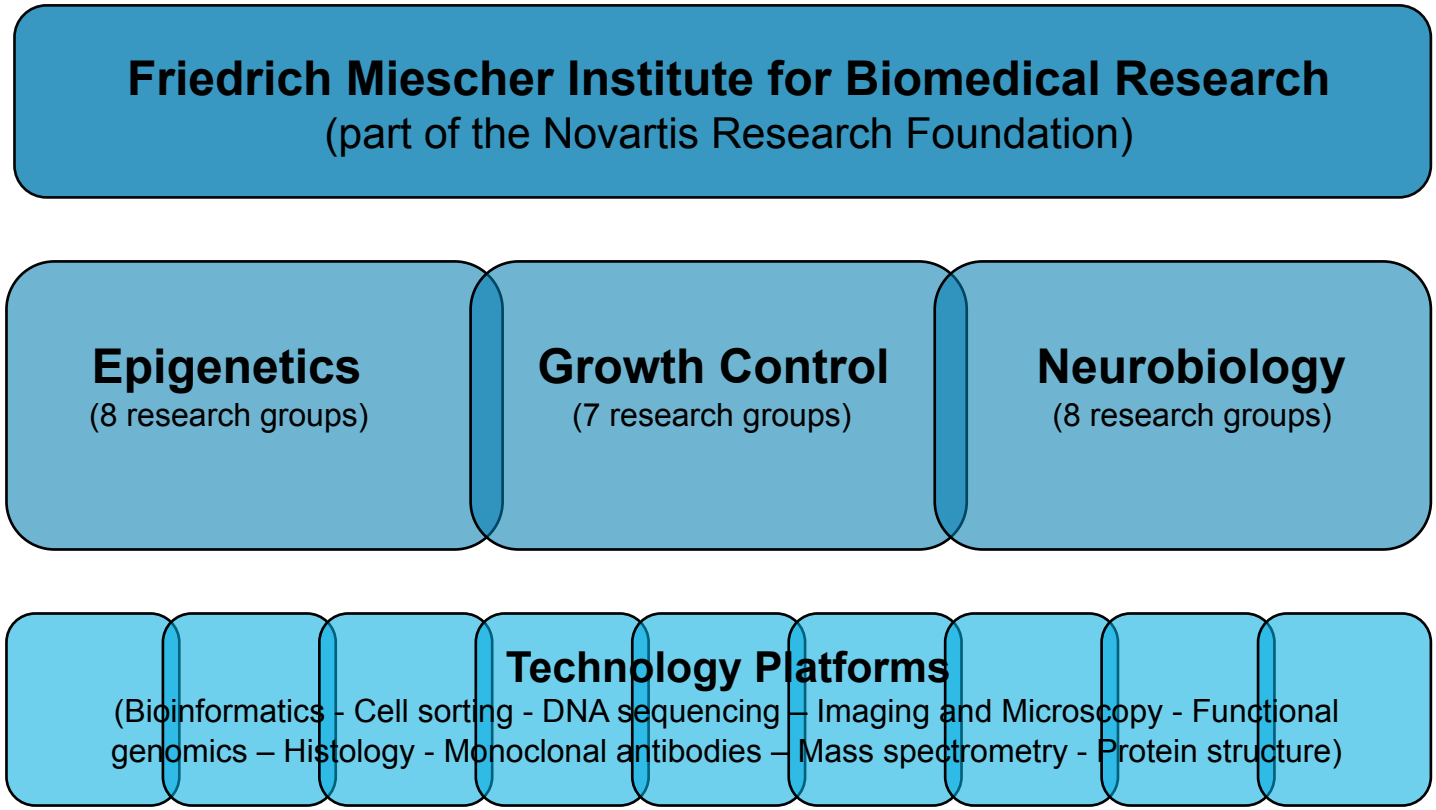
# Do-It-Yourself Bioinformatics with the FMI Galaxy Server

Hans-Rudolf Hotz (hrh@fmi.ch)

Friedrich Miescher Institute for Biomedical Research
Basel, Switzerland
http://www.fmi.ch/




**~ 300 employees (~40 nationalities)**
**99 PhD students and 86 Post Docs**

FMI
Friedrich Miescher Institute
for Biomedical Research

# Bioinformatics / Computational Biology

**four FTE providing support for....**

**the “average” lab scientist, using computers to:**

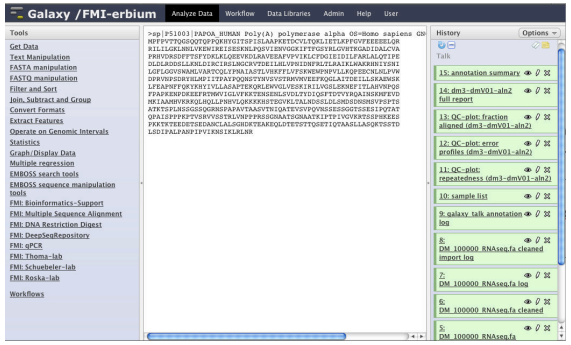
?

**the “modern” lab scientist, using computers to:**

**draw plasmids**

**do BLAST searches**

**use Excel**




**analyze NGS data with Bioconductor**

**write Perl scripts**

FMI
Friedrich Miescher Institute for Biomedical Research

# the FMI Galaxy Servers:

## *http://galaxy.fmi.ch*

- dedicated, virtualized linux server
- 4 cores, 32GB RAM
- storage via NFS (IBM DS4000)
- changeset: 3545 (March)
- external authentication

## *http://galaxy-dev.fmi.ch*

- IBM x3755
- 2 dual core AMD, 16GB RAM
- storage via NFS (IBM DS4000)
- changeset: 3545

## *temporary test servers*

on desktops/laptops and/or
different port numbers

FMI
Friedrich Miescher Institute
for Biomedical Research

# the FMI Galaxy Server (*http://galaxy.fmi.ch*)

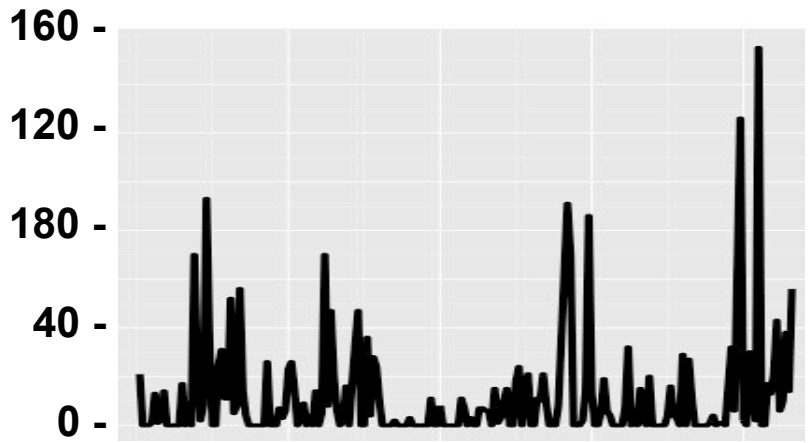
Tools

- Get Data
- Text Manipulation
- FASTA manipulation
- FASTQ manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Multiple regression
- EMBOSS search tools
- EMBOSS sequence manipulation tools
- FMI: Bioinformatics-Support
- FMI: Multiple Sequence Alignment
- FMI: DNA Restriction Digest
- FMI: DeepSeqRepository
- FMI: qPCR
- FMI: Thoma-lab
- FMI: Schuebeler-lab
- FMI: Roska-lab

**67 users (have logged in at least once)**

**~ 13 jobs executed / day (Oct. till May)**

| | |
|---|---|
| 24 % | upload.py |
| 4.2 % | gops_intersect.py |
| 4.1 % | cutWrapper.py |
| 3.3 % | fixedValueColumn.py |
| 2.8 % | *motif_search.pl* |
| 2.6 % | data_source.py |
| 2.6 % | remove_beginning.pl |
| | ... |




number of jobs /day

FMI
Friedrich Miescher Institute for Biomedical Research

# self-written/added tools

*motif search*

*uploading data*

*Next Gen Sequencing*

***accessing data***
***"outside of Galaxy"***

# motif search

**request:** show me all human UniProtKB/Swiss-Prot entries which contain the following motif: “W[ST]XW”

→ **EMBOSS:fuzzpro**

**.....what about a pre-installed copy of human UniProtKB/Swiss-Prot entries?**
**.....can you provide information about subcellular location?**
**.....provide domain and 3D-structure info**
**.....and I want to know the sequence to the left and right of the motif**

## Motif Search

**Your motif:**

W[ST]XW

see below for examples

**protein file:**

- (•) human (reviewed) UniProtKB
- ( ) mouse (reviewed) UniProtKB
- ( ) C. elegans (reviewed) UniProtKB
- ( ) P. falciparum (isolate 3D7) (reviewed) UniProtKB
- ( ) "human TSR proteins"
- ( ) "pompep" - limited results mark-up

**display additional ammino acids to the left and the right (in lower case) of the found motif (max 6 amino acids, not possible for pompep):**

4

**display proteins without hits:**

- (•) "no"
- ( ) "yes" - you could end up with >20,000 lines!

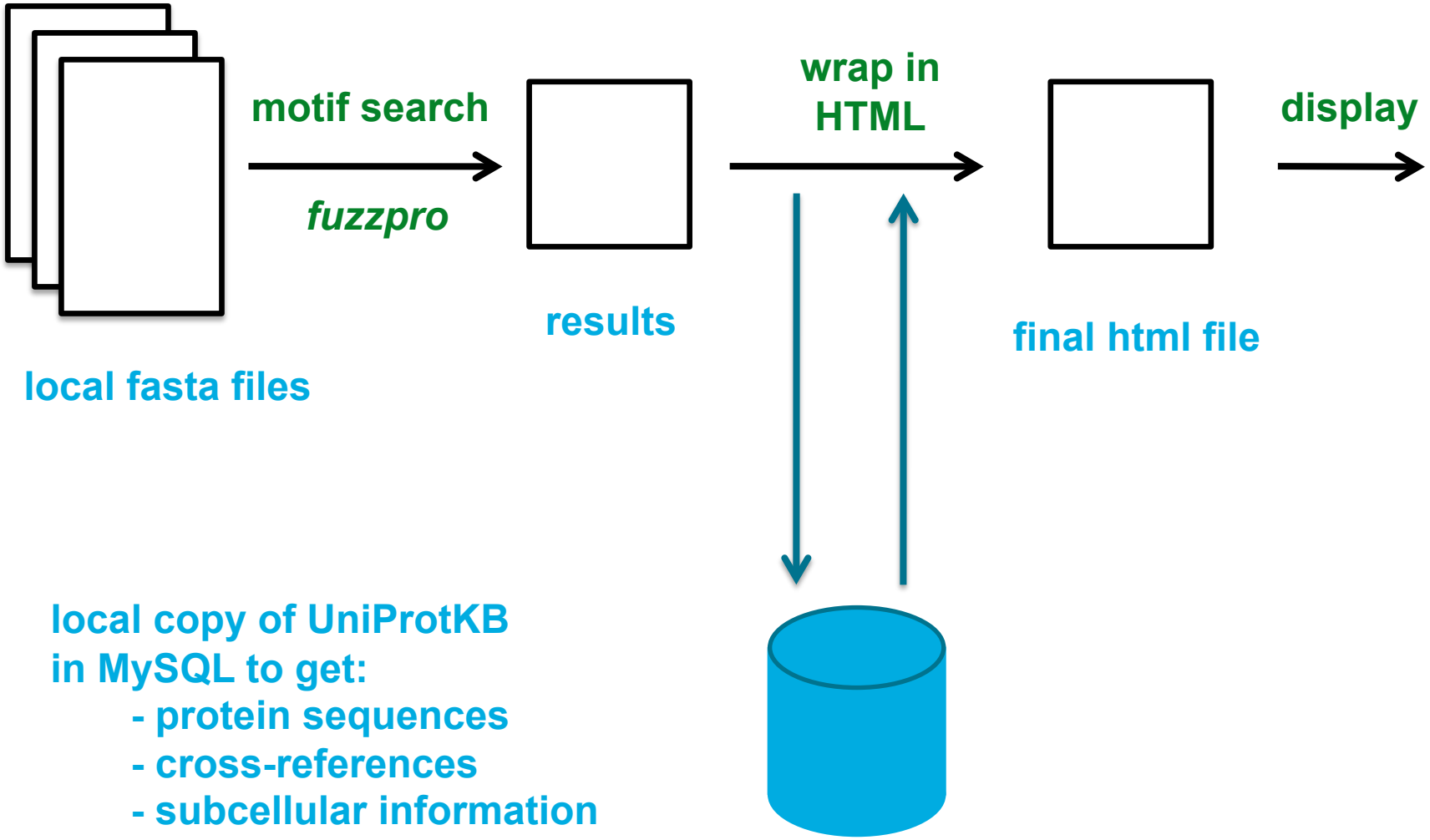
Execute

FMI
Friedrich Miescher Institute
for Biomedical Research

Result for "**W[ST]XW**" search in UniProtKB/Swiss-Prot (release 2010_04) for ***Homo sapiens***

| Accession | Description | start | end | sequence | overlaps with | subcellular location | 3D-structure info |
|---|---|---|---|---|---|---|---|
| P46098 | 5-hydroxytryptamine receptor 3A | 472 | 475 | lvml**WSIW**qya | - | Cell junction, synapse, postsy | - |
| Q99758 | ATP-binding cassette sub-family A member 3 | 305 | 308 | swlh**WSAW**fllf | - | Membrane; Multi-pass membrane | - |
| P22570 | NADPH:adrenodoxin oxidoreductase, mitochondrial | 11 | 14 | rwwg**WSAW**prtr | - | Mitochondrion matrix. | - |
| Q5T2P9 | Arf-GAP with GTPase, ANK repeat and PH domain-containing protein 10 | 646 | 649 | lltg**WTSW**pemp | - | - | - |
| Q5VTM2 | Arf-GAP with GTPase, ANK repeat and PH domain-containing protein 9 | 691 | 694 | lltg**WTSW**pemp | - | - | - |
| Q8N1P7 | Absent in melanoma 1-like protein | 194 | 197 | eypd**WSHW**ggyd | Beta/gamma crystallin 'Greek key' 4 (169 - 211) | - | HSSP |
| P26442 | Autocrine motility factor receptor, isoform 1 | 153 | 156 | sgvd**WTAW**gggr | - | Membrane; Single-pass type I m | - |
| P04424 | Argininosuccinate lyase | 169 | 172 | qpir**WSHW**ilsh | - | - | 1AOS 1K62 |
| Q9HBZ2 | Aryl hydrocarbon receptor nuclear translocator 2 | 657 | 660 | psev**WSQW**qsqh | - | Nucleus (Potential). | - |
| P15848 | Arylsulfatase B | 319 | 322 | rgrk**WSLW**eggv | - | Lysosome. | 1FSU |
| Q9ULK2 | Ataxin-7-like protein 1 | 19 | 22 | lgkp**WSSW**idaa | - | - | - |

