

# Using Galaxy to provide Tools for the Analysis of diverse Local Datasets

## 6 Key Insights

Hans-Rudolf Hotz (hrh@fmi.ch)  
Friedrich Miescher Institute for Biomedical Research  
Basel, Switzerland

## background

# Friedrich Miescher Institute

- part of the Novartis Research Foundation
- affiliated institute of Basel University

**314 employees**

(incl. 96 PhD students, 95 Post Docs)

**Epigenetics**

(8 research groups)

**Growth Control**

(7 research groups)

**Neurobiology**

(8 research groups)

## Technology Platforms

Computational Biology – Cell Sorting – Imaging and Microscopy –  
Functional Genomics – Histology – Mass Spectrometry – Protein Structure

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

# background

we have been using Galaxy since early 2008

initially to provide access to a local BioMart database

fall 2009: providing access to our NGS pipeline

see my talk at last year's Galaxy conference

plan for 2010: a new set of tools for microarray analysis

should have been this year's talk

not much progress



no talk



*We are hoping that you can add to the conference by discussing how you have used and/or extended Galaxy in novel and/or widely useful ways. You have had a local installation of Galaxy for far longer than most places (3+ years?), and that long-term view may also have led to some key insights.*



## **6 Key Insights**

- all based on our experience  
(and topics discussed on [galaxy-dev])
- 4 about accessing local data



**Key Insight 1** use Galaxy for the right job  
.....and set up your Galaxy  
server according your needs

who is going to use Galaxy?  
(who is going to use it in the future?)

 include the requested tools

 set up the right hardware

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

**Key Insight 1** use Galaxy for the right job  
.....and set up your Galaxy  
server according your needs  
.....but don't use it just for  
yourself !

*“Galaxy = bringing developers and biologists  
together. Reproducible science is our goal.”*

 **Don't add your crazy tool**  
(only you understand how to use it)

 **Keep it simple !**

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## **Key Insight 2** everything is possible in Galaxy

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

**....and don't blame Galaxy, if your crazy tool doesn't work. You can always write a wrapper.**

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## **Key Insight 2** everything is possible in Galaxy

**Example: How do you access a local MySQL DB containing microarray annotation?**

**a quote from a recent discussion:**

*“conveyor belt captain’ has this excellent RDB module to directly connect to a MySQL DB - why is there no such thing in Galaxy?”*

 **blame Galaxy**

 **or create your own solution**

**FMI**

Friedrich Miescher Institute  
for Biomedical Research



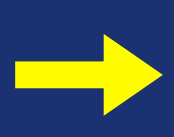
## **Key Insight 2** everything is possible in Galaxy

**Example: How do you access a local MySQL DB containing microarray annotation?**

**talk to your user and find out what information will actually be asked for**

**write a little Perl script using DBI**

 **Keep it simple !**

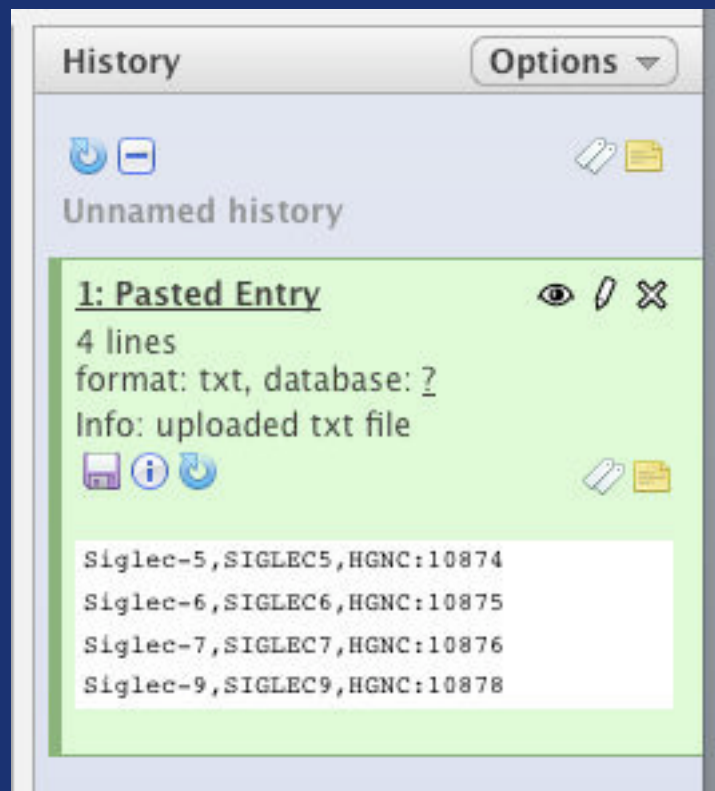
 **if adequate, upload it to the Tool Shed**

