Molgenis Compute based pipelines

Freerk van Dijk
Genomics Coordination Center
University Medical Center Groningen
freerk.van.dijk@gmail.com

umcg

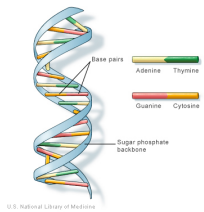
molgenis

Content

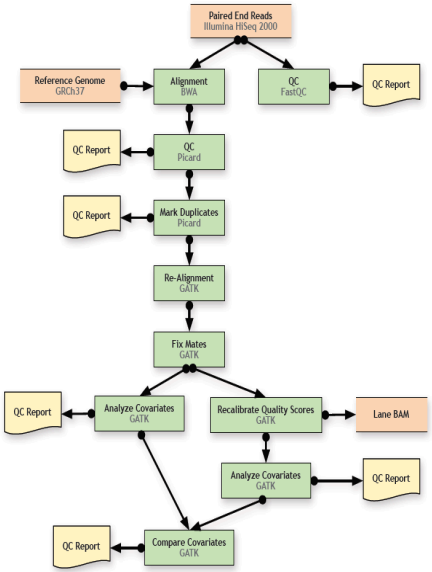
- General background
- Molgenis Compute commandline
- Molgenis ComputeDB (GUI)
- Pipelines
- Future work

General background

General background / example: NGS alignment workflow



HiSeq



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

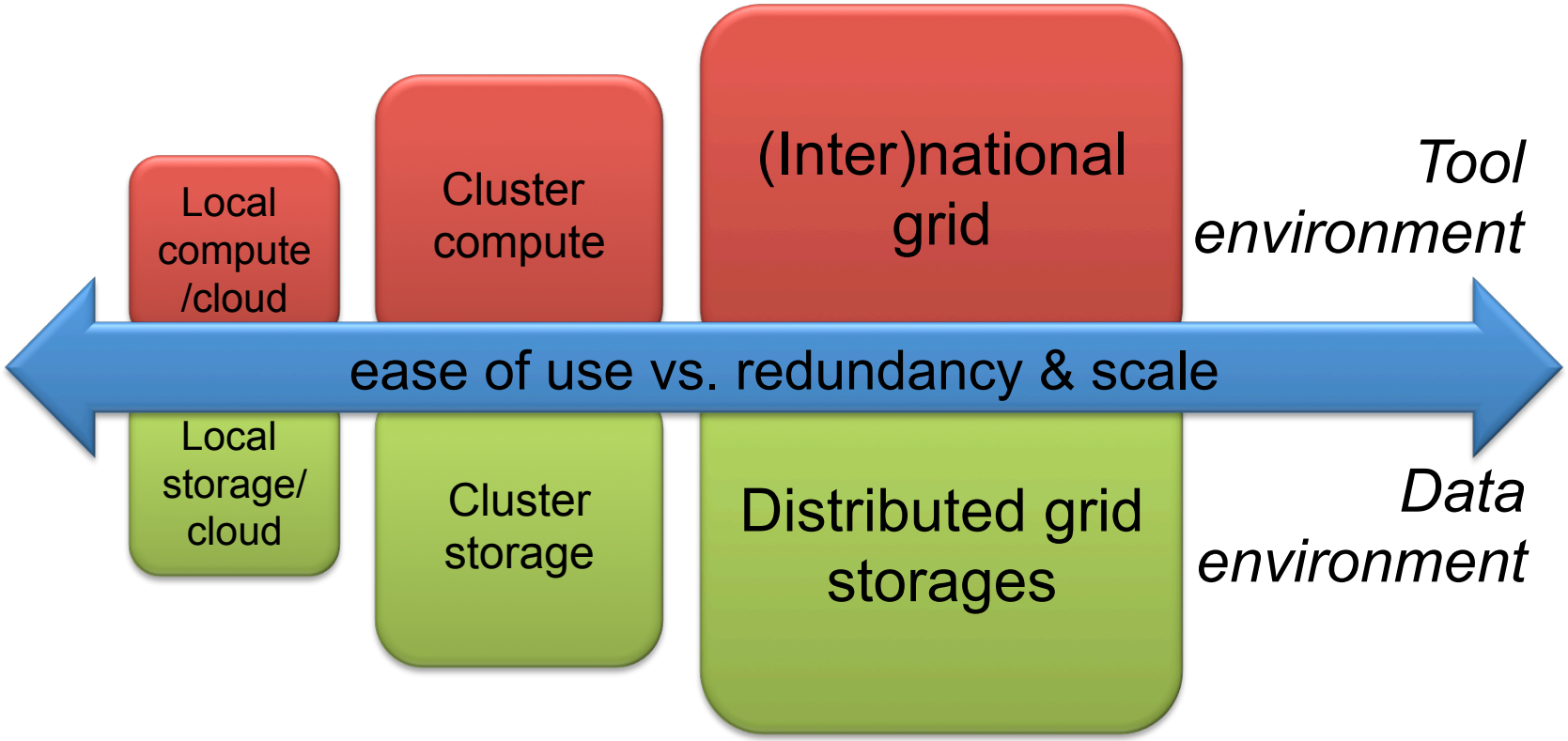
Raw data → ***alignment workflow*** → Result data