# how does it work.....




local copy of UniProtKB
in MySQL to get:

- protein sequences
- cross-references
- subcellular information
- domain information

FMI
Friedrich Miescher Institute
for Biomedical Research

# motif search

## similar tool for genomes

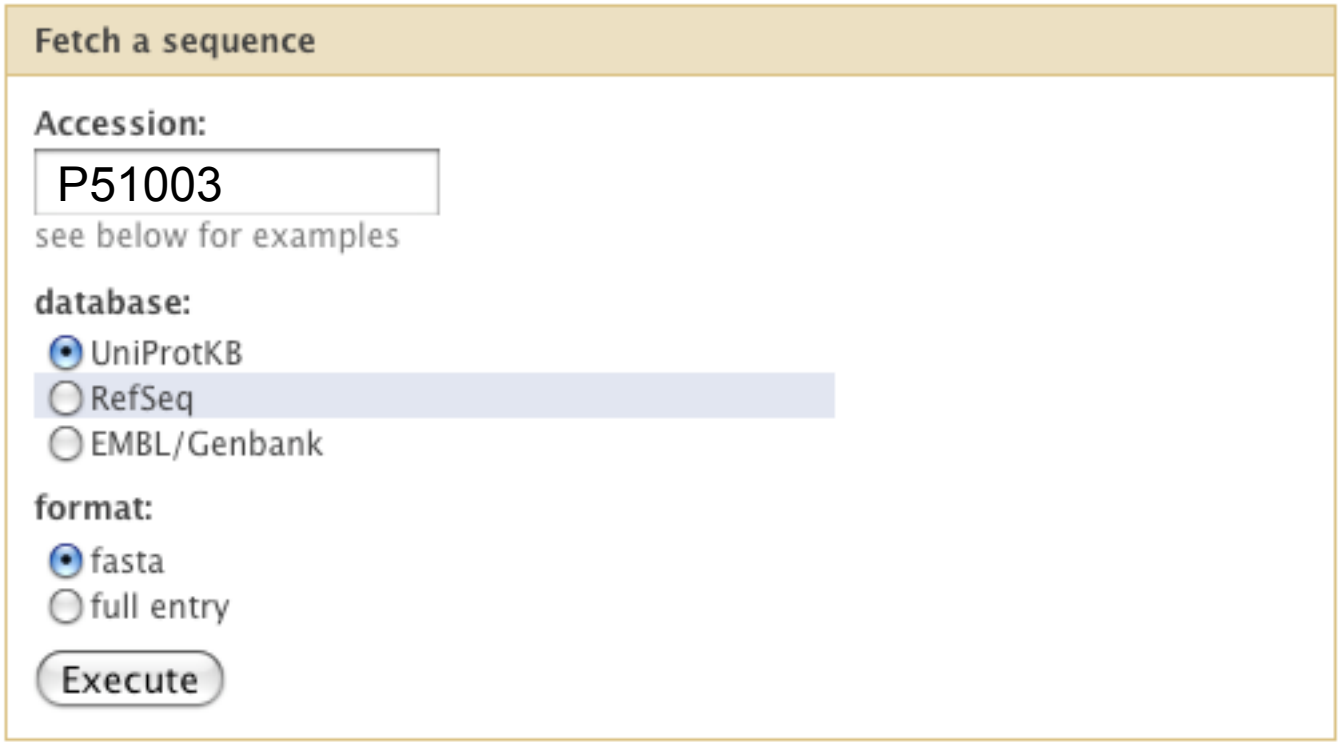
- using EMBOSS:fuzznuc
- connecting to public ensembl database
  to get neighboring genes, exon/intron, etc

→ too slow!

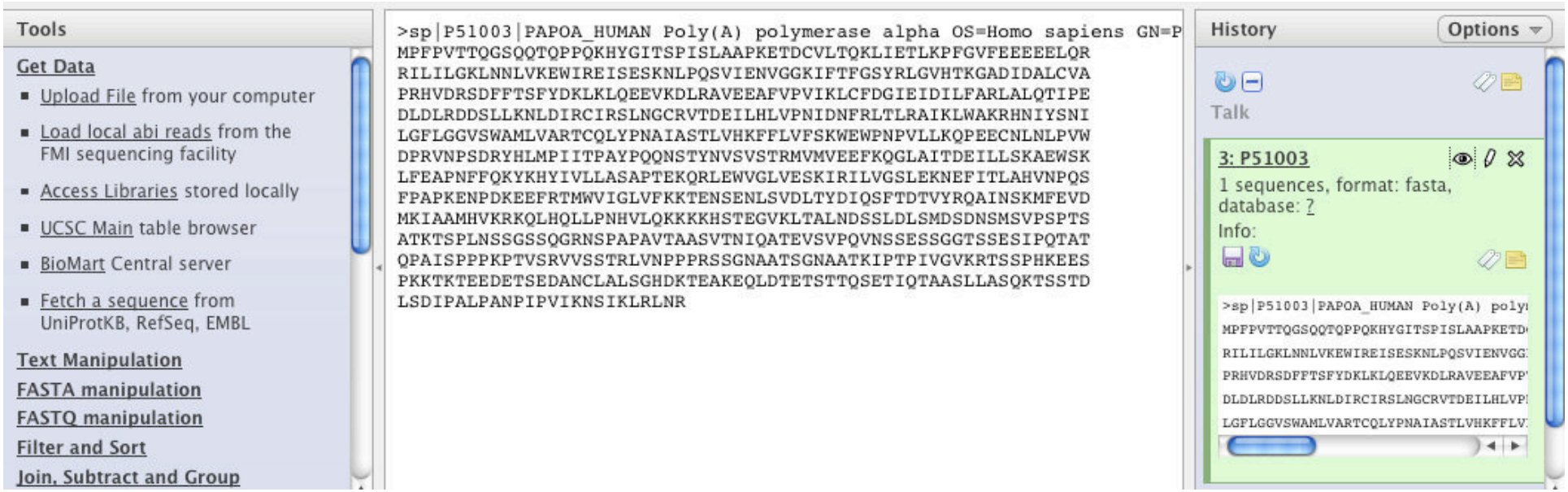
- currently re-writing using the
  Bioconductor package : Biostring

# uploading data

## individual ‘public’ sequences



# uploading data



# uploading data

## ‘ABI PRISM 3730’ sequencer data (yes, they are still in use.....)

- everybody (ie. each desktop) has access to the reads
  from his/her research group

```
drwxr-x---+  2 dnaseq gpielage  3484 2010-05-05 13:30 pielage.j
drwxr-x---+  2 dnaseq grijli    1134 2010-04-19 13:30 rijli.f
drwxr-x---+ 11 dnaseq groska   10201 2010-03-26 09:00 roska.b
drwxr-x---+ 16 dnaseq gschub    9712 2010-05-04 14:30 schubeler.d
```

```
02620_D03_193434SCD_013.ab1
02620_D03_193434SCD_013.seq
02620_D04_193435SCD_013.ab1
02620_D04_193435SCD_013.seq
02620_D05_193436SCD_021.seq
02620_D05_193436SCD_021.ab1
```

→ individual file uploads possible,
but not very efficient

FMI
Friedrich Miescher Institute
for Biomedical Research

# uploading data

**request:**

- keep the directory structure
- maintain security

**problem:**

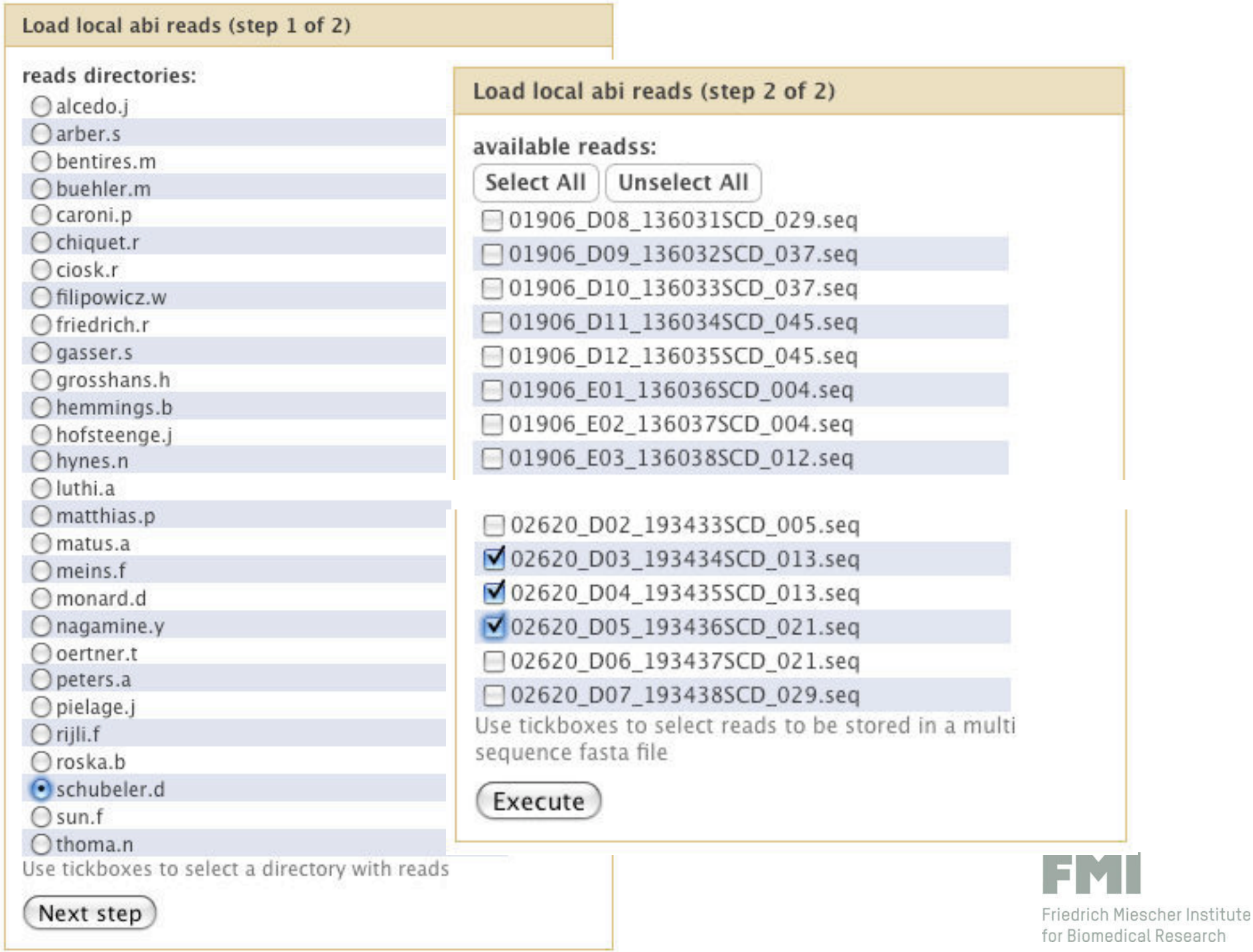
- using ‘Data libraries’ and assigning ‘roles’ to each file seemed un-doable
- providing different tools (ie. GUI) for each user is not possible

