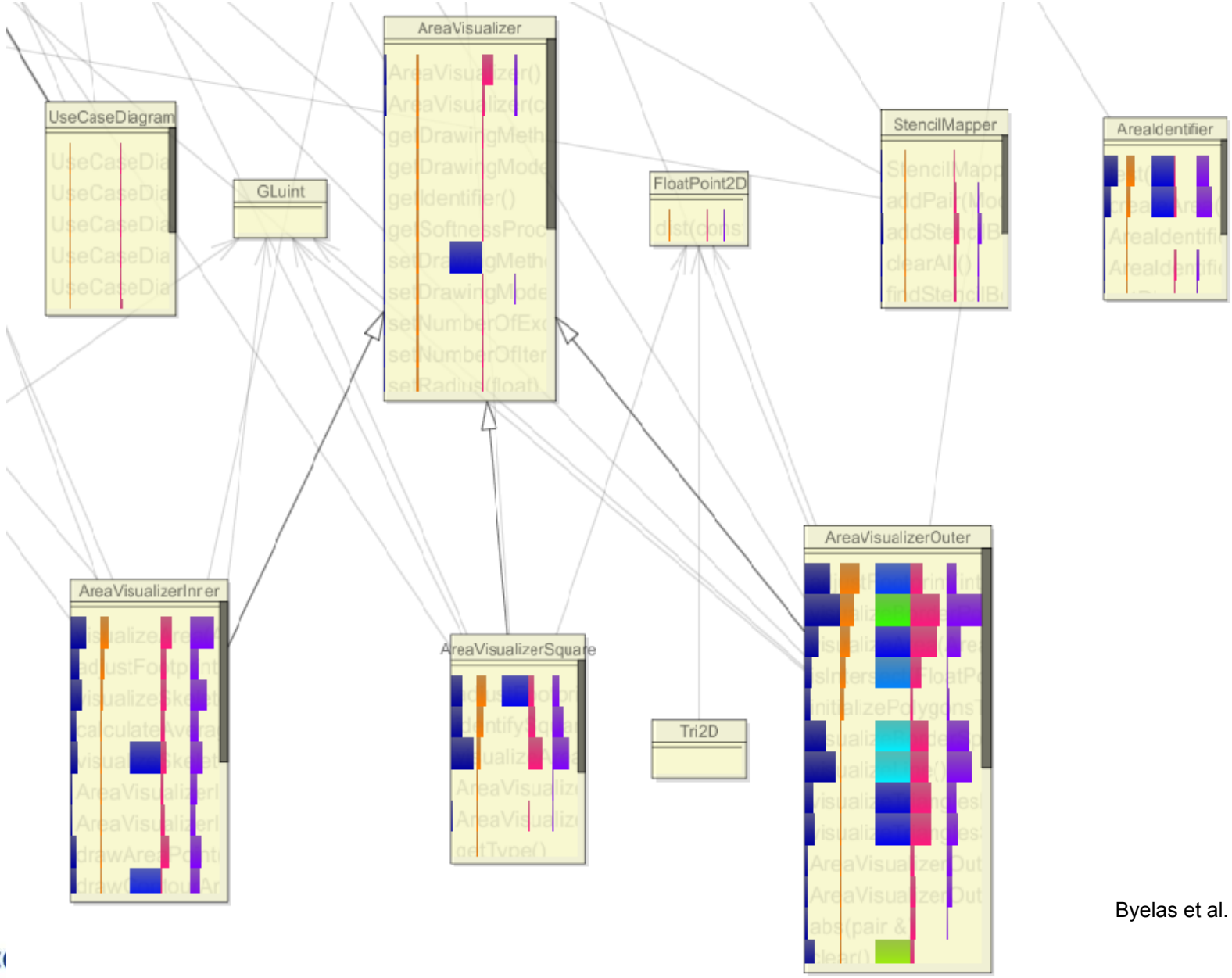
A) Galaxy workflow element



Byelas et al.