10-100 samples

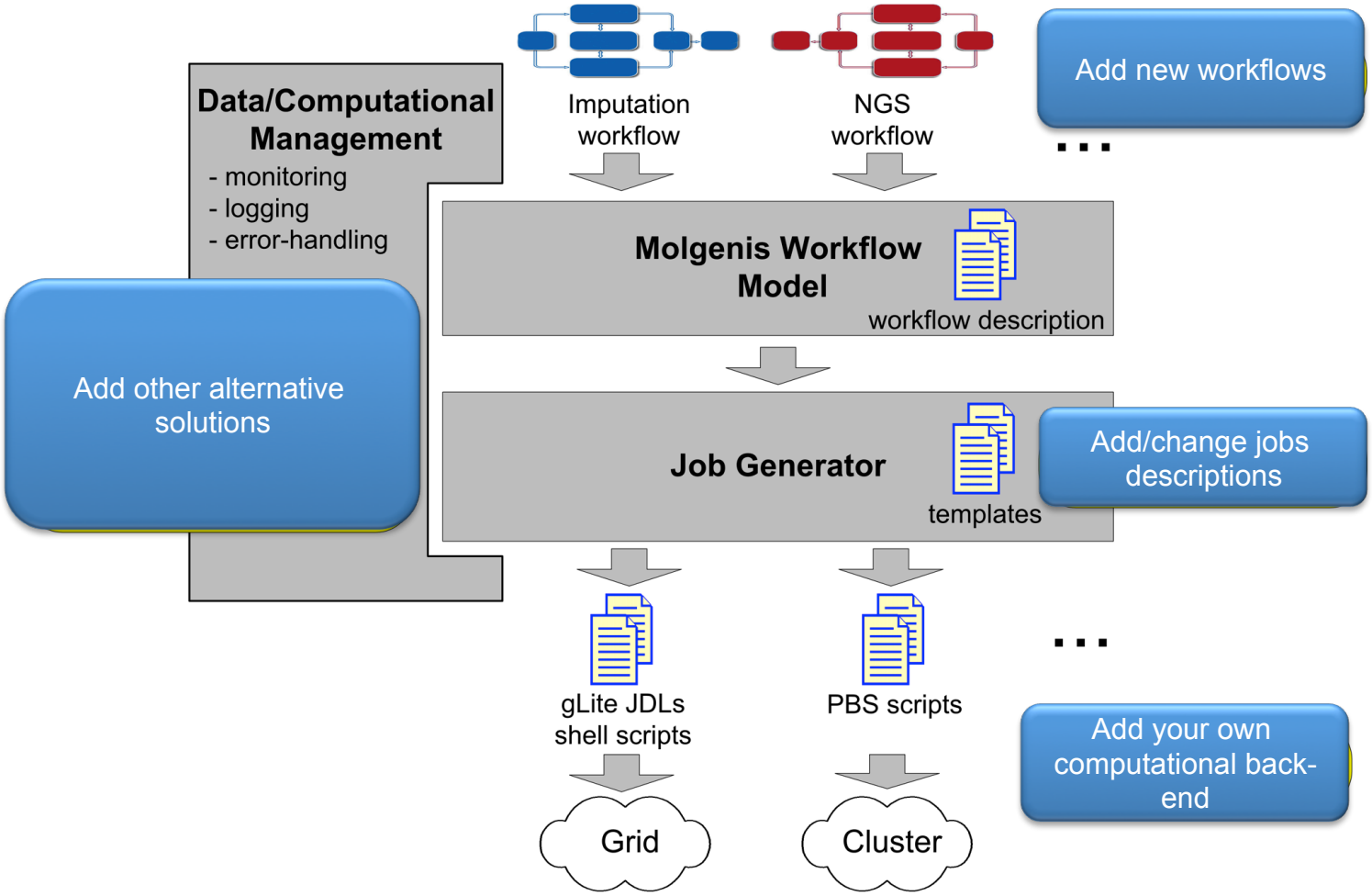
20 – 200 days

80 – 800 GB

Computational environments



Command-line generator