**solution:**

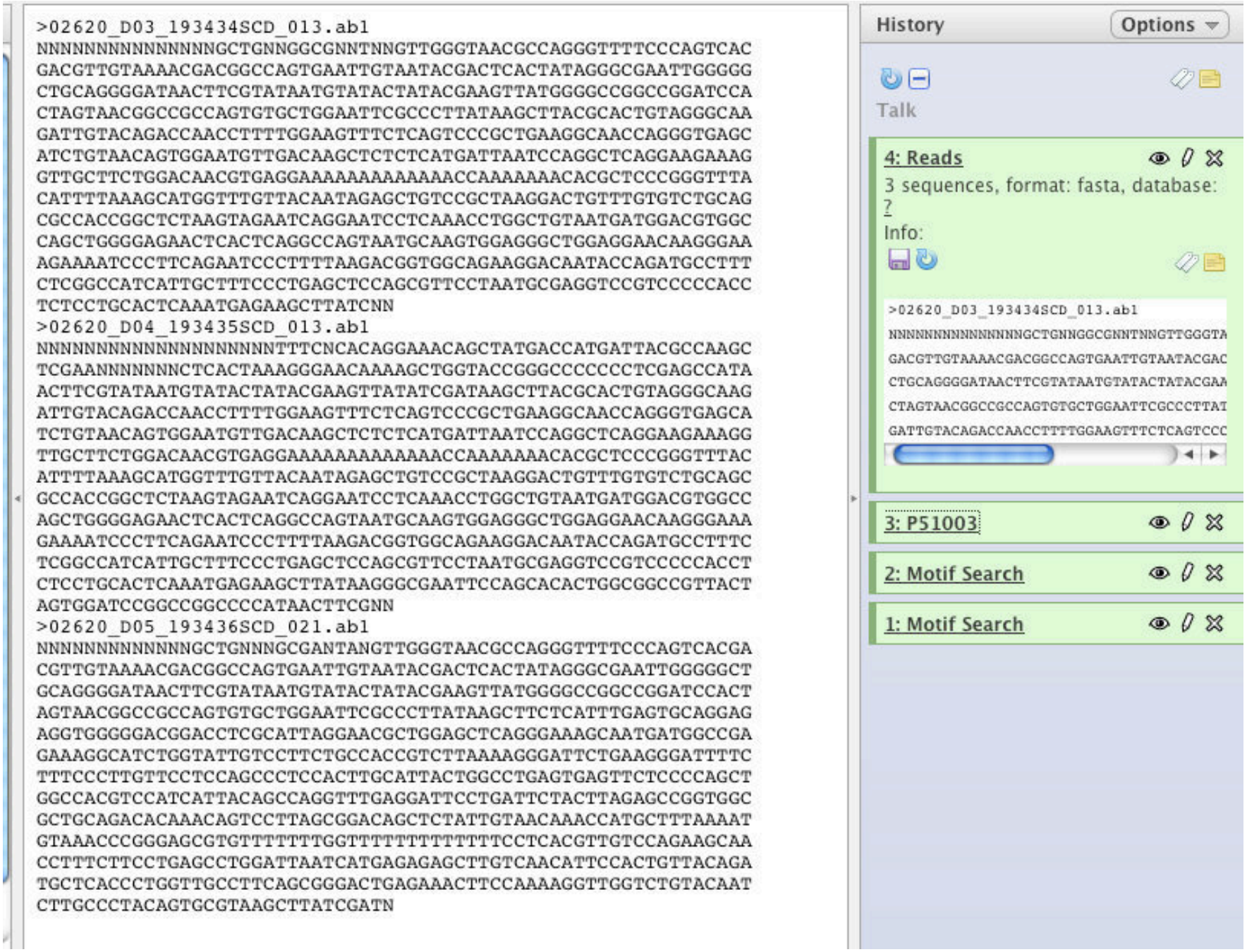
- give “Galaxy” access to all the reads
- use `$userEmail` to provide security

FMI
Friedrich Miescher Institute
for Biomedical Research

```
<tool id="load_reads" name="Load local abi reads" version="1">
  <description>from the FMI sequencing facility</description>
  <command interpreter="perl">load_abi_reads.pl $reads $dir $userEmail
$output
  </command>
  <inputs>
    <page>
      <param name="dir" type="select" label="reads directories"
       help="Use tickboxes to select a directory with reads"
       display="radio"
       dynamic_options="ds_readDirOptions()"/>
    </page>
    <page>
      <param name="reads" type="select" label="available reads"
       help="Use tickboxes to select reads"
       display="checkboxes" multiple="true"
       dynamic_options="ds_readsOptions(dir)"/>
    </page>
  </inputs>
  <outputs>
    <data format="fasta" name="output" label="Reads" />
  </outputs>
  <code file="more_code_for_load_abi_reads.py" />
  <help>
  </help>
</tool>
```

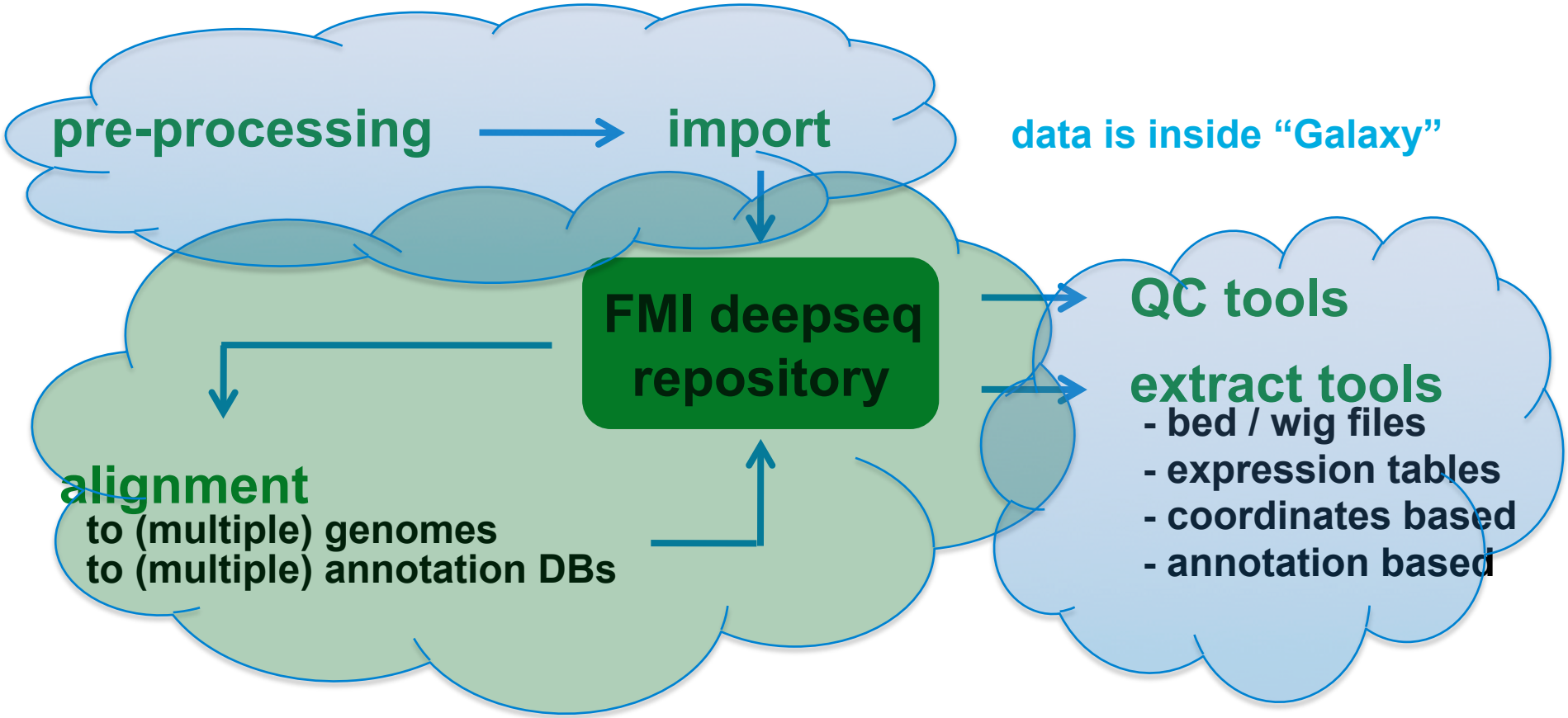



FMI
Friedrich Miescher Institute for Biomedical Research


>02620_D03_193434SCD_013.ab1
NNNNNNNNNNNNNNNGCTGNNGGCGNNTNNGTTGGGTAACGCCAGGGTTTTCCCAGTCAC
GACGTTGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGGGG
CTGCAGGGGATAACTTCGTATAATGTATACTATACGAAGTTATGGGGCCGGCCGGATCCA
CTAGTAACGGCCGCCAGTGTGCTGGAATTCGCCCTTATAAGCTTACGCACTGTAGGGCAA
GATTGTACAGACCAACCTTTTGGAAGTTTCTCAGTCCCGCTGAAGGCAACCAGGGTGAGC
ATCTGTAACAGTGGAATGTTGACAAGCTCTCTCATGATTAATCCAGGCTCAGGAAGAAAG
GTTGCTTCTGGACAACGTGAGGAAAAAAAAAAAAACCAAAAAAACACGCTCCCGGGTTTA
CATTTTAAAGCATGGTTTGTTACAATAGAGCTGTCCGCTAAGGACTGTTTGTGTCTGCAG
CGCCACCGGCTCTAAGTAGAATCAGGAATCCTCAAACCTGGCTGTAATGATGGACGTGGC
CAGCTGGGGAGAACTCACTCAGGCCAGTAATGCAAGTGGAGGGCTGGAGGAACAAGGGAA
AGAAAATCCCTTCAGAATCCCTTTTAAGACGGTGGCAGAAGGACAATACCAGATGCCTTT
CTCGGCCATCATTGCTTTCCCTGAGCTCCAGCGTTCCTAATGCGAGGTCCGTCCCCCACC
TCTCCTGCACTCAAATGAGAAGCTTATCNN
>02620_D04_193435SCD_013.ab1
NNNNNNNNNNNNNNNNNNNNTTTCNCACAGGAAACAGCTATGACCATGATTACGCCAAGC
TCGAANNNNNNNCTCACTAAAGGGAACAAAAGCTGGTACCGGGCCCCCCCTCGAGCCATA
ACTTCGTATAATGTATACTATACGAAGTTATATCGATAAGCTTACGCACTGTAGGGCAAG
ATTGTACAGACCAACCTTTTGGAAGTTTCTCAGTCCCGCTGAAGGCAACCAGGGTGAGCA
TCTGTAACAGTGGAATGTTGACAAGCTCTCTCATGATTAATCCAGGCTCAGGAAGAAAGG
TTGCTTCTGGACAACGTGAGGAAAAAAAAAAAAACCAAAAAAACACGCTCCCGGGTTTAC
ATTTTAAAGCATGGTTTGTTACAATAGAGCTGTCCGCTAAGGACTGTTTGTGTCTGCAGC
GCCACCGGCTCTAAGTAGAATCAGGAATCCTCAAACCTGGCTGTAATGATGGACGTGGCC
AGCTGGGGAGAACTCACTCAGGCCAGTAATGCAAGTGGAGGGCTGGAGGAACAAGGGAAA
GAAAATCCCTTCAGAATCCCTTTTAAGACGGTGGCAGAAGGACAATACCAGATGCCTTTC
TCGGCCATCATTGCTTTCCCTGAGCTCCAGCGTTCCTAATGCGAGGTCCGTCCCCCACCT
CTCCTGCACTCAAATGAGAAGCTTATAAGGGCGAATTCCAGCACACTGGCGGCCGTTACT
AGTGGATCCGGCCGGCCCCATAACTTCGNN
>02620_D05_193436SCD_021.ab1
NNNNNNNNNNNNNGCTGNNNGCGANTANGTTGGGTAACGCCAGGGTTTTCCCAGTCACGA
CGTTGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGGGGCT
GCAGGGGATAACTTCGTATAATGTATACTATACGAAGTTATGGGGCCGGCCGGATCCACT
AGTAACGGCCGCCAGTGTGCTGGAATTCGCCCTTATAAGCTTCTCATTTGAGTGCAGGAG
AGGTGGGGGACGGACCTCGCATTAGGAACGCTGGAGCTCAGGGAAAGCAATGATGGCCGA
GAAAGGCATCTGGTATTGTCCTTCTGCCACCGTCTTAAAAGGGATTCTGAAGGGATTTTC
TTTCCCTTGTTCCTCCAGCCCTCCACTTGCATTACTGGCCTGAGTGAGTTCTCCCCAGCT
GGCCACGTCCATCATTACAGCCAGGTTTGAGGATTCCTGATTCTACTTAGAGCCGGTGGC
GCTGCAGACACAAACAGTCCTTAGCGGACAGCTCTATTGTAACAAACCATGCTTTAAAAT
GTAAACCCGGGAGCGTGTTTTTTTGGTTTTTTTTTTTTTCCTCACGTTGTCCAGAAGCAA
CCTTTCTTCCTGAGCCTGGATTAATCATGAGAGAGCTTGTCAACATTCCACTGTTACAGA
TGCTCACCCTGGTTGCCTTCAGCGGGACTGAGAAACTTCCAAAAGGTTGGTCTGTACAAT
CTTGCCCTACAGTGCGTAAGCTTATCGATN
History
Options
Talk
4: Reads
3 sequences, format: fasta, database: ?
Info:
>02620_D03_193434SCD_013.ab1
NNNNNNNNNNNNNNNGCTGNNGGCGNNTNNGTTGGGTA
GACGTTGTAAAACGACGGCCAGTGAATTGTAATACGAC
CTGCAGGGGATAACTTCGTATAATGTATACTATACGAA
CTAGTAACGGCCGCCAGTGTGCTGGAATTCGCCCTTAT
GATTGTACAGACCAACCTTTTGGAAGTTTCTCAGTCCC
3: P51003
2: Motif Search
1: Motif Search

# Next Gen Sequencing

## FMI pipeline for tag based applications (Illumina GA II)




*....just a bunch of Perl scripts!*

data is outside Galaxy

FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing

**FMI: DeepSeqRepository**

- Remove adaptor sequences from 5'/3' ends of reads
- Import sample into deepSeq repository
- Trigger annoation (alignment) of a sample against reference sequences
- Select samples from the deepSeq repository
- Create quality control plots (repeatedness, error rate and aligned fraction)
- Create full report that classifies reads into biological types
- Extract data from the repository (expression levels, etc.)

