FMI
Friedrich Miescher Institute
for Biomedical Research

Using Galaxy to provide a NGS Analysis Platform

GTC's NGS & Bioinformatics Summit Europe
October 7-8, 2013 in Berlin, Germany.

(public version)

Hans-Rudolf Hotz (hrh@fmi.ch)

Friedrich Miescher Institute for Biomedical Research
Basel, Switzerland

Friedrich Miescher Institute

- funded by the Novartis Research Foundation
- affiliated institute of Basel University

311 employees

(incl. 90 PhD students, 94 Post Docs)

Epigenetics
(7 research groups)

Cancer
(7 research groups)

Neurobiology
(8 research groups)

Technology Platforms

Computational Biology – Cell Sorting – Imaging and Microscopy – *C. elegans*
Functional Genomics – Histology – Mass Spectrometry – Protein Structure

- member of Swiss Institute of Bioinformatics

SIB Swiss Institute of Bioinformatics

UNI BASEL

NOVARTIS

FMI Friedrich Miescher Institute for Biomedical Research

Analyzing NGS data in a Bioinformatics Core Facility is *fascinating* because

- scientists keep coming up with new kind experiments

- new algorithms to deal with NGS data are developed continuously

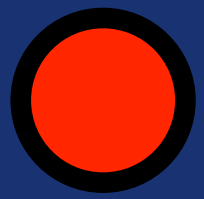
- there is a new (improved) sequencing instrument on the market every few months

FMI
Friedrich Miescher Institute
for Biomedical Research

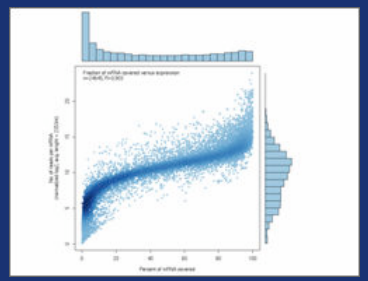
Analyzing NGS data in a Bioinformatics Core Facility is *difficult* because

people with different background/training are interested in using NGS

the "average" lab scientist is looking for the red button to press

bizarre output from the sequencer →  → **publication in *Nature***

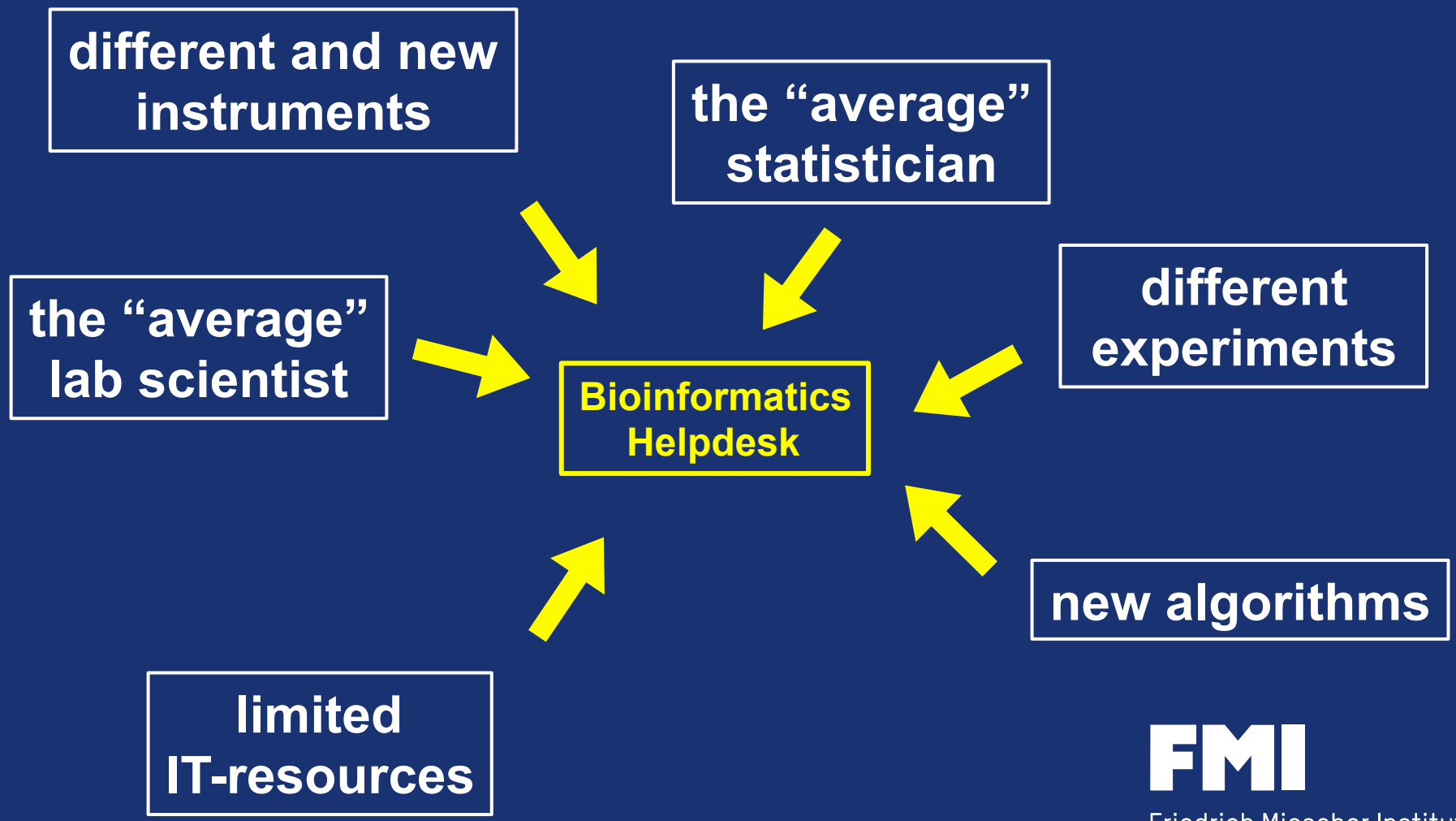
the "average" statistician is creating wonderful blots.....