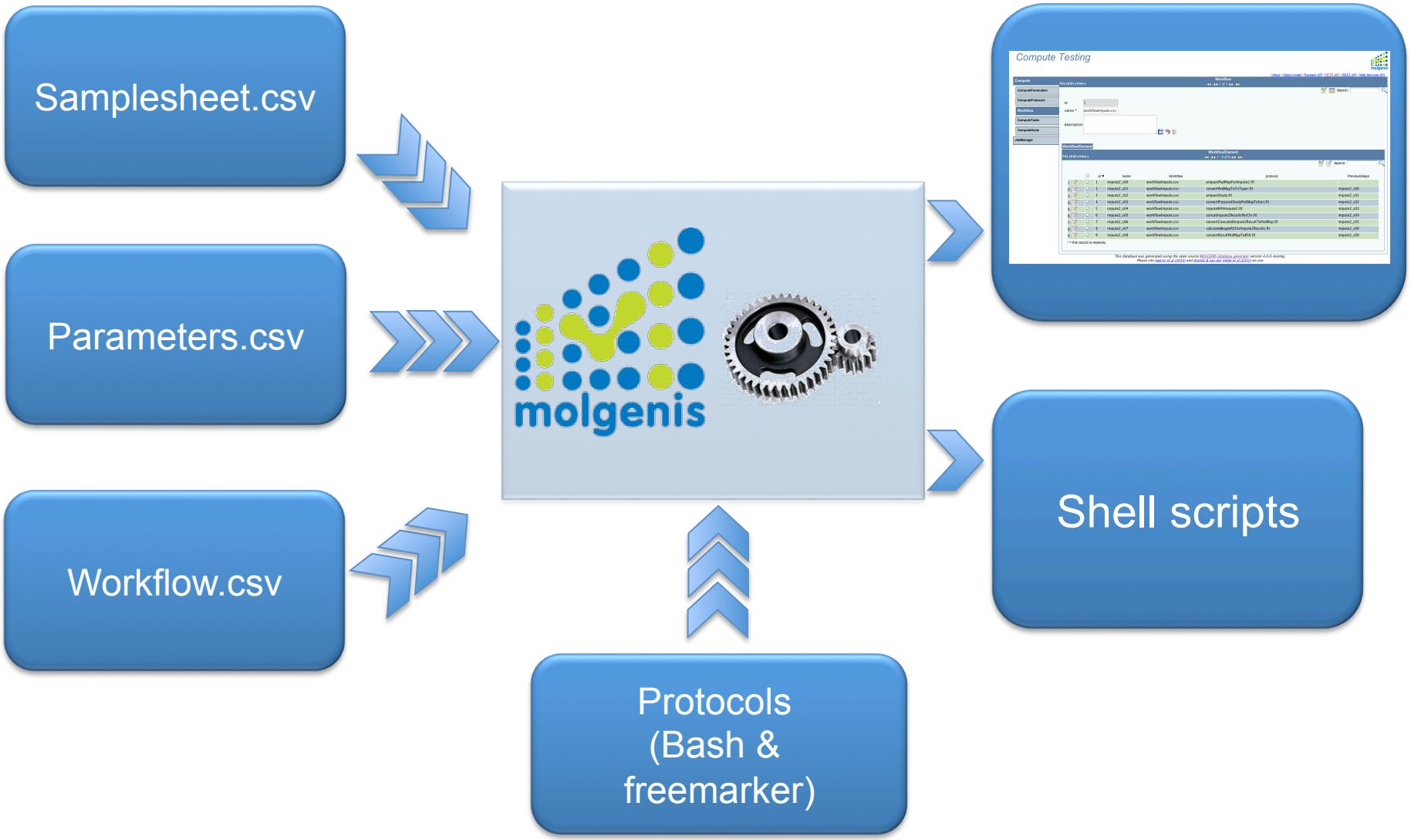
- Generates jobs (scripts) from model described in files
- Suitable for workflows (PBS cluster) and single jobs (gLite grid)