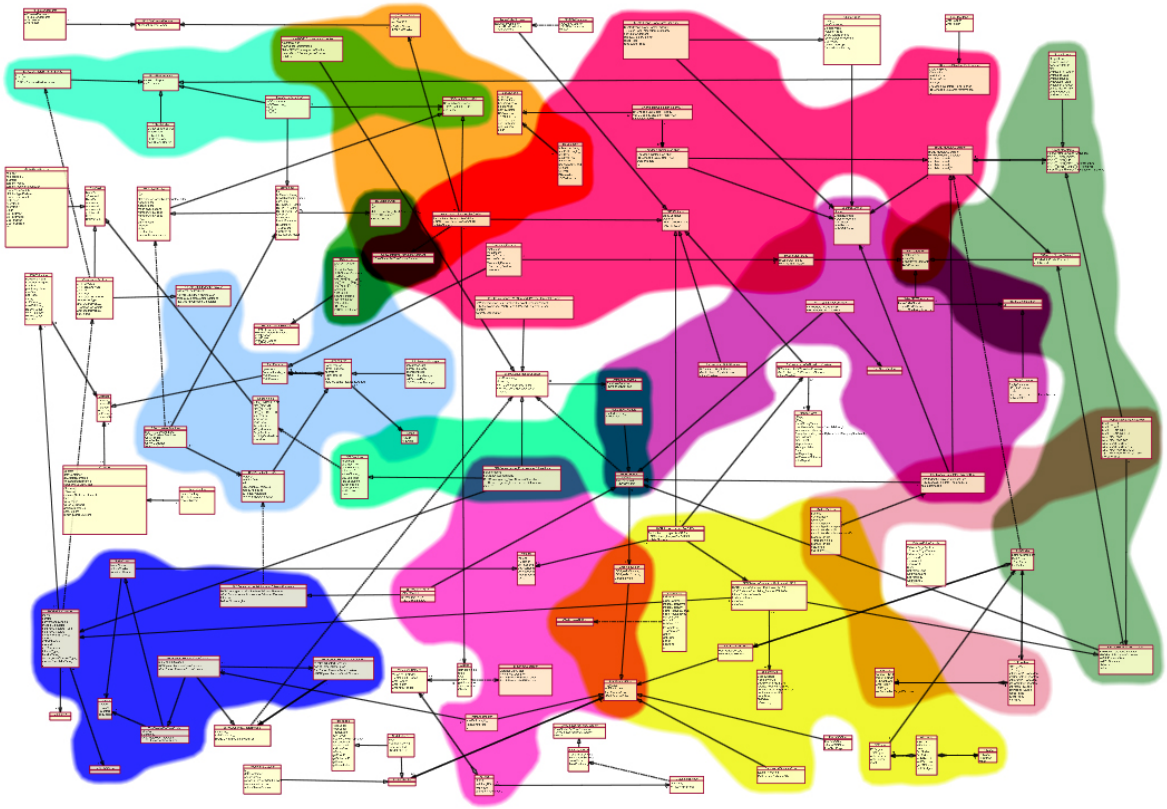
B) Adding metrics for UML class attributes

UML class diagram with 5 attribute level metrics



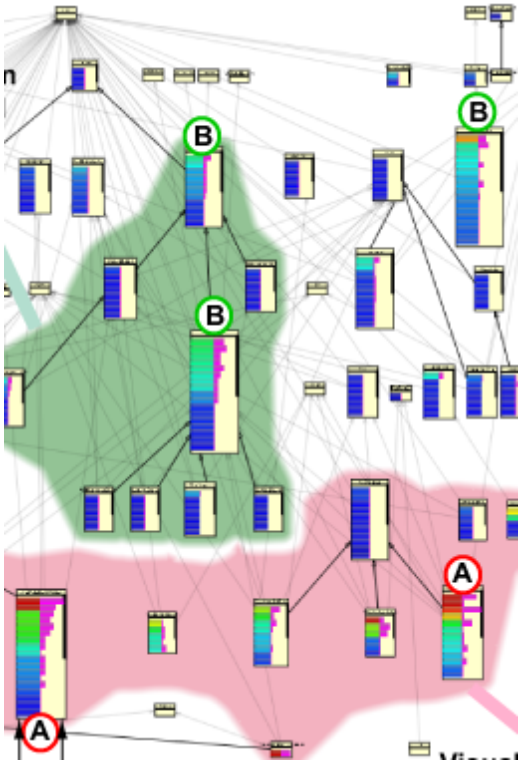
Byelas et al.