**the pipeline ‘starts’ with a fasta file:**

**‘upload file’**

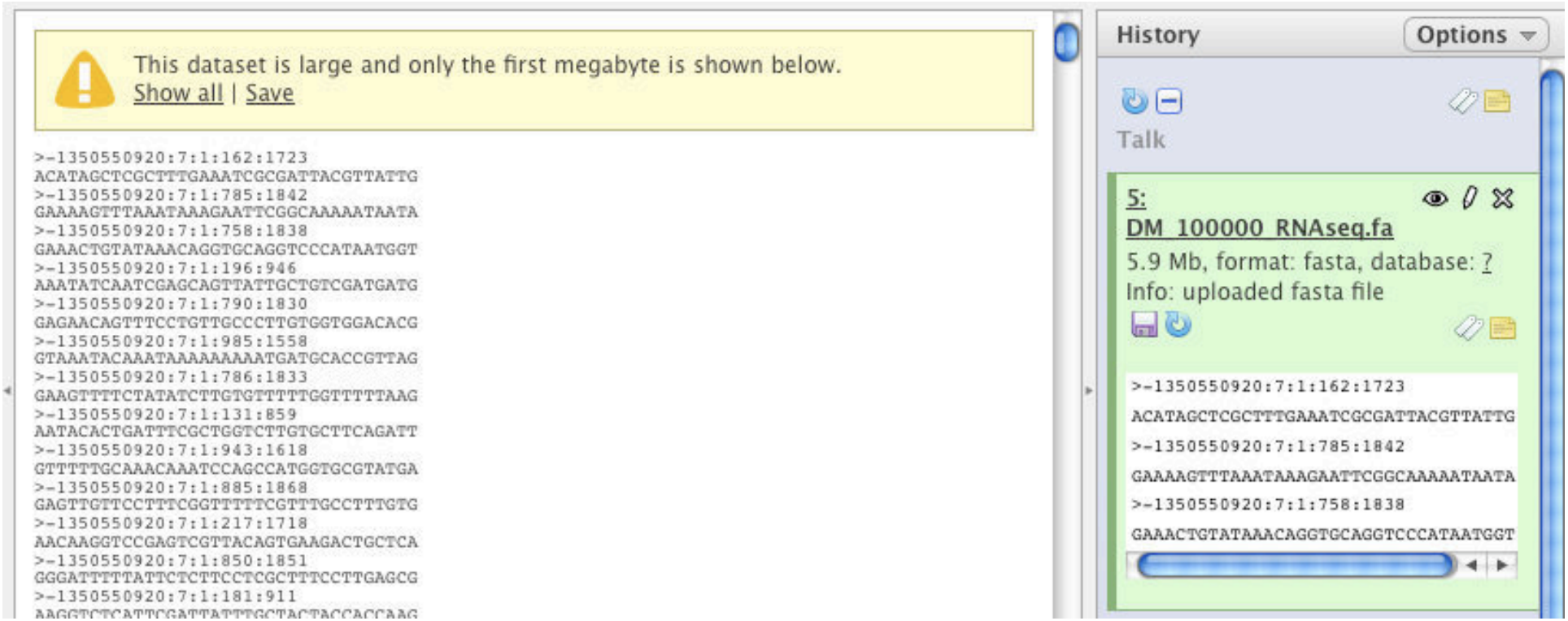
**‘Access Libraries’**

→ **data is inside “Galaxy”**

FMI
Friedrich Miescher Institute
for Biomedical Research

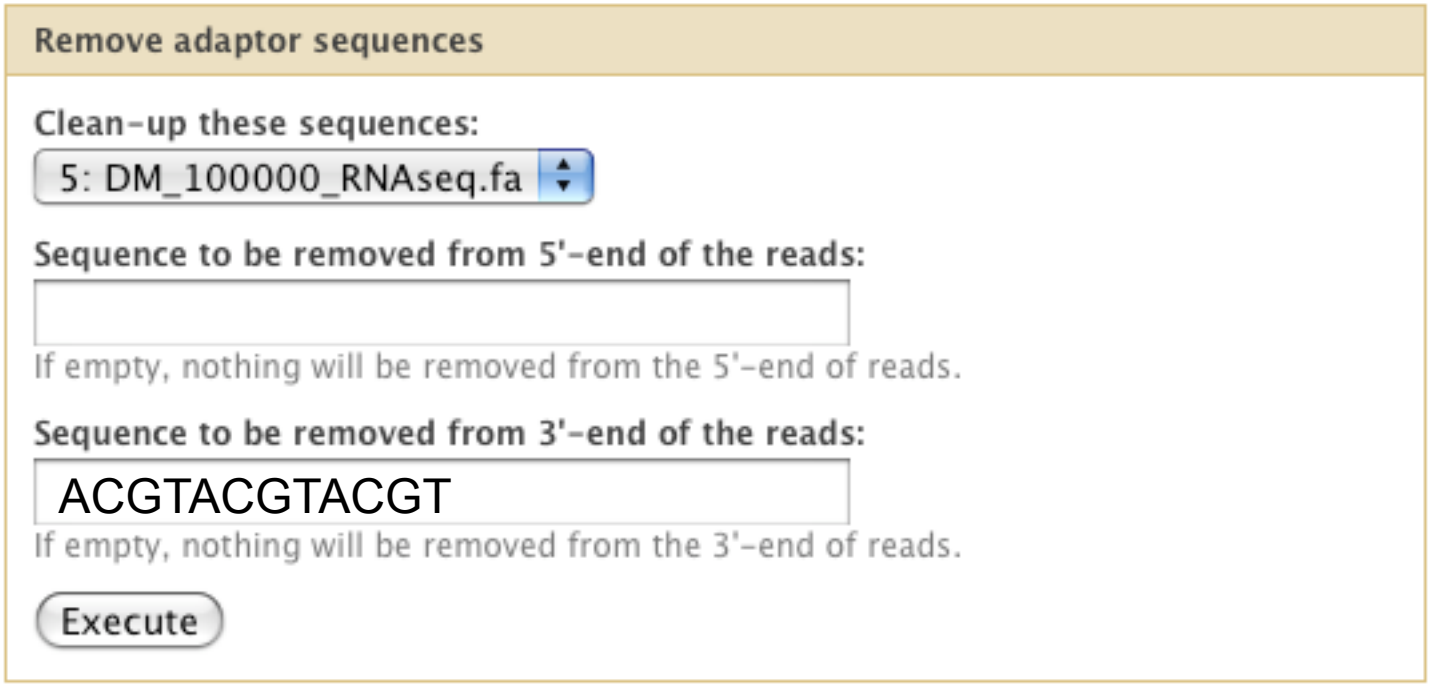
# Next Gen Sequencing

## data is inside “Galaxy”




FMI
Friedrich Miescher Institute
for Biomedical Research

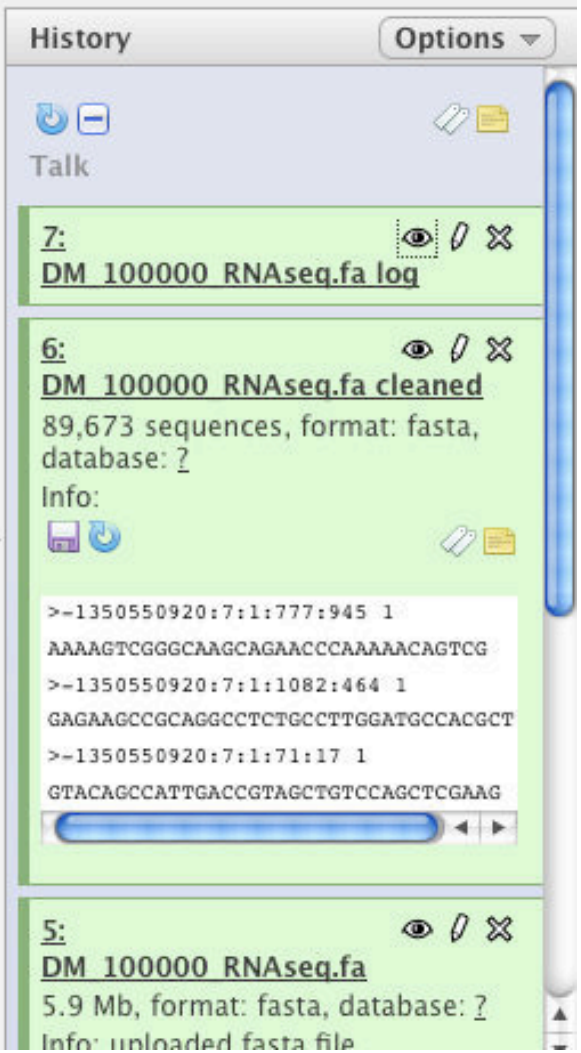
# Next Gen Sequencing



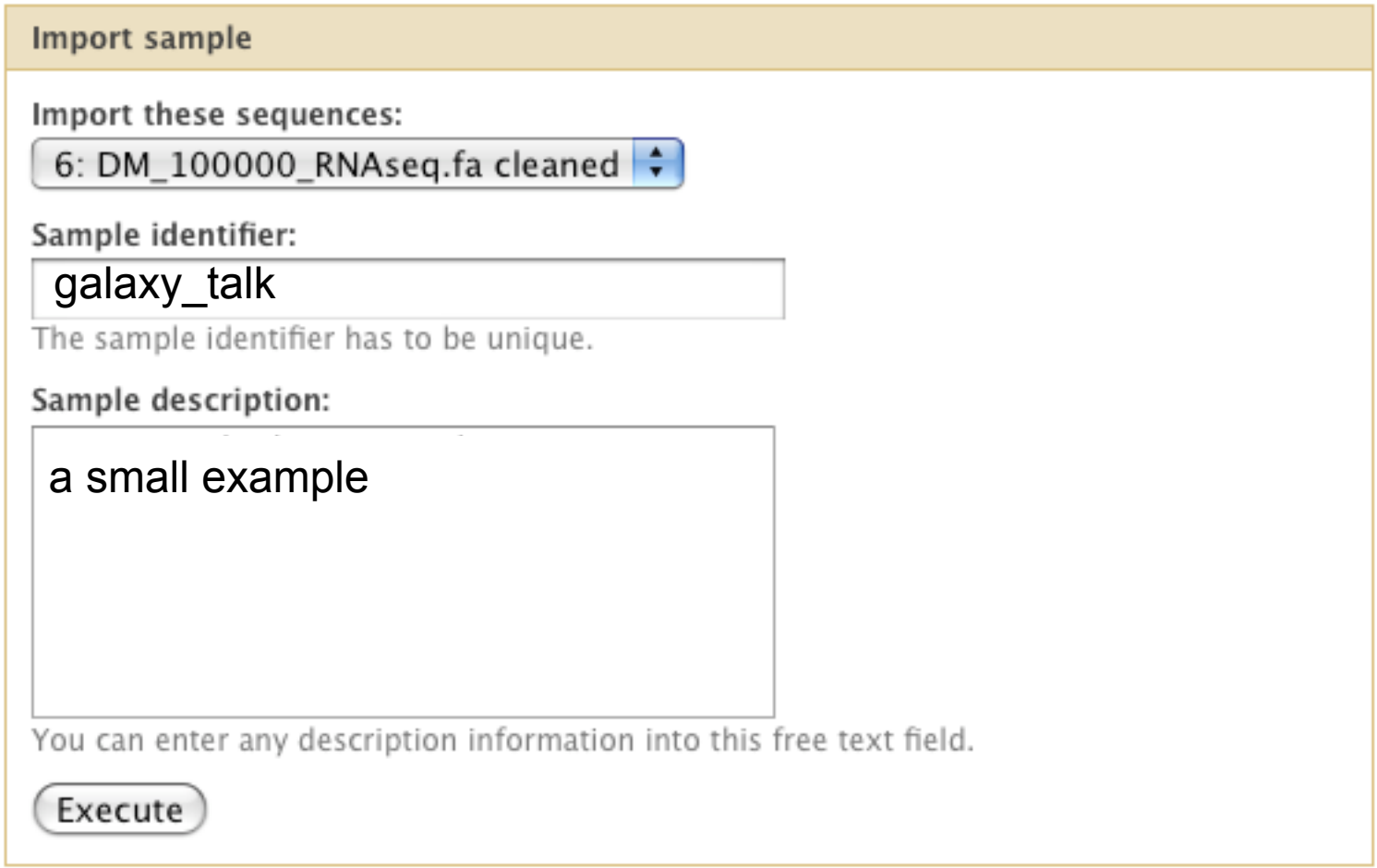
# Next Gen Sequencing

## data is inside “Galaxy”

```
# SUMMARY:
# Number of sequences:
#                          input sequences :    100000 (100.0%)
#                         unique sequences :     89948 ( 89.9%)
#
# Adaptor removal:
#              full matches to 5'-adaptor :         0 (   0.0%)
#    inexact suffix matches to 5'-adaptor :         0 (   0.0%)
#      exact suffix matches to 5'-adaptor :         0 (   0.0%)
#              full matches to 3'-adaptor :        23 (   0.0%)
#    inexact prefix matches to 3'-adaptor :       116 (   0.1%)
#      exact prefix matches to 3'-adaptor :     32615 ( 32.6%)
#
#             total matches to 5'-adaptor :         0 (   0.0%)
#             total matches to 3'-adaptor :     32754 ( 32.8%)
#
#         unique sequences (w/o adaptors) :     89946 ( 89.9%)
#
# Sequence filtering:
#          5'-adaptor alignment too short :         0 (   0.0%)
#          3'-adaptor alignment too short :         0 (   0.0%)
#                               too short :        16 (   0.0%)
#            too many non-base characters :       223 (   0.2%)
#                              low entropy :        44 (   0.0%)
#
# Final sequences:
#                                  passed :     99717 ( 99.7%)
```