→ **...nobody understands**

FMI
Friedrich Miescher Institute for Biomedical Research

and the Bioinformatics Helpdesk is caught in the middle....



FMI
Friedrich Miescher Institute for Biomedical Research

.....looking for a solution

limited IT-resources	can be solved (with money)
new instruments new algorithms different experiments	follow the “literature” and test the new open source tools yourself → learn R/Bioconductor → flexible environment
the “average” lab scientist the “average” statistician	turn a command line tool (R script) into a ‘red button’

the ‘red button’ should be as similar as possible to the command line process

FMI
Friedrich Miescher Institute
for Biomedical Research

the solution:

Galaxy / FMI-Xenon1

Analyze Data | Workflow | Shared Data | Lab | Visualization | Admin | Help | User

Tools — Options

- Get Data
- Text Manipulation
- FASTA manipulation
- FASTQ manipulation
- Filter and Sort
- Join, Subtract and Group
- Unix Tools
- Convert Formats
- Extract Features
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Multiple regression
- Multiple Alignments

EMBOSS TOOLS

- EMBOSS search tools
- EMBOSS sequence manipulation tools

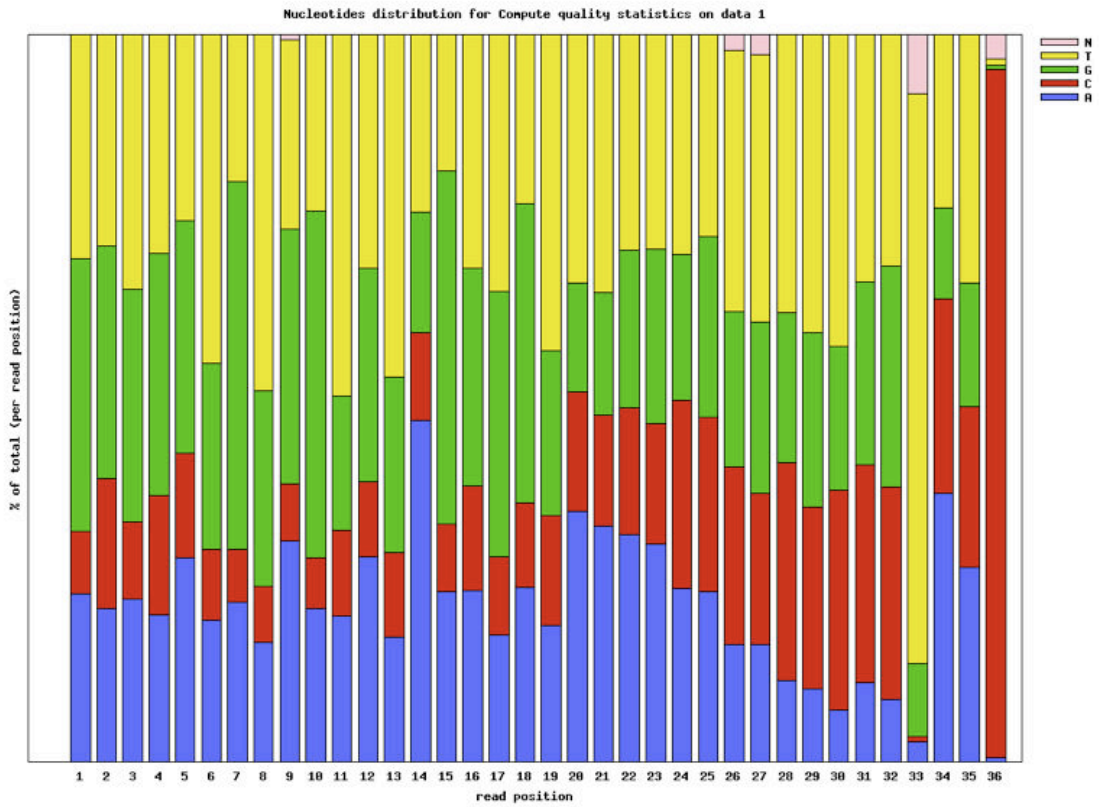
GALAXY NGS TOOLS

- NGS: Peak Calling

FMI TOOLS

- FMI: Bioinformatics-Support
- FMI: Functional Genomics
- FMI: DNA Restriction Digest
- FMI: DeepSeqRepository
- FMI: qPCR
- FMI: Thoma-lab
- FMI: Schuebeler-lab
- FMI: Roska-lab

Workflows



History — Options

poster 1 — 380.9 Mb

- 3: Draw nucleotides distribution chart on data 2
- 2: Compute quality statistics on data 1
- 1: example_1.fq

http://galaxyproject.org

a flexible environment which allows you to turn command line tools into 'red buttons'

FMI
Friedrich Miescher Institute for Biomedical Research

http://galaxyproject.org

Galaxy

"Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses."

The Galaxy Team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University.

The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

FMI

Friedrich Miescher Institute for Biomedical Research

http://galaxyproject.org

Galaxy

....and I am NOT part of the Galaxy Team!

I am just a member of the worldwide community of many Galaxy users, adopters, developers, evangelists, etc.

FMI

Friedrich Miescher Institute
for Biomedical Research

what does Galaxy?