Metrics on different levels of detail



Byelas et al.

A) Showing workflow areas

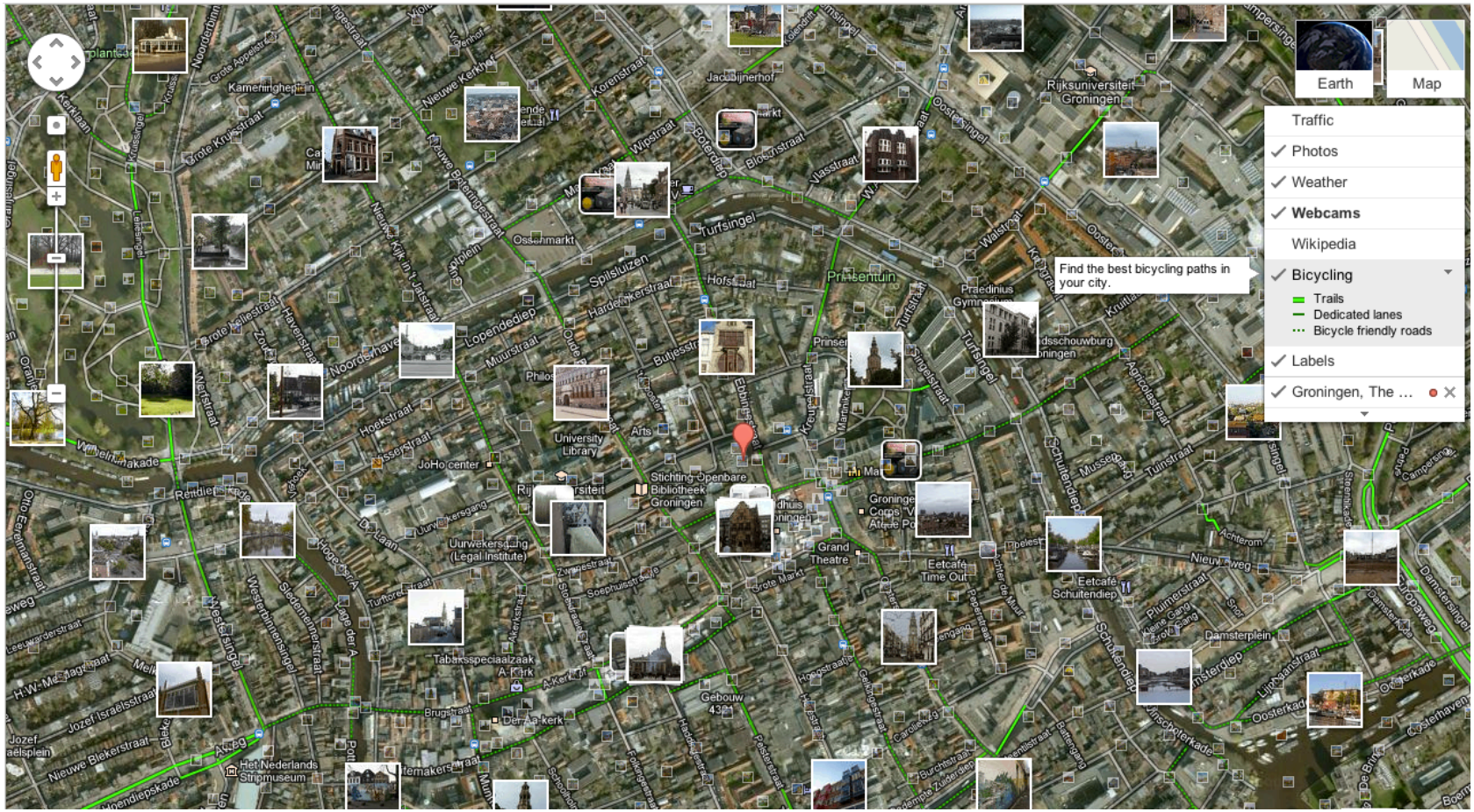


Byelas et al.