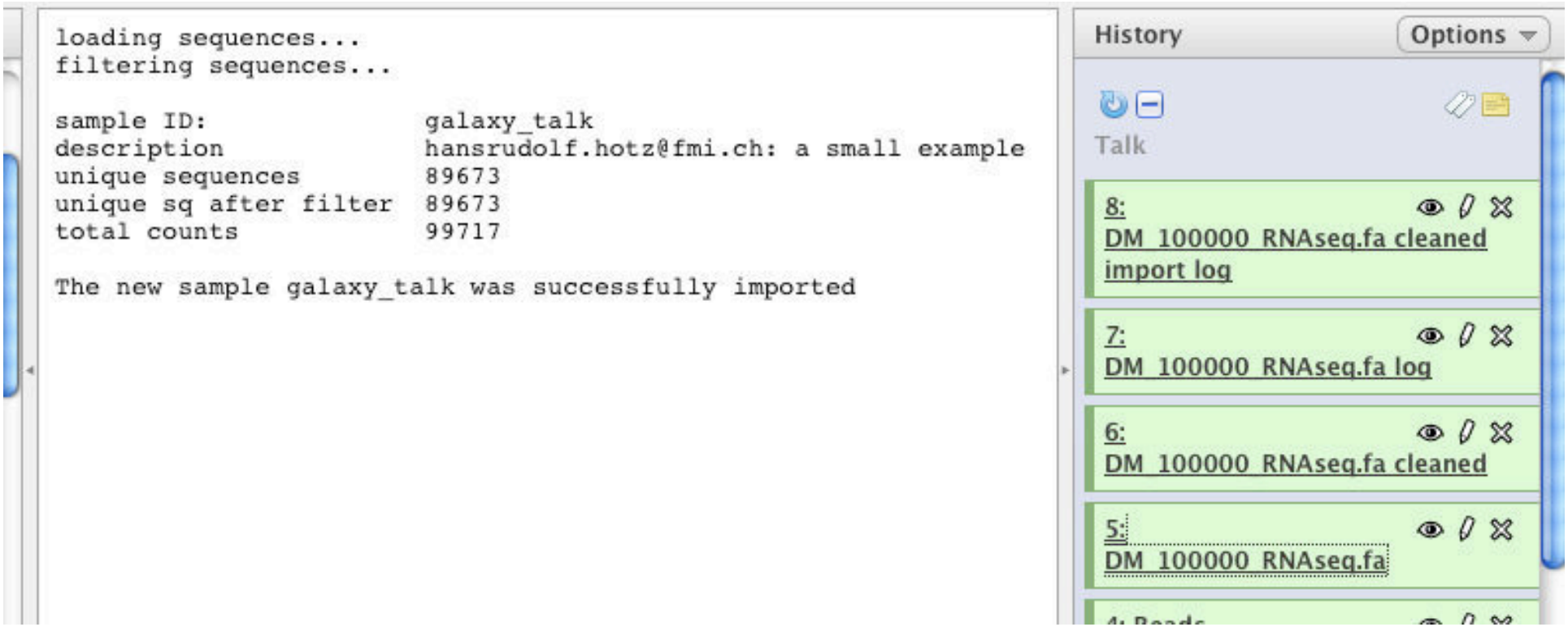
# Next Gen Sequencing




FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing

## data is outside “Galaxy”




FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing

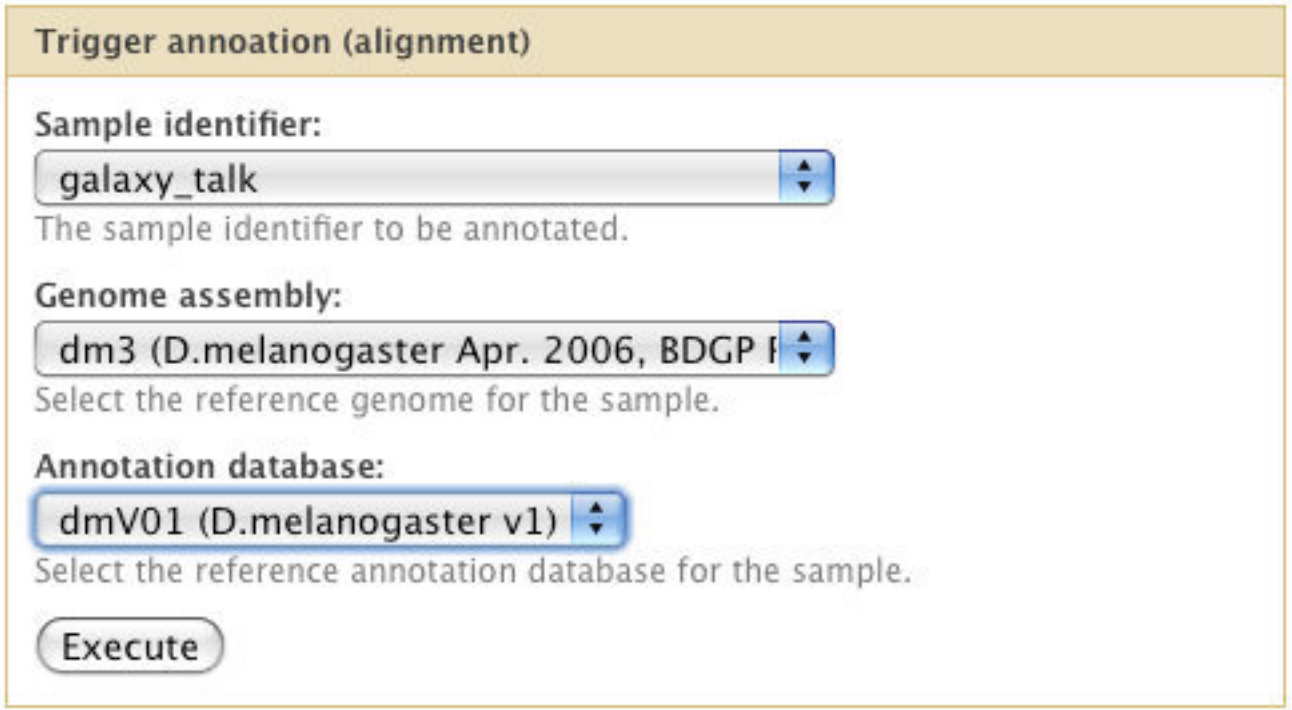
data is outside “Galaxy”

```
    ...
    galaxy   gbioinfo 4096 2010-05-11 14:50 galaxy_talk
    galaxy   gbioinfo 4096 2010-05-06 13:46 galaxy_workshop
    galaxy   gbioinfo 4096 2009-11-18 17:46 GM_8_com_H3K9me3
    galaxy   gbioinfo 4096 2010-05-04 15:55 ME_K27me3
    haruhotz gbioinfo 4096 2010-05-05 09:03 ME_K27me3_2
    lienertf gbioinfo 4096 2009-12-08 09:08 CA_H3K9me2a_quant
    ...
cd galaxy_talk

    galaxy gbioinfo      40 2010-05-06 13:46 description.txt
    galaxy gbioinfo    4096 2010-05-06 13:48 errorLogs
    galaxy gbioinfo    4096 2010-05-06 13:48 fullReports
    galaxy gbioinfo    4096 2010-05-06 13:48 mappings
    galaxy gbioinfo 6304977 2010-05-06 13:46 seqs.tab
    galaxy gbioinfo 6304977 2010-05-06 13:46 seqs_unfi.tab

head description.txt

hansrudolf.hotz@fmi.ch: a small example
```

FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing

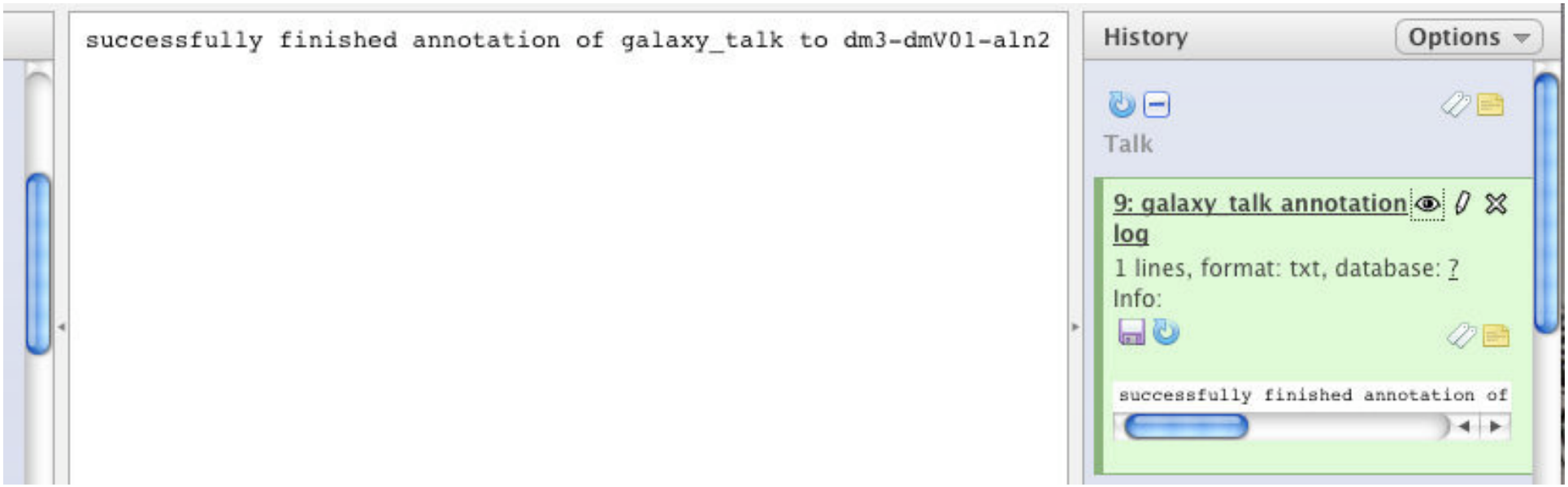
**data is outside “Galaxy”**



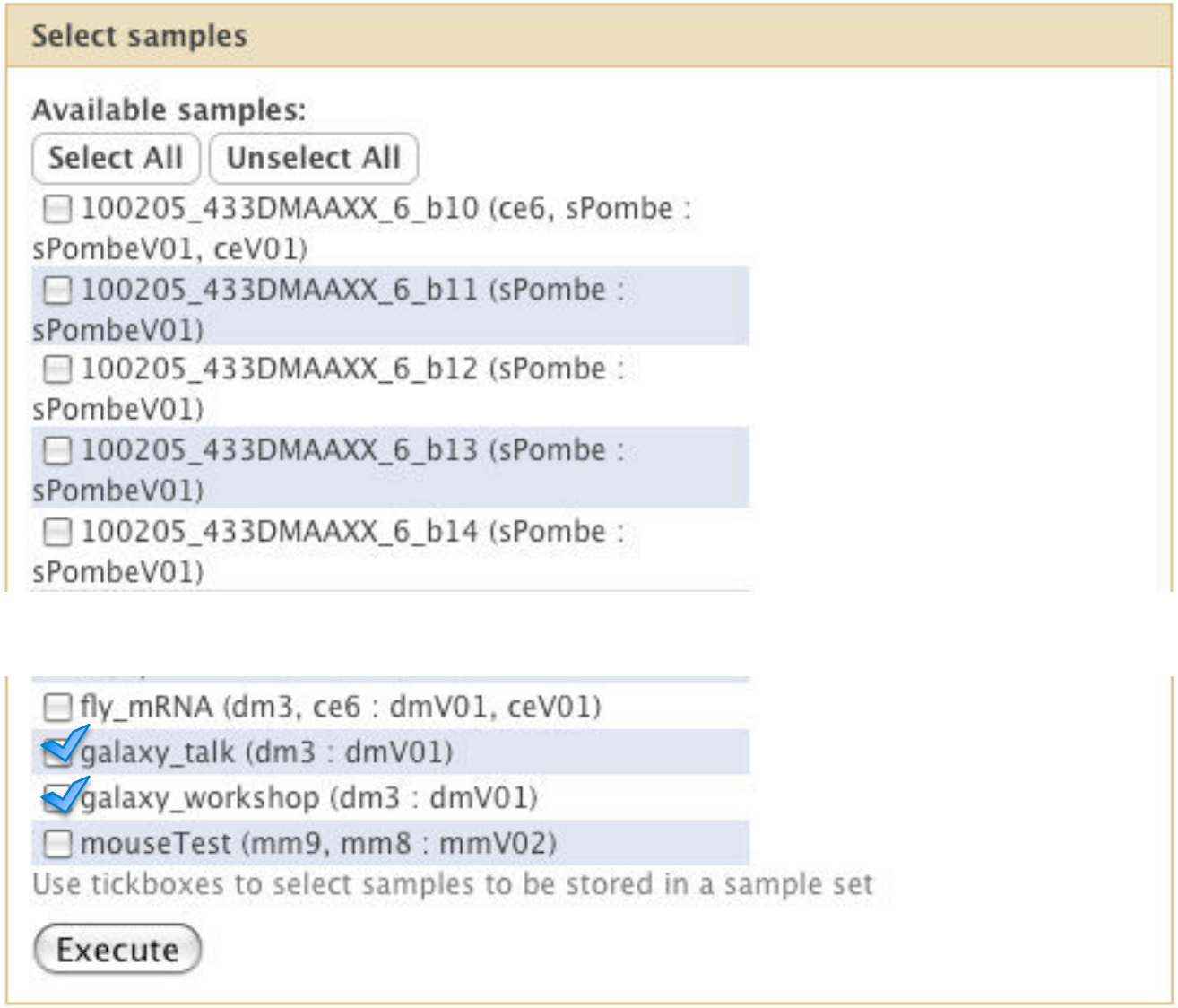

FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing

## data is outside “Galaxy”

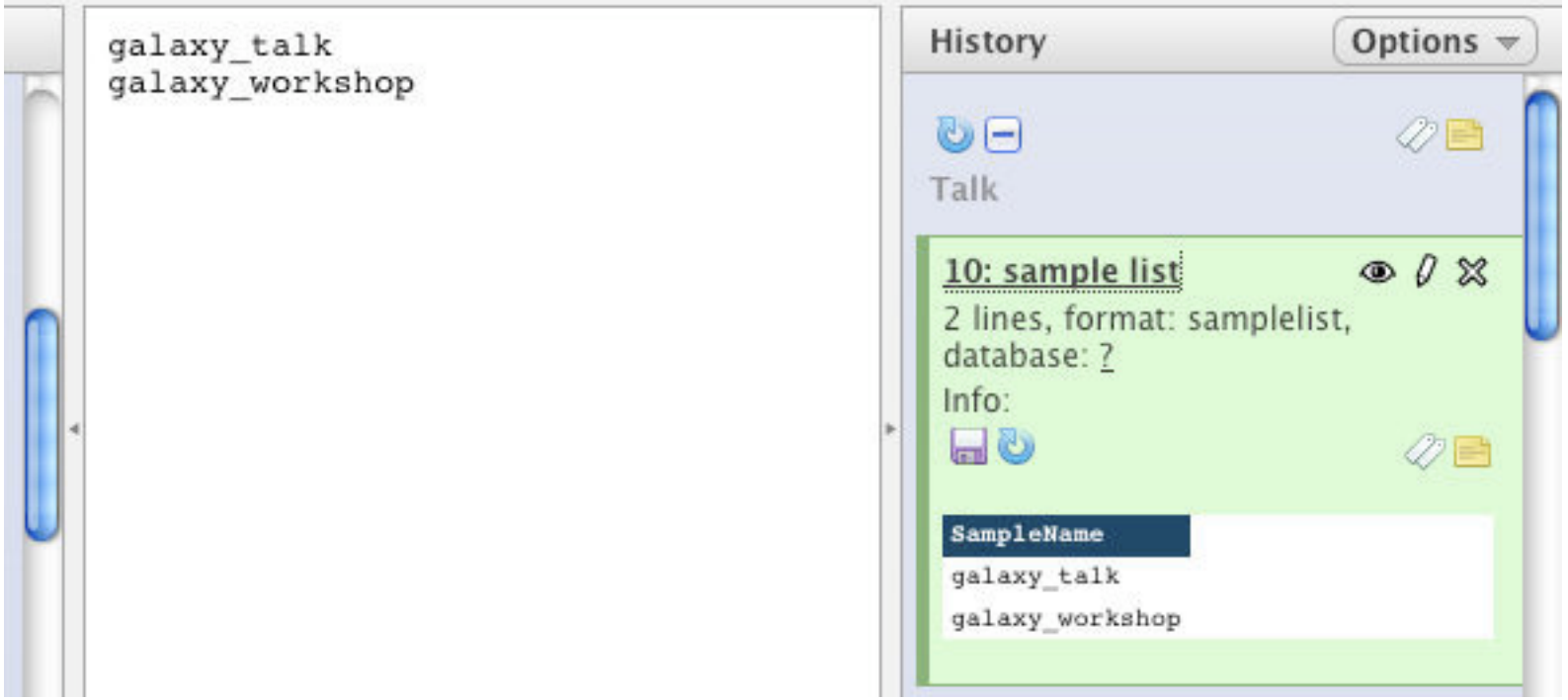


# Next Gen Sequencing



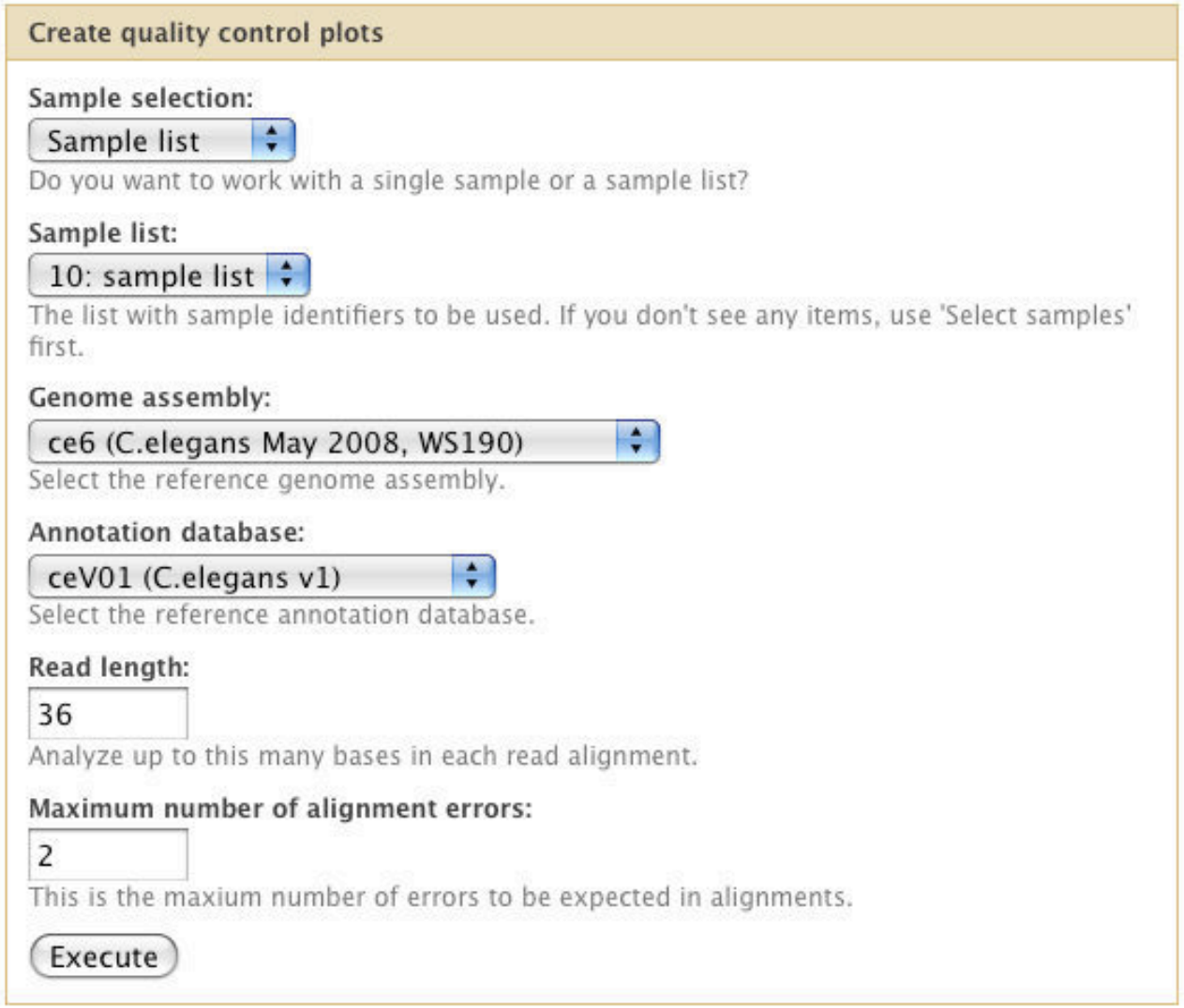

FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing



# Next Gen Sequencing

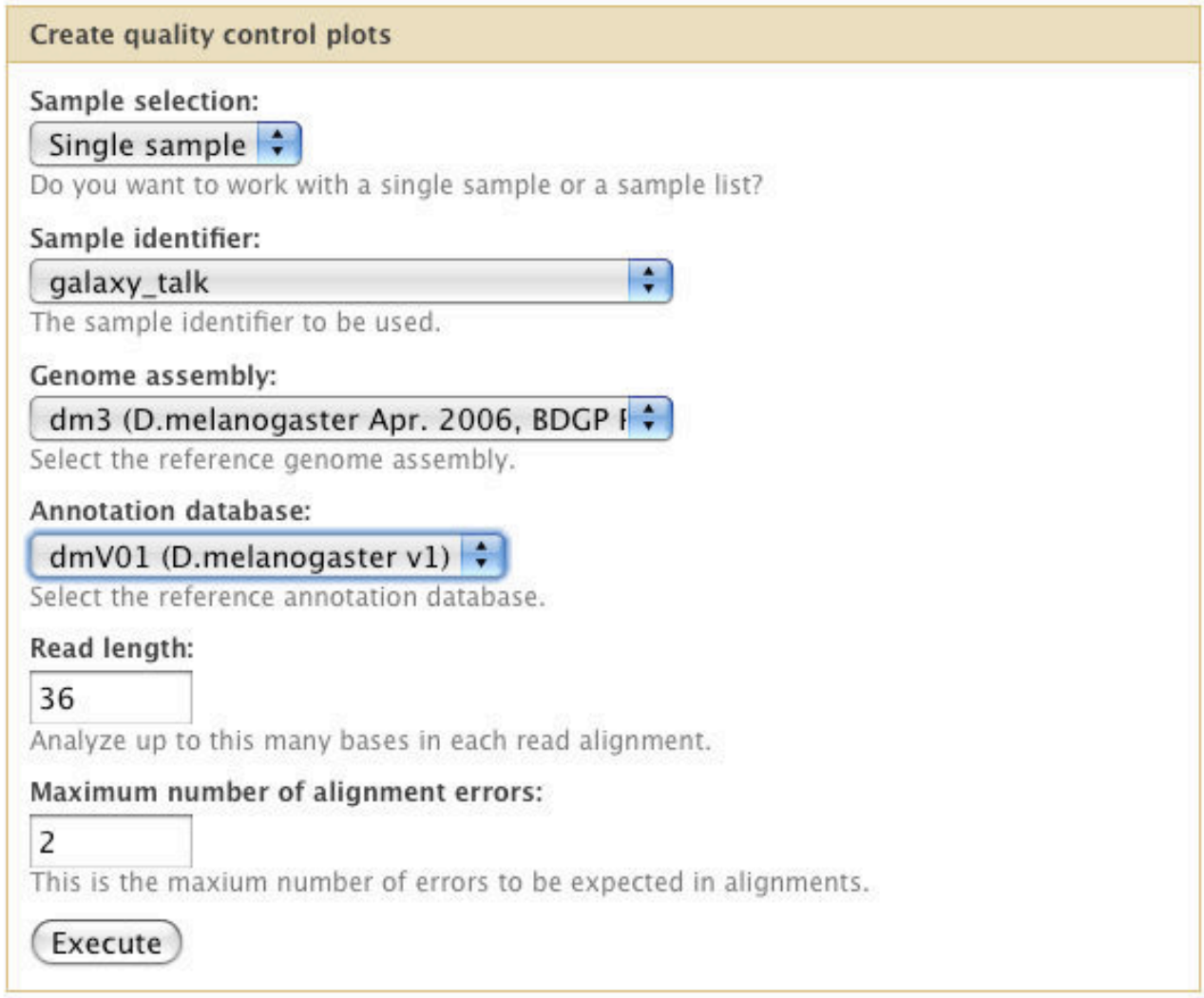
## data is outside “Galaxy”




FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing

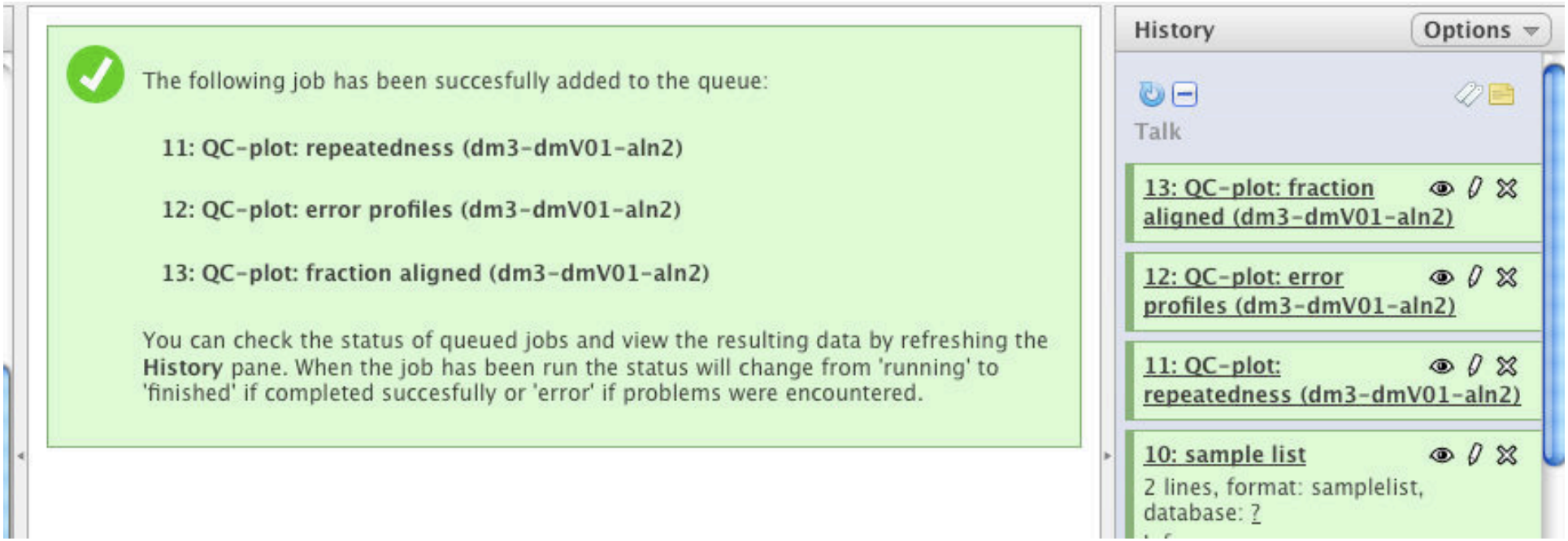
## data is outside “Galaxy”