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 2 everything is possible in Galaxy

Example: How do you access a local MySQL DB containing microarray annotation?



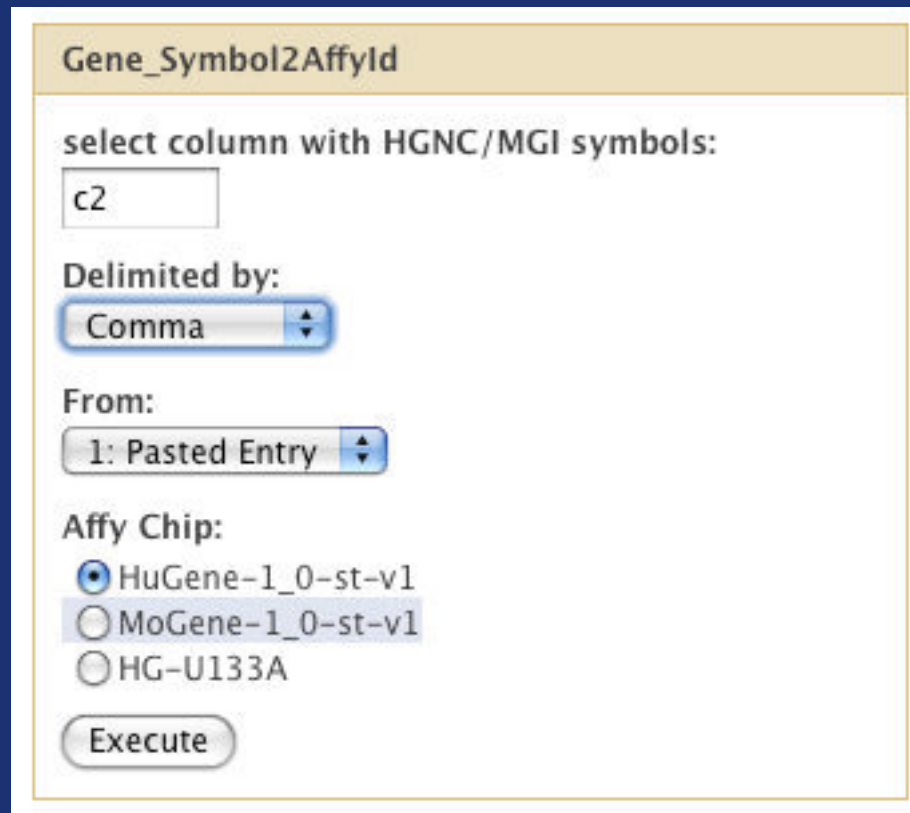
fetching affylds for a list of gene symbols

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 2 everything is possible in Galaxy

Example: How do you access a local MySQL DB containing microarray annotation?



The screenshot shows the configuration interface for the 'Gene\_Symbol2AffyId' tool in Galaxy. The tool title is 'Gene\_Symbol2AffyId'. Below the title, there are several configuration options:

- select column with HGNC/MGI symbols:** A text input field containing 'c2'.
- Delimited by:** A dropdown menu set to 'Comma'.
- From:** A dropdown menu set to '1: Pasted Entry'.
- Affy Chip:** A group of radio buttons with three options: 'HuGene-1\_0-st-v1' (selected), 'MoGene-1\_0-st-v1', and 'HG-U133A'.
- Execute:** A button to run the tool.