B) Metrics on different levels

Conclusion

Multi-scale visualization for workflow structure



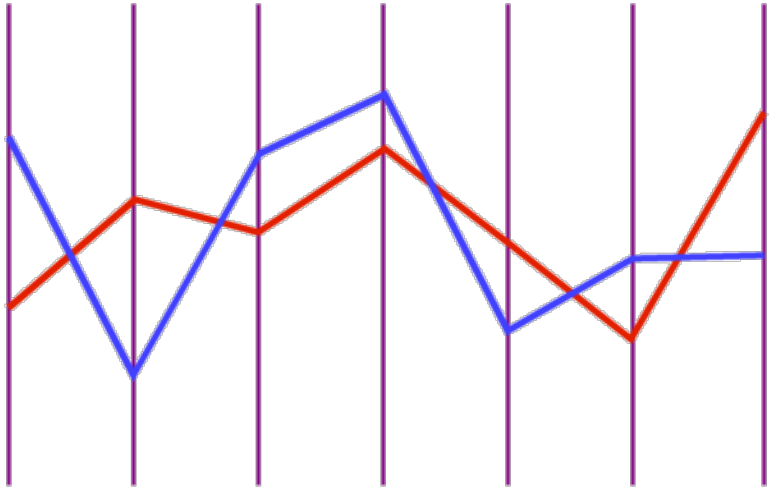
Google Maps

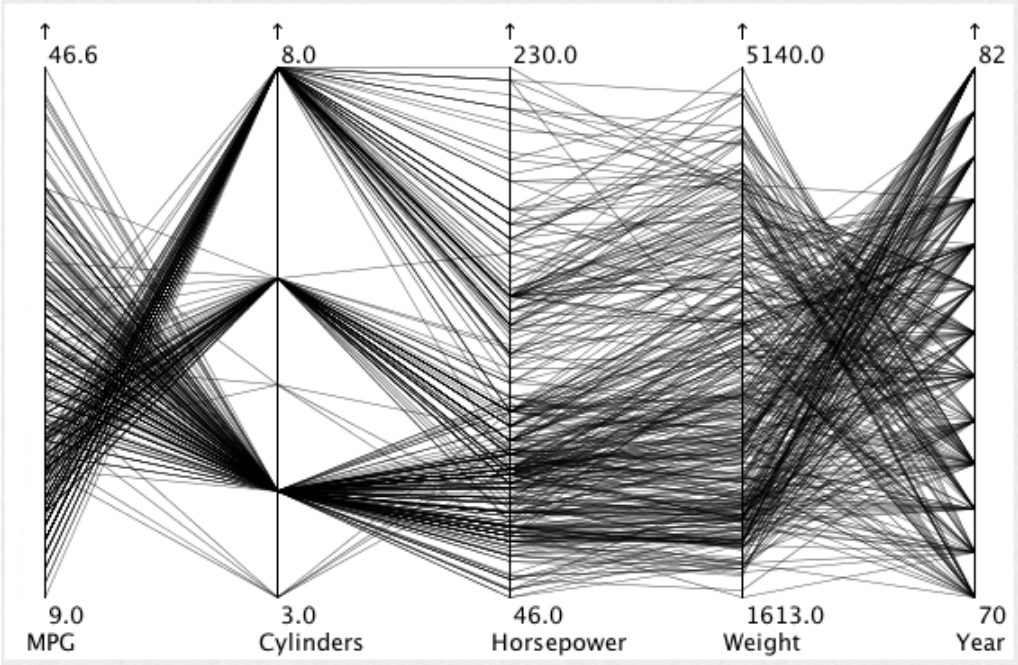
- Can a workflow be shown in the same way?
- Easy to hide/highlight different aspects