Galaxy

- provides a GUI (i.e. the ‘red button’) to (Bioinformatics) command line tools in your web browser
- manages/stores your (raw) data and results
- allows you to create workflows
- allows sharing and reproducing your analysis

FMI
Friedrich Miescher Institute
for Biomedical Research

Use Galaxy

Galaxy

http://wiki.galaxyproject.org/BigPicture/Choices

public and free web service: *http://usegalaxy.org/*

deploy your own Galaxy server:

- local server
- cloud (Amazon Machine Images)
- Galaxy appliance (offered by BioTeam)

FMI
Friedrich Miescher Institute
for Biomedical Research

why are we using Galaxy

Galaxy

- open source software
- it provides a standard set of tools
- we can add our own scripts and tools
  → *turn open source tools into a 'red button'*
- the Galaxy community is huge and the software is established (first publication in 2005)
- a local installation is simple to set up
- it is flexible (you can adjust it to your needs)

in use at the FMI since 2007

FMI
Friedrich Miescher Institute
for Biomedical Research

it is really simple to install

Galaxy

requirements:

- a Mac or linux PC with Python and Mercurial

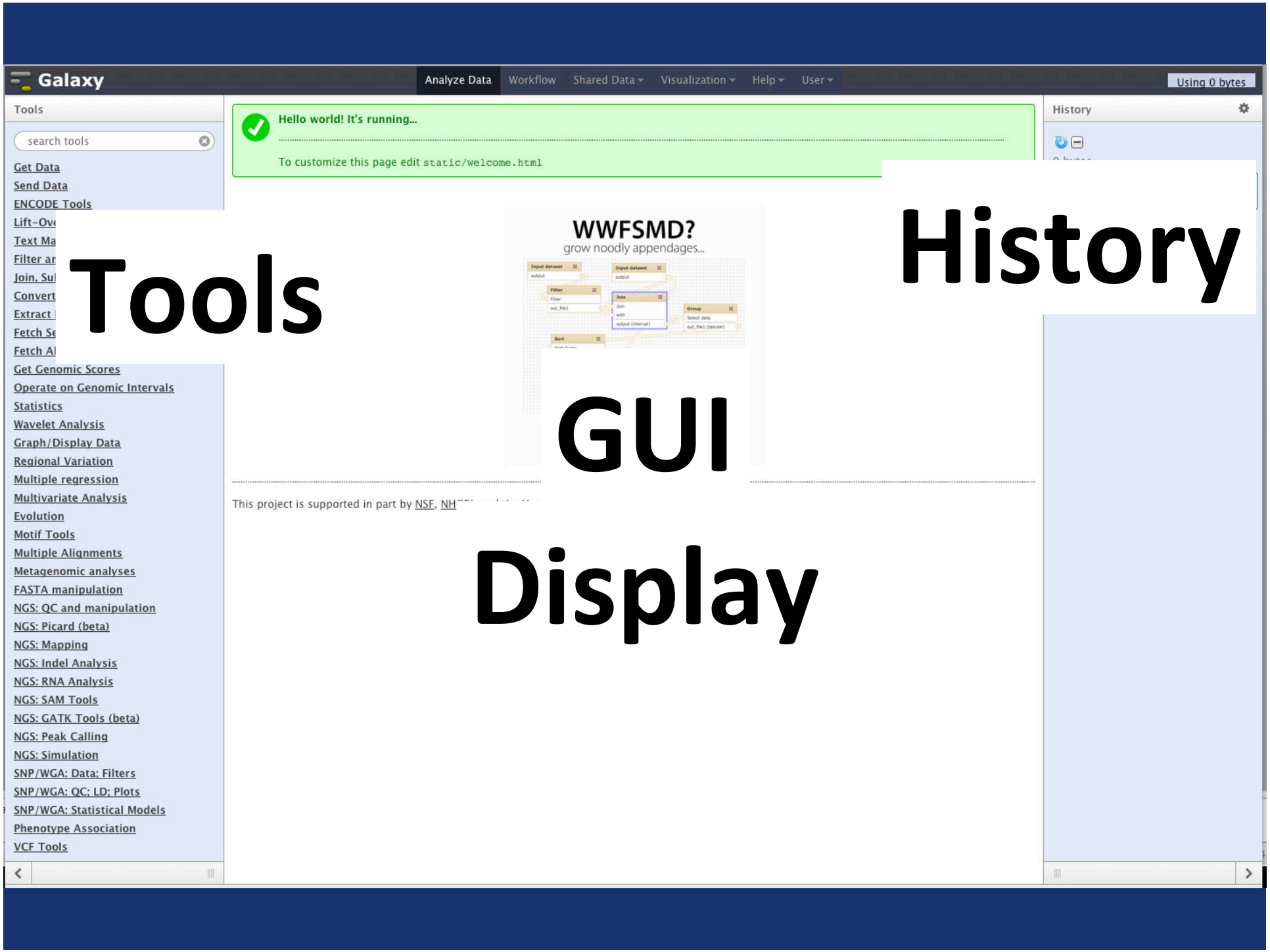
just 3 commands:

- hg clone https://bitbucket.org/galaxy/galaxy-dist/
- cd galaxy-dist
- sh run.sh

...and it is ready in your web browser at:

http://localhost:8080

FMI
Friedrich Miescher Institute
for Biomedical Research


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 bytes
Tools
search tools
Get Data
Send Data
ENCODE Tools
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: Picard (beta)
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: Simulation
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Phenotype Association
VCF Tools
Hello world! It's running...
To customize this page edit static/welcome.html
WWFSMD?
grow noodly appendages...
This project is supported in part by NSF, NH
History
Tools
GUI
Display

how does it work

Galaxy

Tools

search tools

Get Data

- Upload File fr
- UCSC Main ta

Upload File (version 1.1.3)

File Format:

Auto-detect

Which format? See help below

File:

Browse...