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

# Key Insight 2 everything is possible in Galaxy

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

FMI TOOLS

- FMI: Bioinformatics-Support
- FMI: Functional Genomics
  - Gene Symbol2Affyld mapping
- FMI: DNA Restriction Digest
- FMI: DeepSeqRepository
- FMI: qPCR
- FMI: Thoma-lab

```
SIGLEC5 8038877
SIGLEC6 8038861
SIGLEC7 8030789
SIGLEC9 8030782
```

History Options

Unnamed history

- 2: Gene Symbol2Affyld on data 1
- 1: Pasted Entry
  - 4 lines
  - format: txt, database: ?
  - Info: uploaded txt file
  - Siglec-5, SIGLEC5, HGNC:10874
  - Siglec-6, SIGLEC6, HGNC:10875
  - Siglec-7, SIGLEC7, HGNC:10876
  - Siglec-9, SIGLEC9, HGNC:10878

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## **Key Insight 3** Galaxy can help you reducing the storage requirements

a quote from a bioinformatics mailing list:

*“We don’t want to use Galaxy because it produces to much data”*

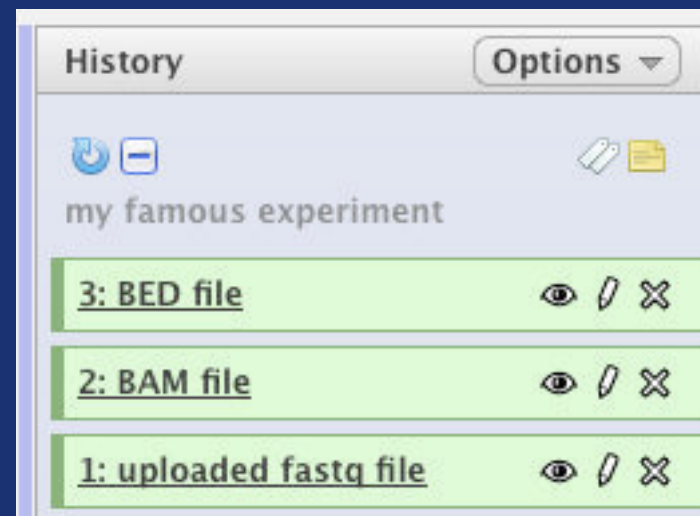
**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 3 Galaxy can help you reducing the storage requirements

a simple NGS workflow

- your famous aligner
- your famous extract tool



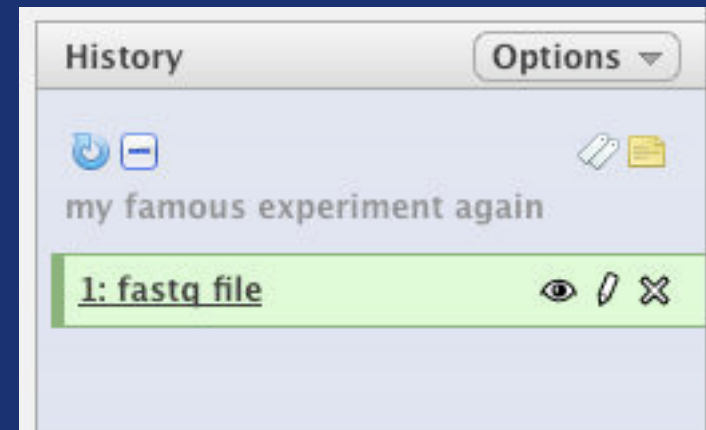
**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 3 Galaxy can help you reducing the storage requirements