Molgenis Compute commandline

umcg

molgenis

Compute overview



Samplesheet.csv

ID	sampleID	Project	Machine	SequencingDate	Run	Flowcell	Lane	Barcode	SequencingT	Manufacture	Enrichmentkit
3	Sample3_2	Testproject2	SN163	130809	501	BD2BWCACXX	4	RPI 01 ATCACG	PE	Agilent	SureSelect_All_Exon_50MB_v5.1
4	Sample5_2	Testproject2	SN163	130809	501	BD2BWCACXX	4	RPI 04 TGACCA	PE	Agilent	SureSelect_All_Exon_50MB_v5.1
5	Sample5_2	Testproject2	SN163	130809	501	BD2BWCACXX	5	RPI 04 TGACCA	PE	Agilent	SureSelect_All_Exon_50MB_v5.1
6	Sample4_2	Testproject2	SN163	130809	501	BD2BWCACXX	4	RPI 08 ACTTGA	PE	Agilent	SureSelect_All_Exon_50MB_v5.1

Parameters.csv

Samplesheet.csv

Parameters.csv

queue	devel				
mem	4				
walltime	23:59:00				
nodes	1				
ppn	1				
defaultInterprete	#!/bin/bash				
stage	module load				
checkStage	module list				
WORKDIR	/gcc/				
root	${WORKDIR}				
tempDir	${WORKDIR}/groups/gaf/tmp02/tmp/				
resDir	${root}/resources/				
toolDir	${root}/tools/				
scriptDir	${toolDir}/scripts/				
gafHome	${root}/groups/gaf/				
tmpDataDir	${root}/groups/gaf/tmp02/				
allRawtmpDataDi	${tmpDataDir}/rawdata/				
allRawNgstmpDat	${allRawtmpDataDir}/ngs/				
allRunDemultiple	${tmpDataDir}/projects/${project}/${run}				
runPrefix	${sequencingStartDate}_${sequencer}_${run}_${flowcell}				

umcg

molgenis

Workflow.csv

step	protocol	dependencies
FastQC	protocols/FastQC.sh	
BwaAlign	protocols/BwaAlign.sh	
SamToBam	protocols/SamToBam.sh	BwaAlign
SortBam	protocols/SortBam.sh	SamToBam;inputSortBam=alignedBam;tmpSortedBam=tmpAlignedSortedBam;tmpSortedBamIdx=tmpAlignedSortedBamIdx;sc
MergeBam	protocols/MergeBam.sh	SortBam;inputMergeBam=alignedSortedBam;inputMergeBamIdx=alignedSortedBamIdx;tmpMergedBam=tmpSampleMergedB
MarkDuplica	protocols/MarkDuplicates.sh	MergeBam

Parameters.csv

Workflow.csv

Example snippet SortBam.sh:

```
#string inputSortBam
#string tmpSortedBam
#string tmpSortedBamIdx


#Run picard, sort BAM file and create index on the fly
java -jar -Xmx3g $PICARD_HOME/${sortSamJar} \
INPUT=${inputSortBam} \
OUTPUT=${tmpSortedBam} \
SORT_ORDER=coordinate \
CREATE_INDEX=true \
VALIDATION_STRINGENCY=LENIENT \
MAX_RECORDS_IN_RAM=2000000 \
TMP_DIR=${tempDir}
```

umcg

molgenis

Protocol structure in more detail

```
//header
      #MOLGENIS walltime=15:00 nodes=1 cores=4 mem=6gb
      #list inputMergeBam
      #string outputMergedBam
//tool management
      module load bwa/${bwaVersion}
//data management
      getFile ${indexfile}
      getFile ${leftbarcodefqgz}
//template of the actual analysis
      bwa aln \
      ${indexfile} ${leftbarcodefqgz} \
      -t ${bwaaligncores} -f ${leftbwaout}
//data management
      putFile ${leftbwaout}
```

umcg

molgenis

Data transfer

- getFile and p
  - are back-e
  - now, we
    - check i
    - do srm/
- Input

```
getFile $
```