TIP: Due to browser limitations, uploading files larger than 2GB is guara
method (below) or FTP (if enabled by the site administrator).

URL/Text:

foo 1 2

Here you may specify a list of URLs (one per line) or paste the contents

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

Click to Search or Select

Execute

History

Options

8 bytes

1: Pasted Entry

1 region

format: interval, database: ?

Info: uploaded interval file

view in GeneTrack

1.Chrom	2.Start	3.End
foo	1	2

FMI

Friedrich Miescher Institute
for Biomedical Research

how does it work

Galaxy

History Options

Analyze Data | Workflow | Shared Data | Help | User | Using 113 bytes

```
##gff-version 2
##bed_to_gff_converter.py
foo     bed2gff region_0     2     2     0     +     .     region_0;
```

History Options

113 bytes

2: BED-to-GFF on data 1

1: Pasted Entry

view in GeneTrack

1.Chrom	2.Start	3.End
foo	1	2

alignment

- AXT to FASTA Converts formatted file to FASTA
- AXT to LAV Converts a formatted file to LAV fo
- BED-to-GFF converter
- FASTA-to-Tabular con
- GFF-to-BED converter
- LAV to BED Converts a LAV formatted file to BED format
- Maf to BED Converts a MAF formatted file to the BED format

BED-to-GFF (version 2.0.0)

Convert this query:

1: (as bed) Pasted Entry

Execute

FMI

Friedrich Miescher Institute for Biomedical Research

Galaxy out of the box

Galaxy

input tools:

- text box / upload file / url
- access to UCSC table browser and ensembl biomart

tools for file conversion and text manipulation

tools for table calculation, basic set-theory and operation on genomic intervals

FMI
Friedrich Miescher Institute
for Biomedical Research

adding more tools

Galaxy

Galaxy Tool Shed

http://wiki.galaxyproject.org/Tool%20Shed

enables sharing of Galaxy tools across the Galaxy community

more than 1000 tools available

handles 3rd party dependencies

your own tool

ideally submitted to the Tool Shed

FMI
Friedrich Miescher Institute for Biomedical Research

adding your own tools

Galaxy

everything is possible in Galaxy

As long as you can run it on the command line, you can incorporate it into Galaxy.

- add the executable or script (R, perl, python, bash, etc)
- write a tool definition file
- add it to the list of tools

FMI
Friedrich Miescher Institute for Biomedical Research

command line to ‘red button’

Galaxy

```
bash-3.2$ ls
bed_to_gff_converter.py   foo.bed
bash-3.2
```

```
bash-3.2$ cat foo.bed
foo     1       2
bash-3.2$
```

```
bash-3.2$ ./bed_to_gff_converter.py foo.bed foo.gff
1 lines converted to GFF version 2.
bash-3.2$
```

```
bash-3.2$ ls
bed_to_gff_converter.py   foo.bed   foo.gff
bash-3.2$ cat foo.gff
##gff-version 2
##bed_to_gff_converter.py
foo  bed2gff region_0    2   2   0   +   .   region_0;
bash-3.2$
```

command line to ‘red button’

Galaxy

```
<tool id="bed2gff1" name="BED-to-GFF" version="2.0.0">
 <description>converter</description>

 <command>bed_to_gff_converter.py $input $output</command>

 <inputs>
  <param format="bed" name="input" type="data" label="Convert this"/>
 </inputs>

 <outputs>
  <data format="gff" name="output" />
 </outputs>

 <help>
This tool converts data from BED format to GFF format
 </help>

</tool>
```

→ no need to define/design a GUI !

FMI
Friedrich Miescher Institute
for Biomedical Research

sort of a ‘red button’

Galaxy

History Options

Analyze Data | Workflow | Shared Data | Help | User | Using 113 bytes

```
##gff-version 2
##bed_to_gff_converter.py
foo     bed2gff region_0     2     2     0     +     .     region_0;
```

History Options

113 bytes

2: BED-to-GFF on data 1

1: Pasted Entry

view in GeneTrack

1.Chrom	2.Start	3.End
foo	1	2

alignment

- AXT to FASTA Converts formatted file to FASTA
- AXT to LAV Converts a formatted file to LAV fo
- BED-to-GFF converter
- FASTA-to-Tabular con
- GFF-to-BED converter
- LAV to BED Converts a LAV formatted file to BED format
- Maf to BED Converts a MAF formatted file to the BED format

BED-to-GFF (version 2.0.0)

Convert this query:

1: (as bed) Pasted Entry

Execute

FMI

Friedrich Miescher Institute for Biomedical Research

a few more highlights

Galaxy

History Lists
Saved Histories

History
Options

Galaxy
Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Workflow Canvas | imported: metagenomic analysis
Options

Input dataset
output

Input dataset
output

Select high quality segments
Reads
Quality scores
output1 (fasta)

Concatenate datasets
Concatenate Dataset
Dataset 1 > Select
out_file1

Megablast
Compare these sequences
output1 (tabular)

FASTA-to-Tabular
Convert these sequences
output (tabular)