make use of the *Galaxy libraries*

*"Link to files without copying into Galaxy"*



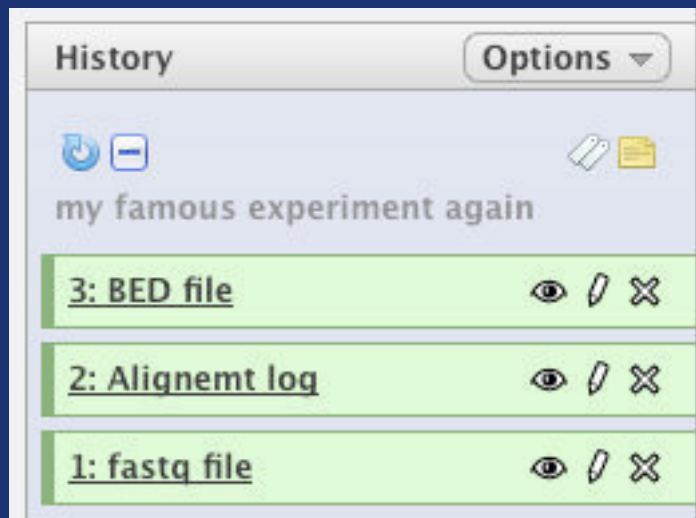
**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 3 Galaxy can help you reducing the storage requirements

### do you really need to store the data in Galaxy?

- do you need the result as a new history item?
- does your tool require a Galaxy history item as input?



- the 'famous aligner' has a wrapper storing the BAM file in the central NGS repository and creating just a log file for Galaxy
- your 'famous extract tool' has a wrapper providing the information about the location of the NGS repository

**FMI**

Friedrich Miescher Institute  
for Biomedical Research



## Key Insight 3 Galaxy can help you reducing the storage requirements

storing data outside of Galaxy makes it easier to share with non-Galaxy users



The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo, the text "Galaxy / FMI Xenon1", and menu items for "Workflow", "Shared Data", "Lab", "Visualization", "Admin", "Help", and "User". The main content area shows a terminal window with the text "successfully finished annotation of sample\_20110518 to dm3-dmV01-aln2". To the right, a "History" panel is visible, featuring an "Options" dropdown and a list of workflow steps:

- 3: BED file (with eye, edit, and delete icons)
- 2: Alignment log (with eye, edit, and delete icons)
- 1: fastq file (with eye, edit, and delete icons)

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 3 Galaxy can help you reducing the storage requirements

successfully finished annotation of sample\_20110518 to dm3-dmV01-aln2

and now the command line geek can do

```
[geek@xenon1 ~]$ extractData.pl -f -s p -m
100 -i mySampleId_20110518 dm3-dmV01-aln2
genome |frag2bed.pl -t -q -U - | head -5
track name='mySampleId_20110518'
chr2L    10493    10528    sq39319  1        +
chr2L    10736    10764    sq74484  1        +
chr2L    11442    11477    sq1340   1        +
chr2L    13799    13834    sq84955  1        +
[geek@xenon1 ~]$
```

## Key Insight 3 Galaxy can help you reducing the storage requirements

### command line

```
extractData.pl -f -s p -m 100 -i  
mySampleId_20110518 dm3-dmV01-aln2 genome |  
frag2bed.pl -t -q -U -
```

### Galaxy tool definition file

```
#elif ($summary.mode=="bed") #extractData.pl  
-f $strand $maxhits $ignCnts  
$sampleSelect.sampleId $genome-$annot-aln2  
genome | frag2bed.pl -t -q $summary.ucsc -  
> $output
```

## Key Insight 4 Galaxy is able to use data outside of ~/database/files/

previous example:

mySampleId\_20110518

```
[geek@xenon1 ~]$ ls /NGS/common_samples/  
...  
mySampleId_20110505  
mySampleId_20110509  
mySampleId_20110518  
...  
[geek@xenon1 ~]$
```

directory listing  
in the repository

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

# Key Insight 4 Galaxy is able to use data outside of ~/database/files/

directory listing in Galaxy

The screenshot shows a web interface for selecting samples. At the top, there is a header 'Select samples' in a light brown box. Below it, the text 'Available samples:' is followed by two buttons: 'Select All' and 'Unselect All'. A list of samples follows, each with a checkbox and a label. The labels include sample IDs and associated data types or versions.