- Generated o

```
getFile $V
imputatio
chr20.map
```

```
#------- data transfer

getRemoteLocation()
{
    ARGS=($@)
    myFile=${ARGS[0]}
    remoteFile=srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP2${myFile:`expr length $TMPDIR`}
    echo $remoteFile
}

getFile()
{
    ARGS=($@)
    NUMBER="${#ARGS[@]}";
    if [ "$NUMBER" -eq "1" ]
    then

    myFile=${ARGS[0]}
    remoteFile=`getRemoteLocation $myFile`

    # 1. myPath = getPath( myFile ) will strip off the file name and return the path
    mkdir -p $(dirname "$myFile")

    # 2. cp srm:.../remoteFile myFile
    echo "srmcp -server_mode=passive $remoteFile file:///$myFile"
    srmcp -server_mode=passive $remoteFile file:///$myFile
    chmod 755 $myFile

    else
    echo "Example usage: getData \"\$TMPDIR/datadir/myfile.txt\""
    fi
}

putFile()
{
    ARGS=($@)
    NUMBER="${#ARGS[@]}";
    if [ "$NUMBER" -eq "1" ]
    then
    myFile=${ARGS[0]}
    remoteFile=`getRemoteLocation $myFile`
    echo "srmrm $remoteFile"
    srmrm $remoteFile
    echo "srmcp -server_mode=passive file:///$myFile $remoteFile"
    srmcp -server_mode=passive file:///$myFile $remoteFile
        returnCode=$?

        echo "srmcopy: ${returnCode}"

        if [ $returnCode -ne 0 ]
        then
            exit 1
        fi
    else
    echo "Example usage: putData \"\$TMPDIR/datadir/myfile.txt\""
    fi
}
```

Generated back-end independent script

```
//header
#MOLGENIS walltime=15:00 nodes=1 cores=4 mem=6
//tool management
module load bwa/0.5.8c_patched
//data management
getFile $WORKDIR/resources/hg19/indices/human_g1k_v37.fa
getFile $WORKDIR/groups/gcc/projects/cardio/run01/rawdata/
121128_SN163_0484_AC1D3HACXX_L8_CAACCT_1.fq.gz
//template of the actual analysis
bwa aln \
human_g1k_v37.fa 121128_SN163_0484_AC1D3HACXX_L8_CAACCT_1.fq.gz -t 4 \
-f 121128_SN163_0484_AC1D3HACXX_L8_CAACCT_1.bwa_align.human_g1k_v37.sai
//data management
putFile $WORKDIR/groups/gcc/projects/cardio/run01/results/
121128_SN163_0484_AC1D3HACXX_L8_CAACCT_1.bwa_align.human_g1k_v37.sai
```

Best practices

- Always output to temporary files and check return codes of software

```
#Get return code from last program call
returnCode=$?

echo -e "\nreturnCode MergeBam: $returnCode\n\n"

if [ $returnCode -eq 0 ]
then
    echo -e "\nMergedBam finished succesfull. Moving temp files to final.\n\n"
    mv ${tmpMergedBam} ${finalMergedBam}
else
    echo -e "\nFailed to move MergeBam results to ${finalMergedBam}\n\n"
    exit -1
fi
```

- Create md5sums

Molgenis ComputeDB (GUI)