Join two Datasets
Join
with

Add column
to Dataset
out_file1

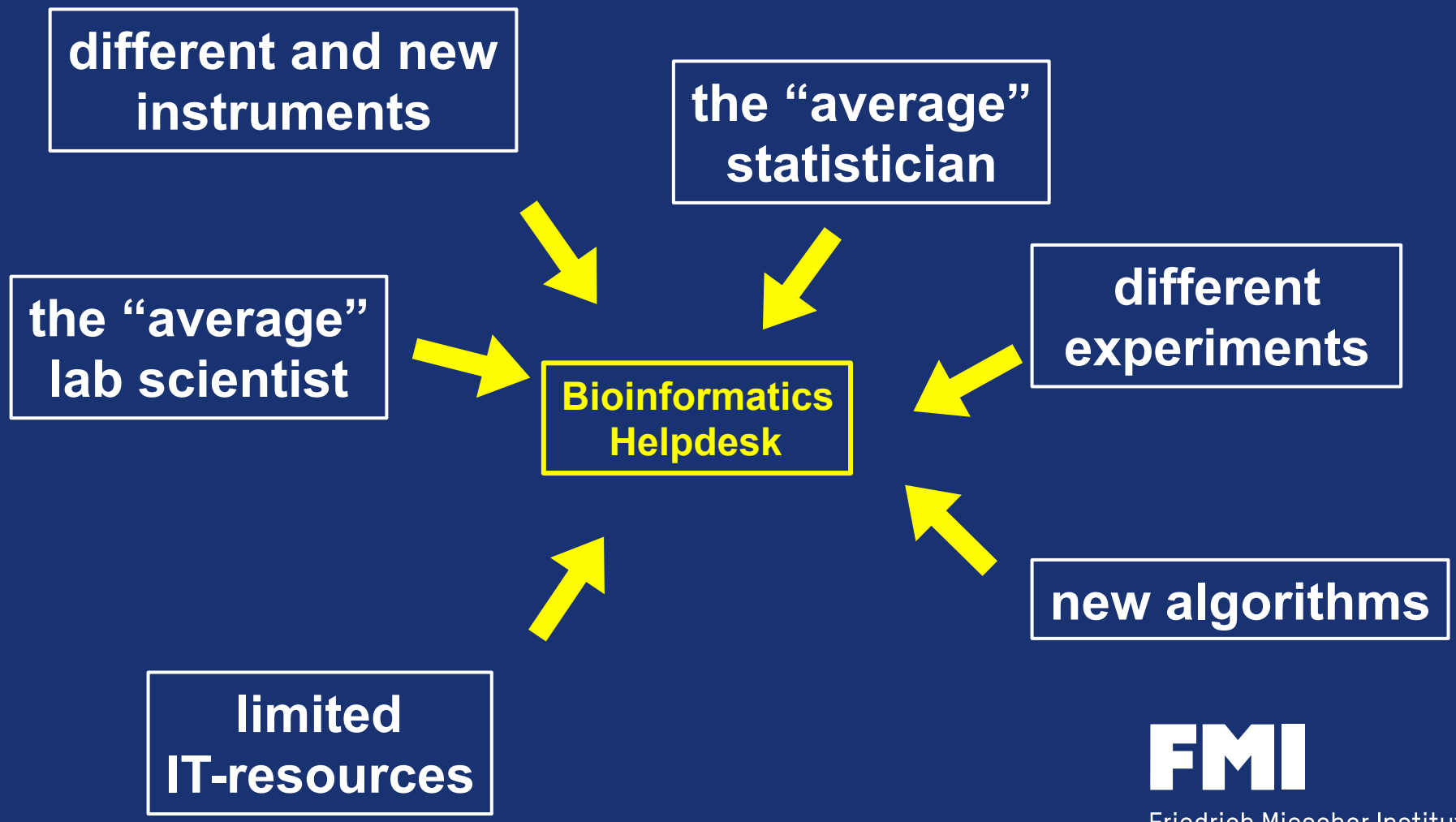
Megablast
Compare these sequences
output1 (tabular)

Tabular-to-FASTA

1: Pasted Entry

FMI
Friedrich Miescher Institute
for Biomedical Research

what happened to the poor guy sitting at the Bioinformatics Helpdesk....