Sample ID	Associated Data
<input type="checkbox"/> mySampleId_20110417	(dm3 : dmV01)
<input type="checkbox"/> mySampleId_20110426	(dm3 : dmV01)
<input type="checkbox"/> mySampleId_20110505	(dm3, hg18 : dmV01, hgV01)
<input type="checkbox"/> mySampleId_20110509	(dm3 : dmV01)
<input type="checkbox"/> mySampleId_20110518	(dm3 : dmV01)

## Key Insight 4 Galaxy is able to use data outside of ~/database/files/

use “dynamic\_options”

```
<inputs>
  <param name="samples" type="select"
    label="Available samples"
    help="Use tickboxes to select samples"
    display="checkboxes" multiple="true"
    dynamic_options="ds_samples()" />
</inputs>
```

```
...
<code file="NGS_code.py" />
```

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 4 Galaxy is able to use data outside of ~/database/files/

“NGS\_code.py”

```
sampleDir = "/NGS/common_samples/"

def ds_samples( ):
    """List available deepseq samples"""
    l = os.listdir(sampleDir)
    l.sort()
    samples = [(s,s,False) for s in l
if os.path.isdir(sampleDir + s)]
    return samples
```

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## Key Insight 5 there is privacy in Galaxy

if you enforce log-in,  
you can use one of the  
pre-defined variables:

```
$userEmail  
($__user_email__)
```

```
<command interpreter="perl">  
  access_data.pl $dir $userEmail $output  
</command>
```

**FMI**

Friedrich Miescher Institute  
for Biomedical Research



## Key Insight 5 there is privacy in Galaxy

“access\_data.pl”

```
my $priv_user = "geek\@fmi.ch";
```

```
my $file = $ARGV[0];
```

```
my $user = $ARGV[1];
```

```
unless ($user eq $priv_user) {  
    print "you don't have access to this data";  
    exit; }  
}
```

```
open (FILE, "$file");
```



**Keep it simple !**

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## **Key Insight 6** you can change the hardware of a production server

<http://galaxy.fmi.ch>  
(only visible within the FMI)

- external authentication
- MySQL
- 'April 8<sup>th</sup>' changeset
- server is used for Galaxy and individual logins

- virtual server
- 4 cores
- 32GB RAM
- python 2.5.2
- storage via NFS



- real server
- four quad-core Intel X7350
- 128GB RAM
- python 2.6.5
- direct attached storage

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

**Key Insight 6** you can change the hardware of a production server, as long as you keep the 'database/' directory and the SQL DB in sync.

**this might sound trivial**

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

## **Key Insight 6** you can change the hardware

**this might sound trivial:**

- make a copy of the MySQL DB
- copy the complete galaxy directory to the new server (make sure you keep the path)
- point the new galaxy server to the MySQL DB copy and start it
  - > due to the higher Python version, news eggs were downloaded
  - > all python code was re-compiled
- test the new server (while the old one is still in use)
- stop the old server
- rsync ~/galaxy\_dist/database/files/
- point the new galaxy server to the 'live' MySQL DB and re-start it

***actually, it is trivial!***

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

# Summary

use Galaxy for the right job

everything is possible in Galaxy

Galaxy can help you reducing the storage requirements

Galaxy is able to use data outside of ~/database/files/

there is privacy in Galaxy

you can change the hardware



**Keep it simple !**

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

# Summary

## Mission

running a Bioinformatics Helpdesk

## Vision

I don't have to do anything

## Strategy

*Use Galaxy!*

**FMI**

Friedrich Miescher Institute  
for Biomedical Research

# Acknowledgment

## *Computational Biology*

- Michael Stadler

## *Functional Genomics*

- Tim Roloff

## *IT Support*

- Thomas Übermeier

*....and all the people from the “Galaxy”*

**FMI**

Friedrich Miescher Institute  
for Biomedical Research