umcg

molgenis

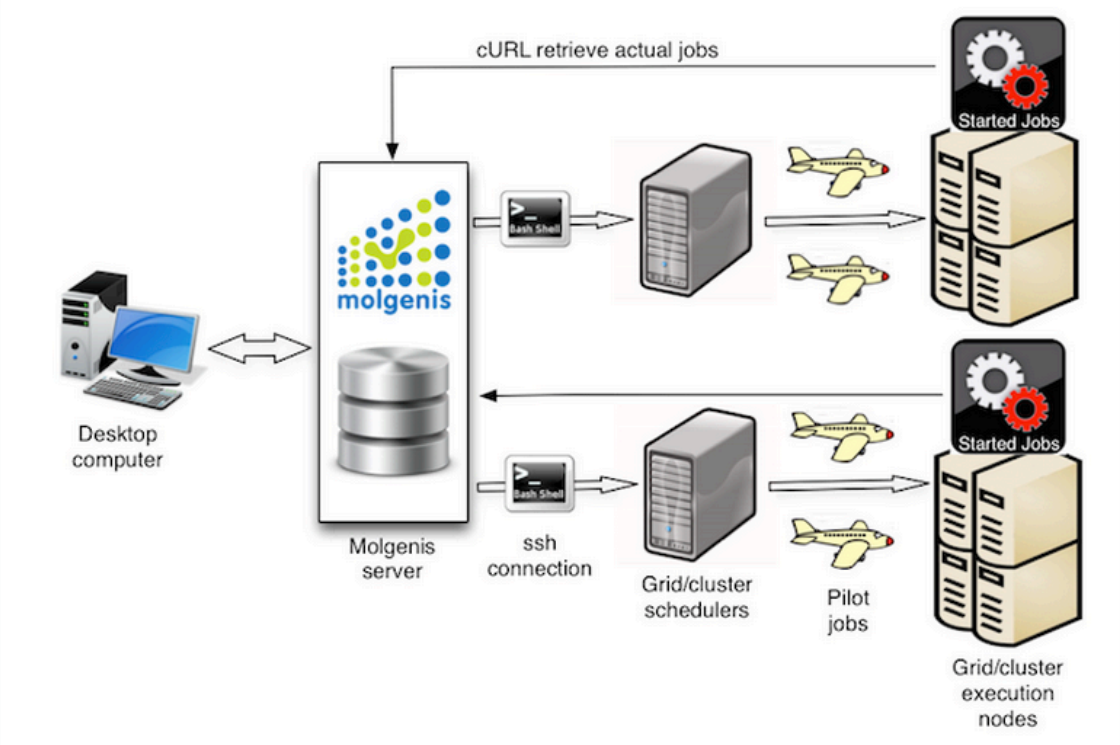
Pilot jobs

molgenis

Home

Sign in

Welcome to Molgenis Compute5db Platinum Grid Edition



This database was generated using the open source MOLGENIS database generator version 4.0.0-testing. Please cite Swertz et al (2010) or Swertz & Jansen (2007) on use.

umcg

molgenis

Login

test18@ui.grid.sara.nl (2013-09-09 12:56:45)

Active

Inactivate

Username

Password

Submit Pilots

Cancel Run

Jobs generated	0	Jobs done	589
Jobs ready	0	Jobs failed	0
Jobs running	22	Jobs cancelled	0
Pilots submitted	1261	Pilots started	902

umcg

molgenis

Dashboard

molgenis

Pilot dashboard | Runs | Tasks | Parameters | Backends | Sign out

Main / Compute

test11@ui.grid.sara.nl (2013-12-20 14:15:56)

Not active

Activate

Cancel Run

Jobs generated	2	Jobs done	0
Jobs ready	0	Jobs failed	0
Jobs running	0	Jobs cancelled	0
Pilots submitted	0	Pilots started	0

This database was generated using the open source MOLGENIS database generator version 4.0.0-testing.
Please cite Swertz et al (2010) or Swertz & Jansen (2007) on use.

Pipelines

Supported analyses

- NGS variant calling pipeline
- RNA-seq analysis
- Liftover genome builds
- … and many more pipelines

- Out-of-the-box imputation pipeline
  - Automatic installation of tools and reference datasets
  - Liftover, phasing and imputation in 3 commands

Future work

Future work

- Finish implementation new NGS pipeline
- Collect and show more statistics to user in dashboard/commandline
- Add more workflows
- Unique *job/workflow* interface to run in different computational back-ends
  - Cluster (PBS, SGE, etc.)
  - Grid (gLite)
  - Cloud (OpenStack) <- Ongoing
- Interface with Galaxy (and other WMS?)
- Visual analytics of data/workflows/runs

Future work

- Explore easybuild
  http://hpcugent.github.io/easybuild/

Contributors: GCC team

- Compute development
  - George Byelas
  - Martijn Dijkstra
- Pipeline development and NGS analysis:
  - Pieter Neerincx
  - Gerben van der Vries
  - Lennart Johansson
  - Roan Kanninga

Github repositories

- Molgenis Compute (& tutorial)
  - https://github.com/molgenis/molgenis-pipelines/
- Imputation out-of-a-box
  - https://github.com/molgenis/molgenis-imputation
- Pipelines
  - https://github.com/molgenis/molgenis-pipelines/tree/master/compute5
- NGS (Under development)
  - https://github.com/molgenis/molgenis-pipelines/tree/master/compute5/NGS_alignment_SNP_calling