FMI
Friedrich Miescher Institute for Biomedical Research

He is a Galaxy administrator

Galaxy

the "average" lab scientist ⟷ the "average" statistician

...and he adds the R script to Galaxy

→ the lab scientist can repeat the analysis

in the 'friendly' web browser and
not on the 'scary' command line

FMI
Friedrich Miescher Institute
for Biomedical Research

He is a Galaxy administrator

Galaxy

different and new instruments

new algorithms

different experiments

...and he adds the new tools to Galaxy

→ and everybody can test them

without any delay

FMI

Friedrich Miescher Institute for Biomedical Research

He is a Galaxy administrator

Galaxy

limited
IT-resources

- no need to buy extra hardware

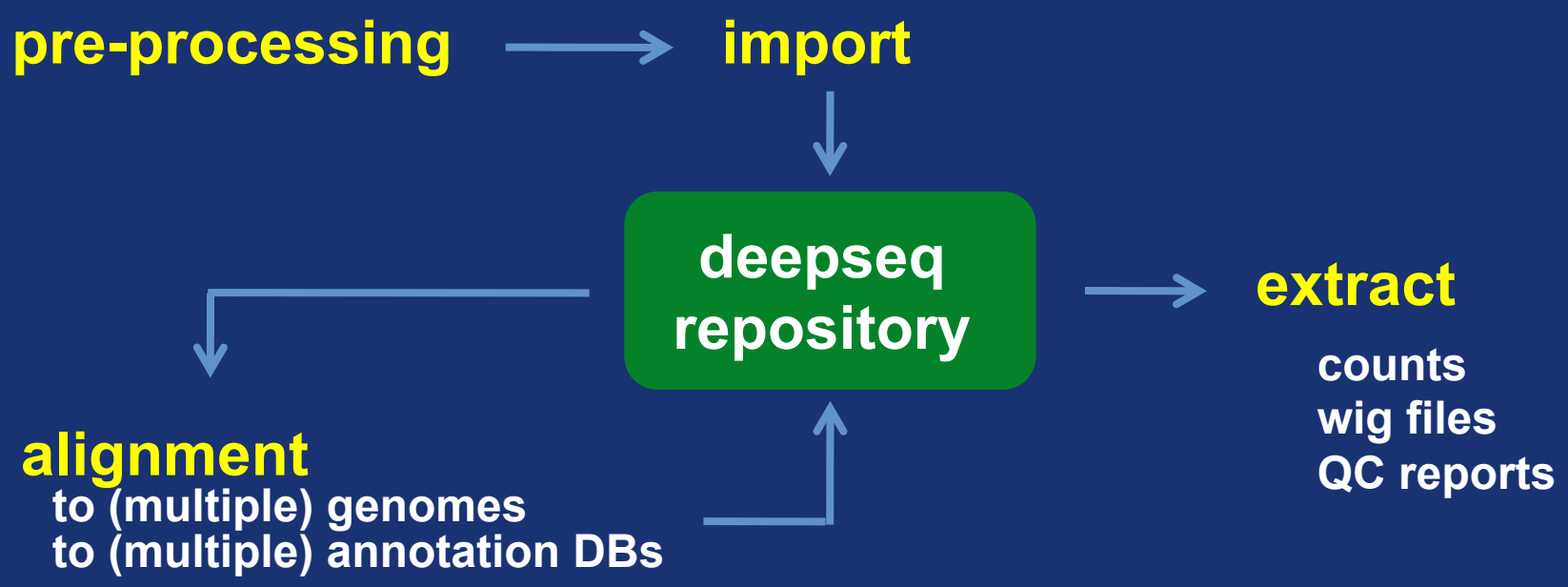
- Galaxy provides tools to track and report jobs (errors are flagged)

- Galaxy provides tools to limit disk space

- Galaxy allows you to share data

FMI
Friedrich Miescher Institute
for Biomedical Research

NGS analysis at the FMI (with the old pipeline)



FMI
Friedrich Miescher Institute for Biomedical Research

The old NGS pipeline has been

....just a bunch of Perl scripts

....just a simple file system

The new NGS pipeline is

a Bioconductor package: **"QuasR"**
(Quantification and Analysis of Short Reads)