umcg

molgenis

Workflow behavior visualization



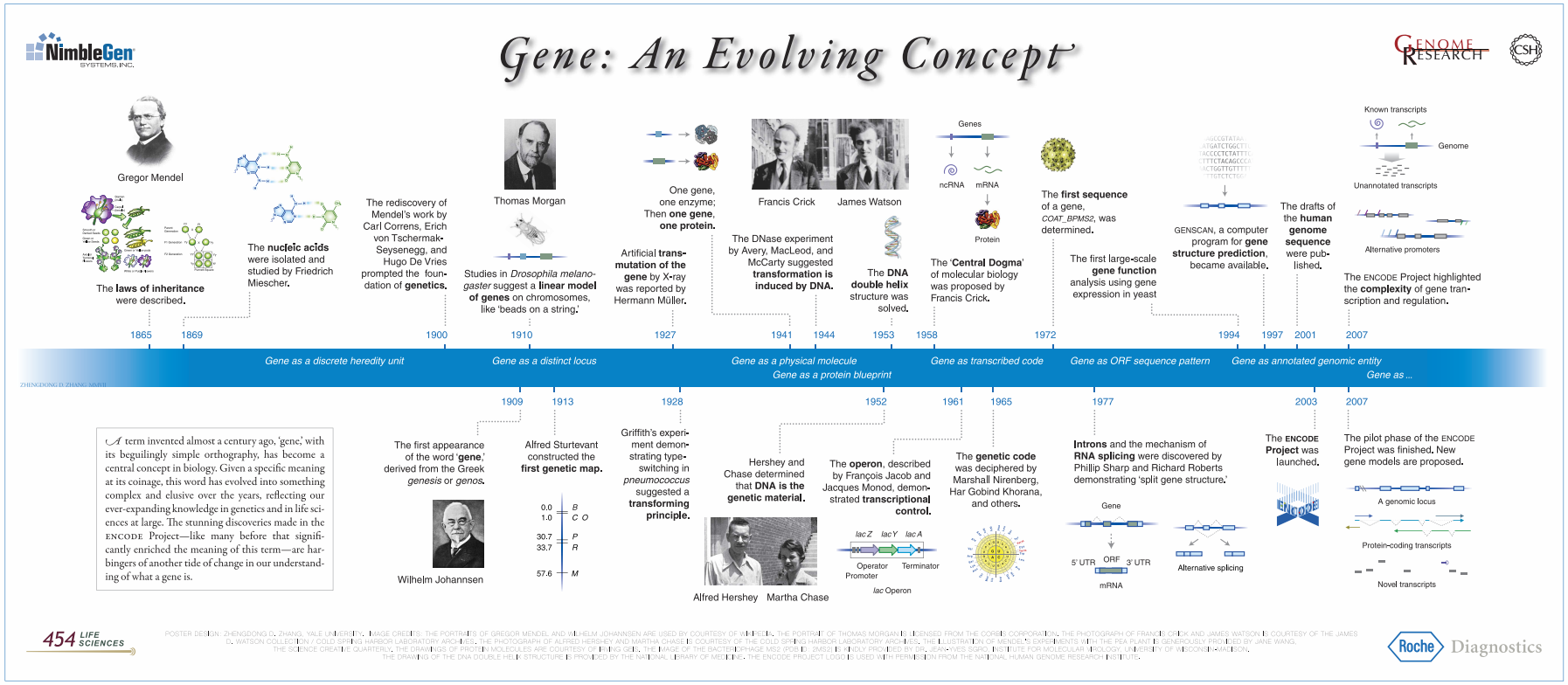


http://eagereyes.org/

Can we use parallel coordinates to show how
- analysis parameters
- execution settings

influence analysis results?

Workflow evolution visualization



- Can we use timelines to show how a workflow has been changed?
- What components were introduced/removed and when/why?

umcg

molgenis

Questions?

http://www.molgenis.org

http://www.molgenis.org/wiki/ComputeStart

h.v.byelas@gmail.com m.a.swertz@gmail.com

H. Byelas and M. Swertz, “Visualization of bioinformatics