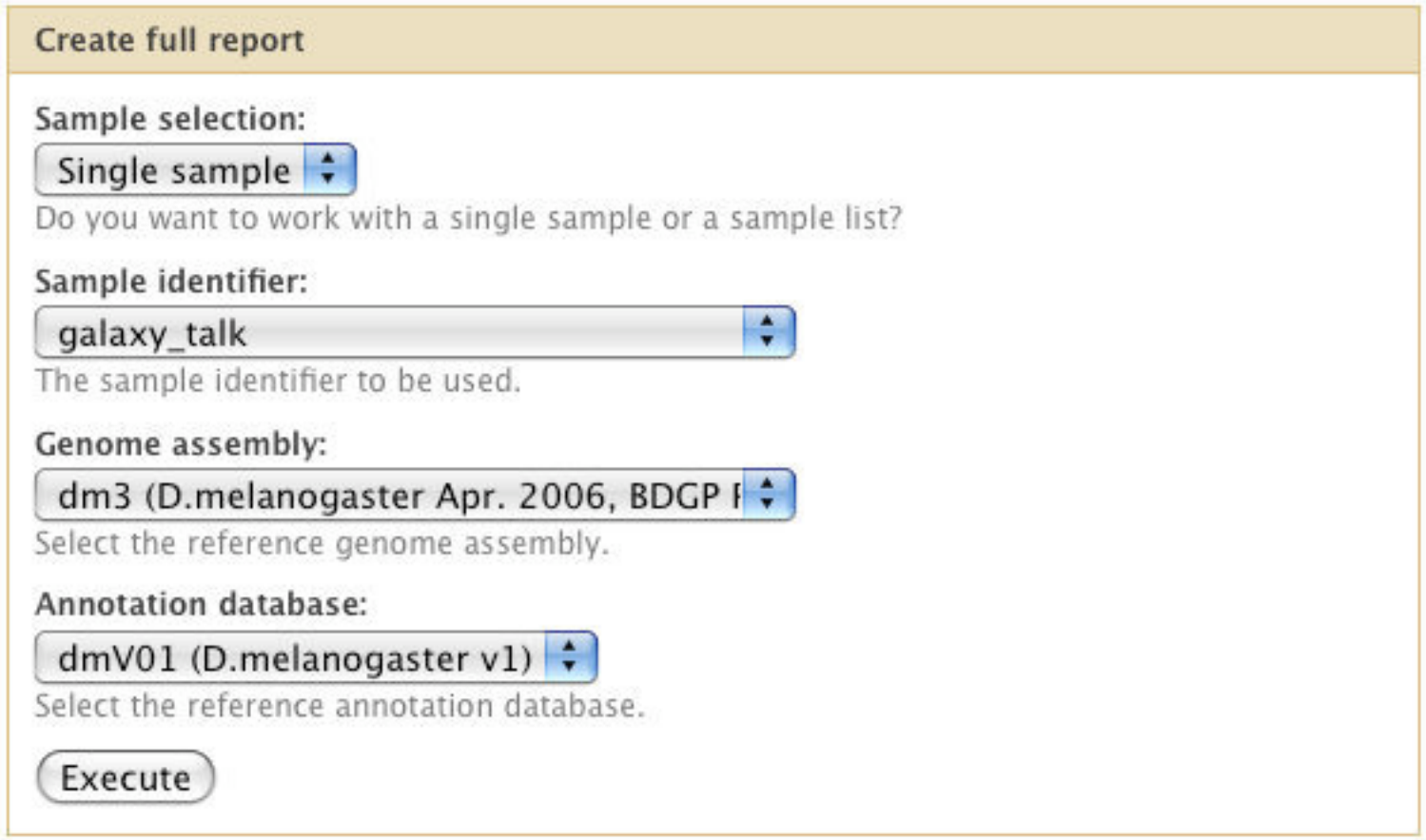

FMI
Friedrich Miescher Institute
for Biomedical Research

# Next Gen Sequencing

## data (result) is inside “Galaxy”



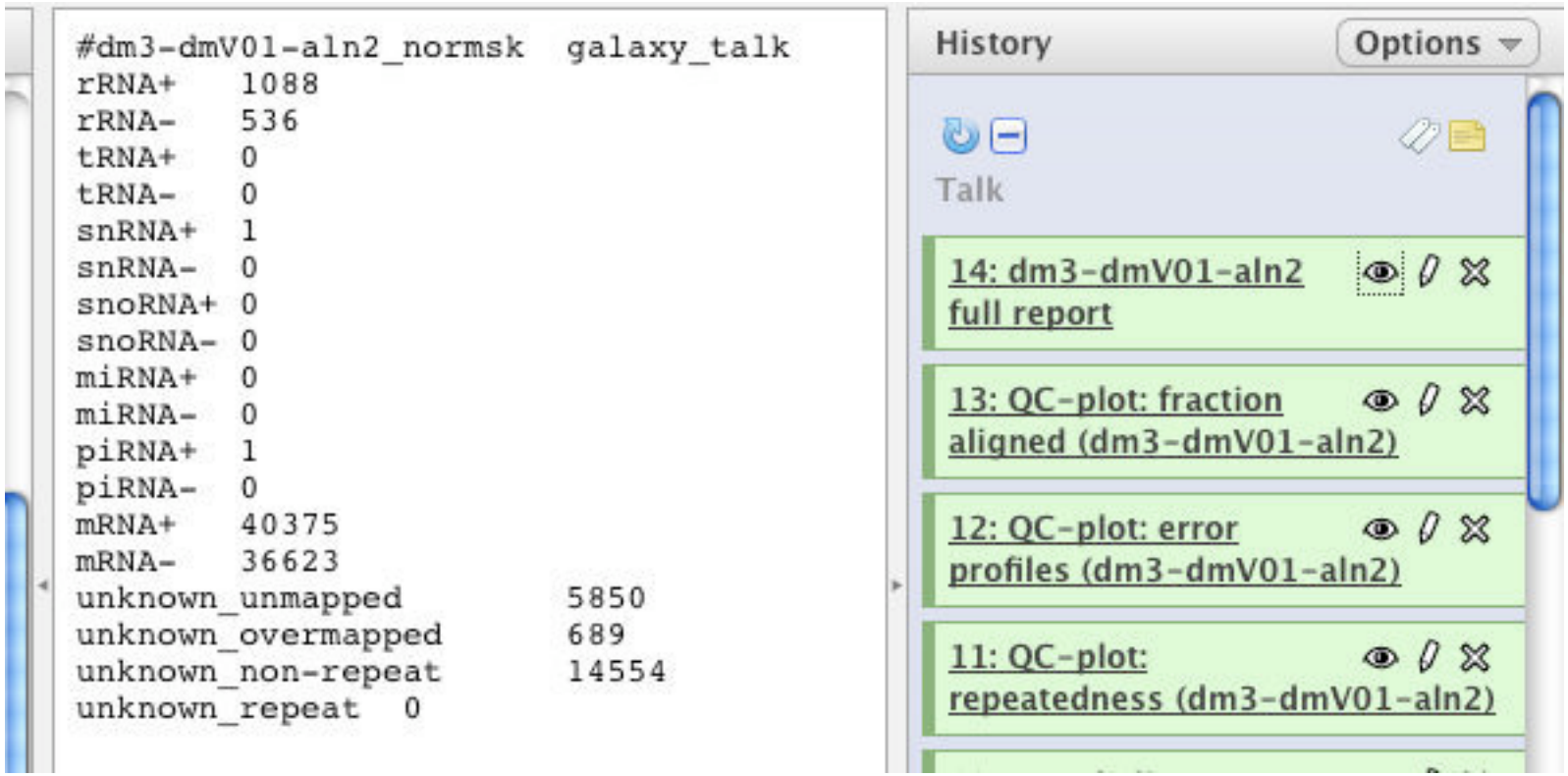
# Next Gen Sequencing




FMI
Friedrich Miescher Institute
for Biomedical Research

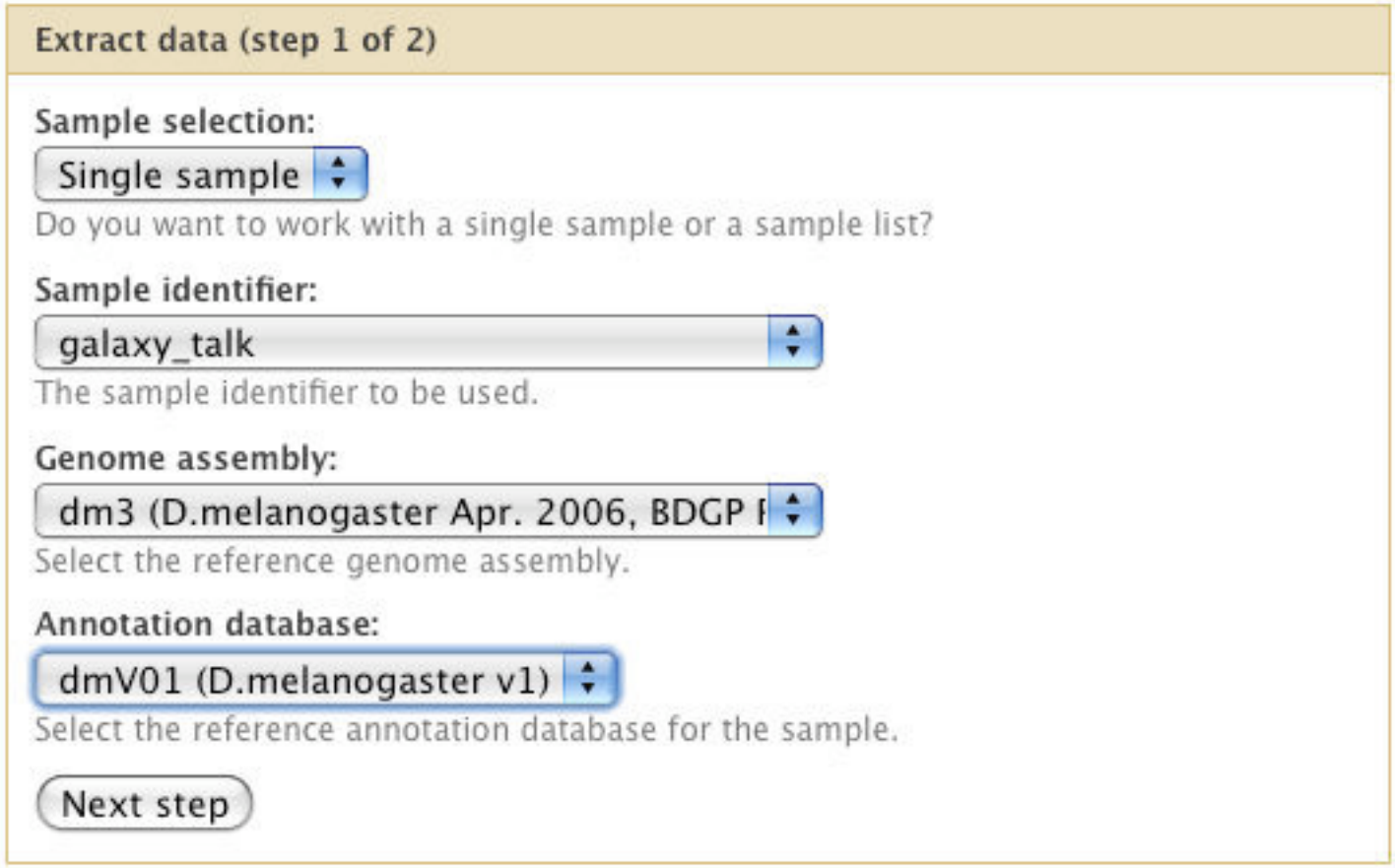
# Next Gen Sequencing

## data (result) is inside “Galaxy”

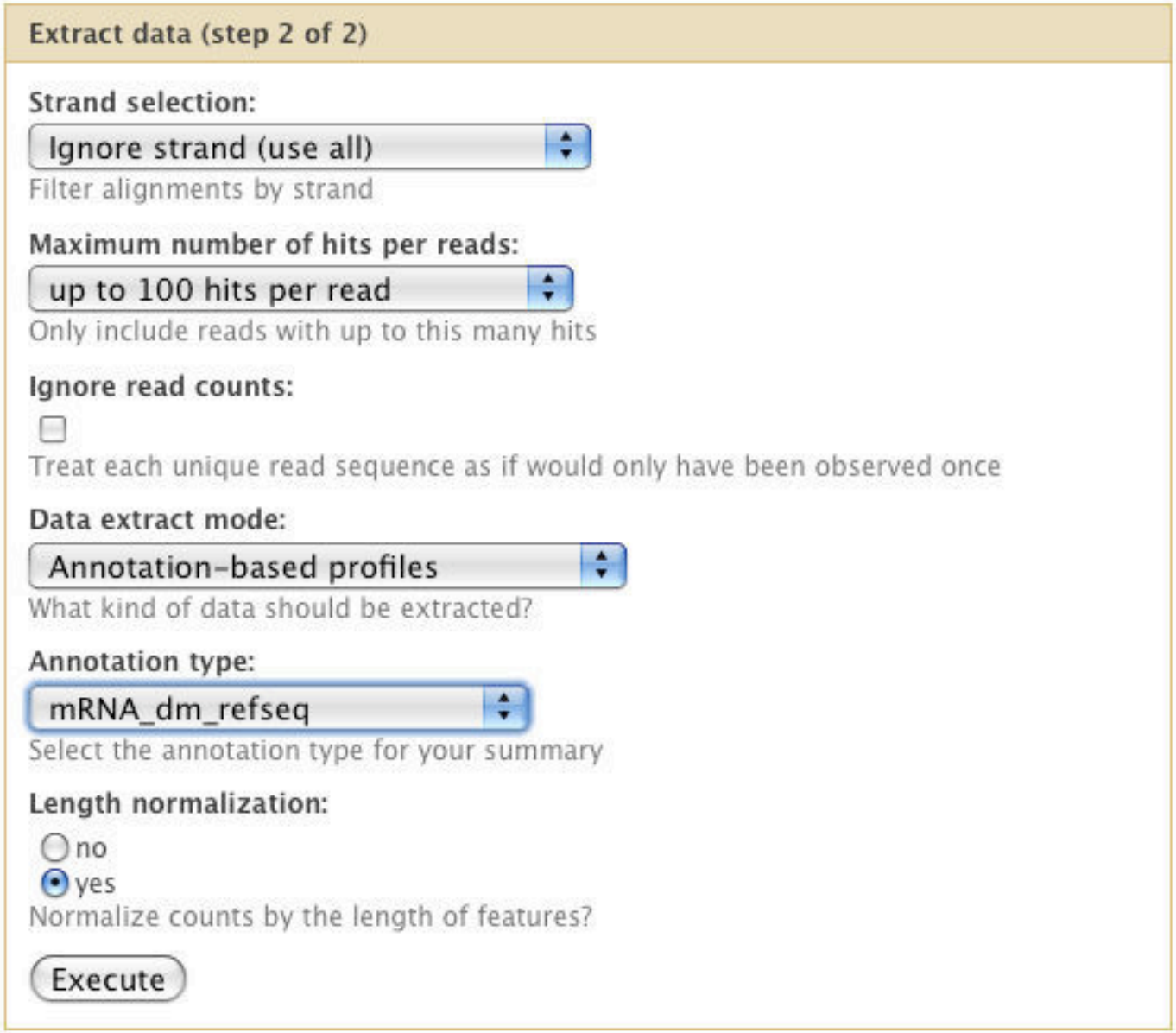


```
#dm3-dmV01-aln2_normsk  galaxy_talk
rRNA+    1088
rRNA-    536
tRNA+    0
tRNA-    0
snRNA+   1
snRNA-   0
snoRNA+  0
snoRNA-  0
miRNA+   0
miRNA-   0
piRNA+   1
piRNA-   0
mRNA+    40375
mRNA-    36623
unknown_unmapped          5850
unknown_overmapped        689
unknown_non-repeat        14554
unknown_repeat   0
```

# Next Gen Sequencing



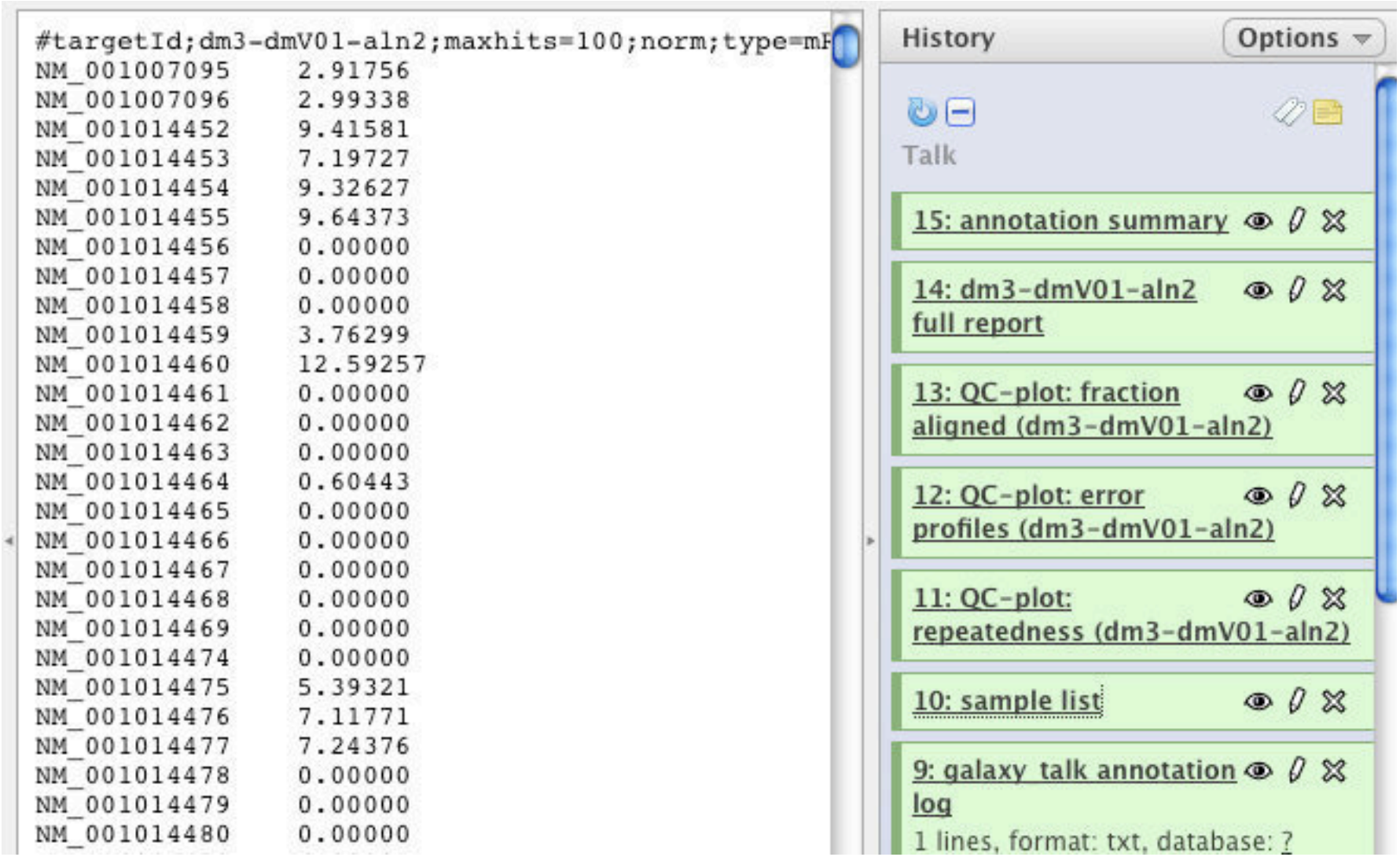
# Next Gen Sequencing




FMI
Friedrich Miescher Institute
for Biomedical Research

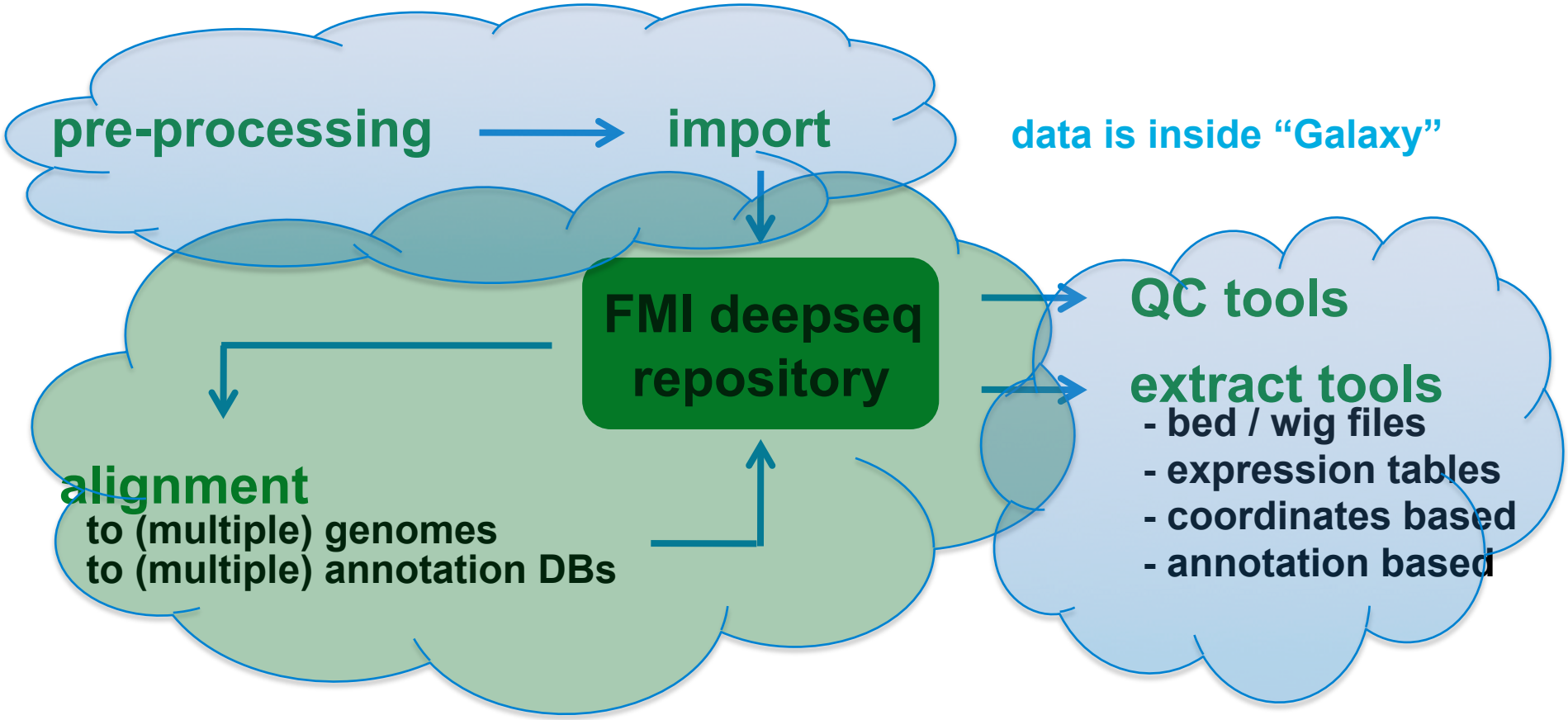
# Next Gen Sequencing

## data (result) is inside “Galaxy”



# Next Gen Sequencing

*....just a bunch of Perl scripts!*




FMI
Friedrich Miescher Institute for Biomedical Research

# Acknowledgment


## *Computational Biology*

- Michael Stadler
- Dimos Gaidatzis

## *Functional Genomics*

- Tim Roloff

## *IT Support*

- Thomas Übermeier
- Jan Welker
- Dean Flanders

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

FMI
Friedrich Miescher Institute
for Biomedical Research