http://www.bioconductor.org/packages/release/bioc/html/QuasR.html

FMI
Friedrich Miescher Institute
for Biomedical Research

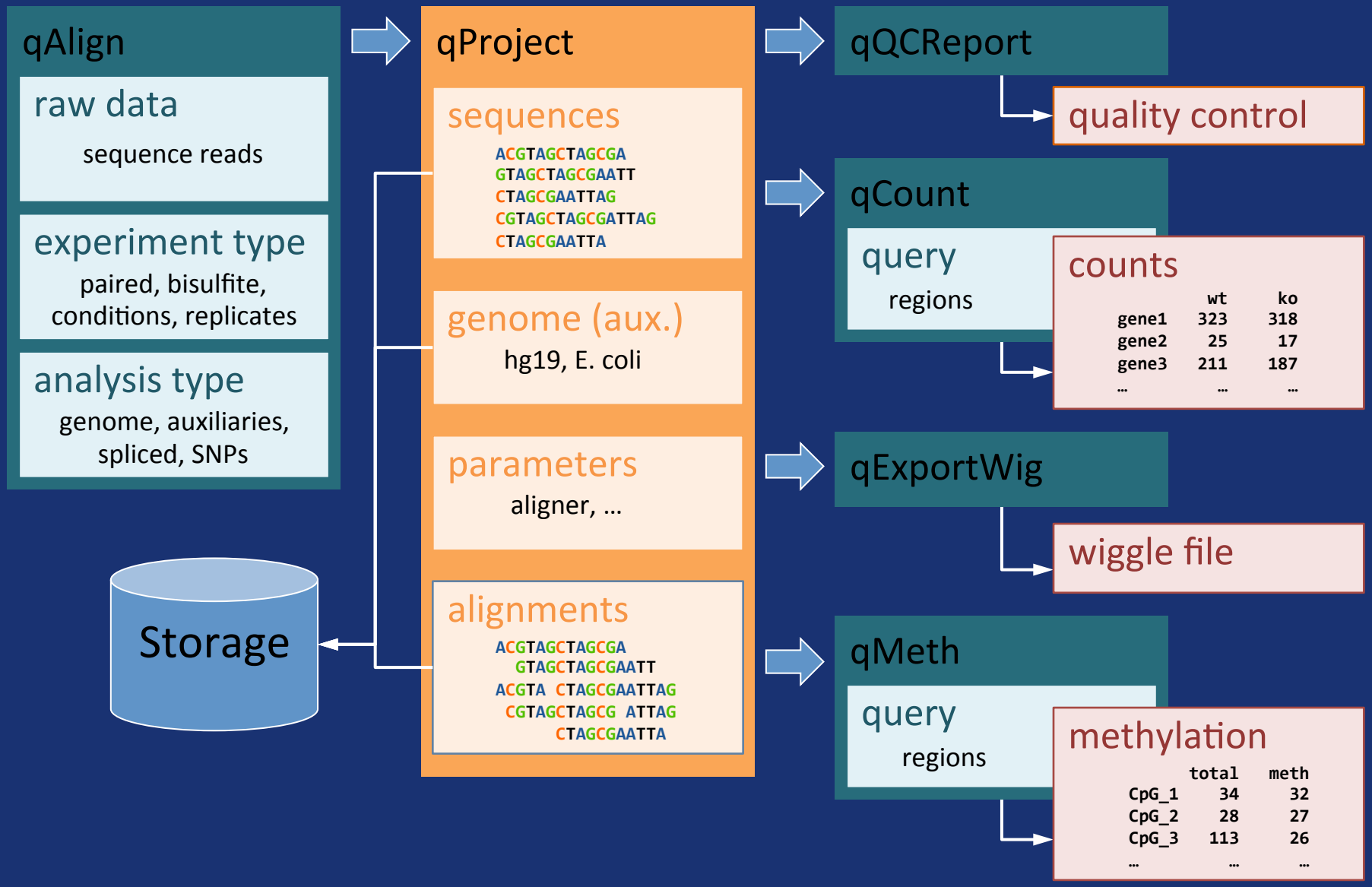
the new NGS pipeline is

new Bioconductor package: **QuasR**
(Quantification and Analysis of Short Reads)

- package that provides an end-to-end analysis solution for tag counting applications
- ships with the aligners Bowtie and SpliceMap
- creates alignments from within R
- provides a set of simple to use functions to create a large variety of count-tables
- provides an additional layer of abstraction on top of pre-existing tools in Bioconductor

FMI
Friedrich Miescher Institute
for Biomedical Research

QuasR Bioconductor Package



it is really simple with QuasR

```
sampleFile <- "data/samples_chip.txt"
genomeFile <- "genomes/hg19.fa"

proj <- qAlign(sampleFile, genome=genomeFile)

qExportWig(proj, binsize=10)
```

but still to scary / complicated

how can Galaxy help?

FMI
Friedrich Miescher Institute
for Biomedical Research

a general NGS workflow

Galaxy

align reads

→ BAM file

extract data

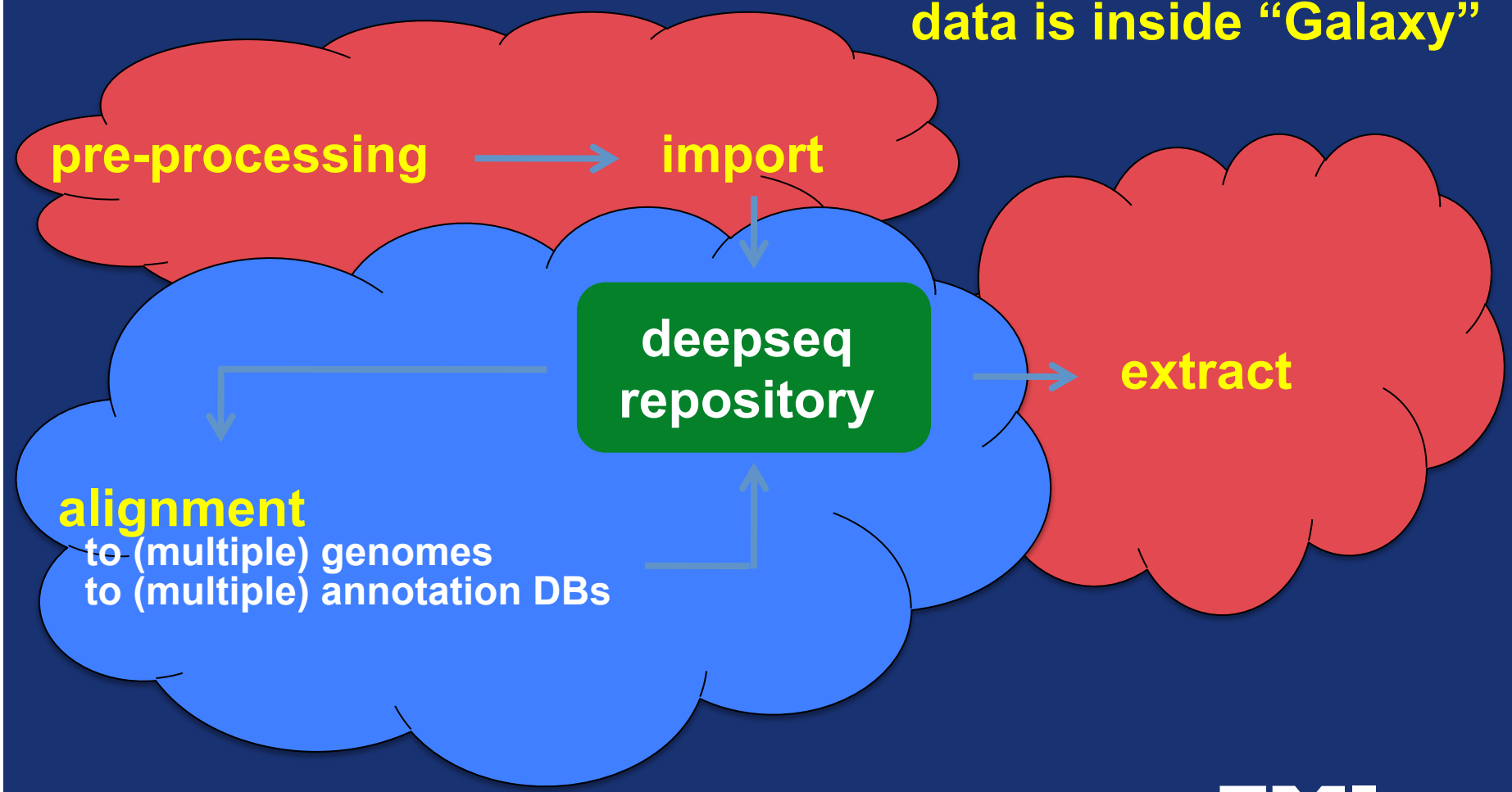
→ WIG file

data is hidden in Galaxy

FMI
Friedrich Miescher Institute
for Biomedical Research

NGS analysis at the FMI

Galaxy

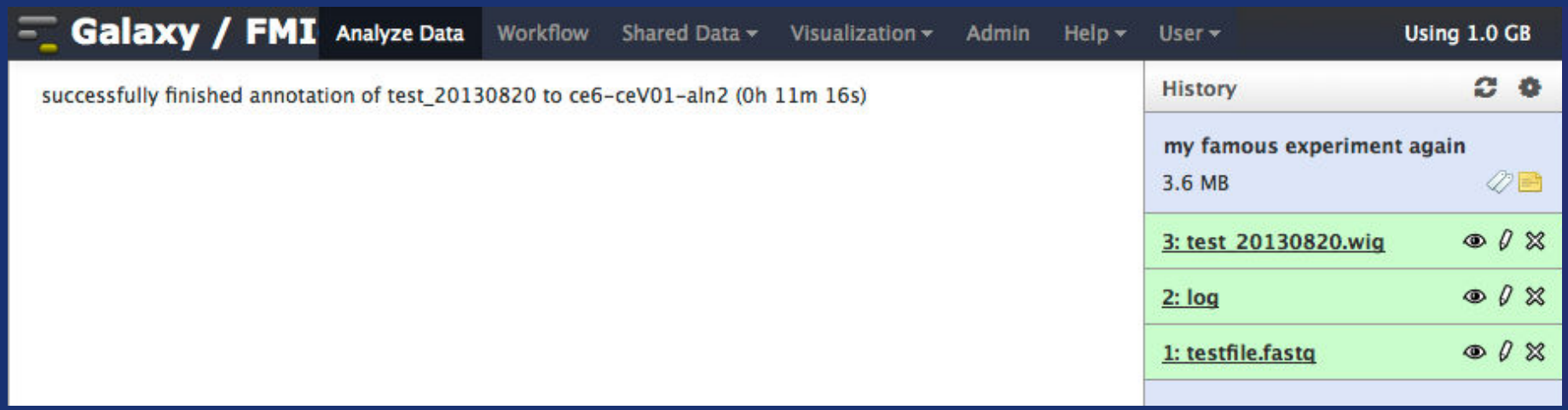


FMI
Friedrich Miescher Institute for Biomedical Research

storing data outside of Galaxy

Galaxy

- the Galaxy ‘aligner’ stores the BAM file in the central NGS repository and creates just a log file for Galaxy
- the Galaxy ‘extract’ tool knows the location of the NGS repository



allows to share with non-Galaxy users

FMI
Friedrich Miescher Institute
for Biomedical Research

allows to share with non-Galaxy users

```
successfully finished annotation of
test_20130820 to ce6-ceV01-aln2
```

and now the ‘command line geek’ can make a BED file

```
[geek@xenon1 ~]$ extractData.pl -f -s p -m
100 -i test_20130820 ce6-ceV01-aln2 genome |
frag2bed.pl -t -q -U - | head -5
track name=’test_20130820'
chr1    10493    10528    sq39319 1        +
chr1    10736    10764    sq74484 1        +
chr1    11442    11477    sq1340  1        +
chr1    13799    13834    sq84955 1        +
[geek@xenon1 ~]$
```

allows to share with non-Galaxy users

command line

```
extractData.pl -f -s p -m 100 -i
test_20130820 ce6-ceV01-aln2 genome |
frag2bed.pl -t -q -U -
> test_20130820-ce6-ceV01-aln2.bed
```

Galaxy tool definition file

```
extractData.pl -f $strand $maxhits $ignCnts
$sampleSelect.sampleId $genome-$annot-aln2
genome | frag2bed.pl -t -q $summary.ucsc -
> $output
```

and doing the same in Galaxy

Galaxy

Extract data (step 1 of 2)

Sample selection:

Single sample

Extract data (step 2 of 2)

Galaxy / FMI Analyze Data Workflow Shared Data Visualization Admin Help User Using 1.0 GB

track name="test_20130820"

chr1	10493	10528	sq39319	1	+
chr1	10736	10764	sq74484	1	+
chr1	11442	11477	sq1340	1	+
chr1	13799	13834	sq84955	1	+
chr1	13993	14028	sq39245	1	+
chr1	14231	14266	sq65314	1	+
chr1	14234	14269	sq7825	1	+
chr1	14241	14276	sq88621	1	+
chr1	1482(				
chr1	1483				

History

my famous experiment again

7.2 MB

4: test_20130820.bed

3: test_20130820.wig

2: log

```
[geek@xenon1 ~]$ extractData.pl -f -s p -m 100 -i test_20130820
ce6-ceV01-aln2 genome |frag2bed.pl -t -q -U - | head -5
```

Galaxy / FMI Analyze Data Workflow Shared Data Visualization Admin Help User Using 1.0 GB

20130820 (ce6) chr1 13,515 - 15,000

14,000

test_20130820.bed

sq84955 sq39245 sq65314 sq7825 sq88621 sq17523 sq16230

Galaxy will give you

Galaxy

a platform where you can offer your local NGS pipeline with a graphical user interface without compromising the freedom of the command line

FMI
Friedrich Miescher Institute
for Biomedical Research

Acknowledgment

Michael Stadler Christian Hundsrucker

Anita Lerch Tim Roloff Lukas Burger

Dimos Gaidatzis Stefan Grzybek

....and all the people from the "Galaxy"

http://galaxyproject.org

http://www.bioconductor.org/packages/release/bioc/html/QuasR.html

Swiss Institute of Bioinformatics

UNI BASEL

NOVARTIS

Friedrich Miescher Institute for Biomedical Research