FMI
Friedrich Miescher Institute
for Biomedical Research

11/15/11

Using Galaxy
to provide a NGS Analysis Platform

Hans-Rudolf Hotz (hrh@fmi.ch)

Friedrich Miescher Institute for Biomedical Research
Basel, Switzerland

Friedrich Miescher Institute

- part of the Novartis Research Foundation
- affiliated institute of Basel University
- member of Swiss Institute of Bioinformatics

316 employees
(incl. 96 PhD students, 95 Post Docs)

Epigenetics	Growth Control	Neurobiology
(8 research groups)	(7 research groups)	(8 research groups)

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

SIB Swiss Institute of Bioinformatics
UNI BASEL
NOVARTIS

FMI
Friedrich Miescher Institute
for Biomedical Research

working with NGS data is *fascinating* because

- there is a different instrument on the market every few months

- Scientists come up with new kind experiments

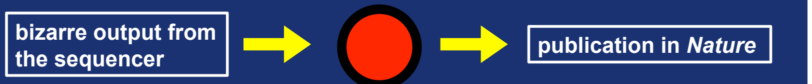
- new algorithms to deal with NGS data are developed continuously

FMI
Friedrich Miescher Institute
for Biomedical Research

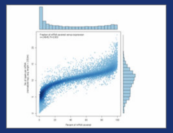
working with NGS data is *difficult* because

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***

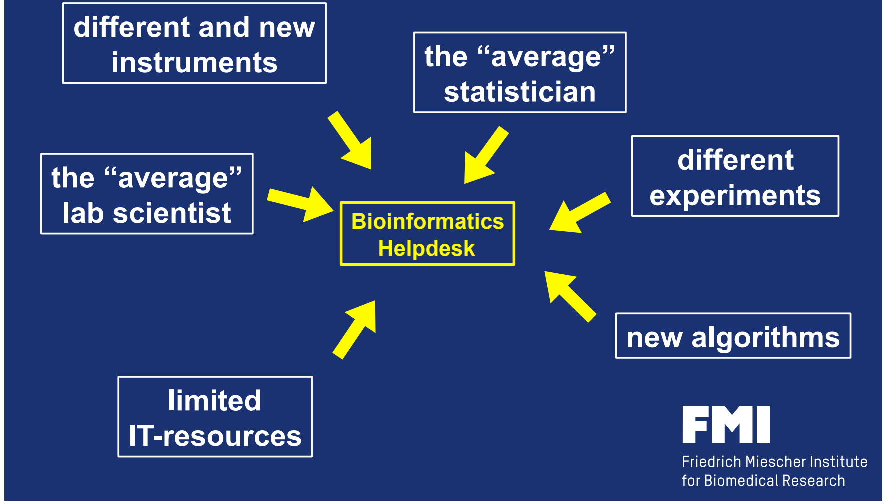
the "average" statistician is creating wonderful blots.....



→ **...nobody understands**

FMI
Friedrich Miescher Institute
for Biomedical Research

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



different and new instruments

the “average” statistician

the “average” lab scientist

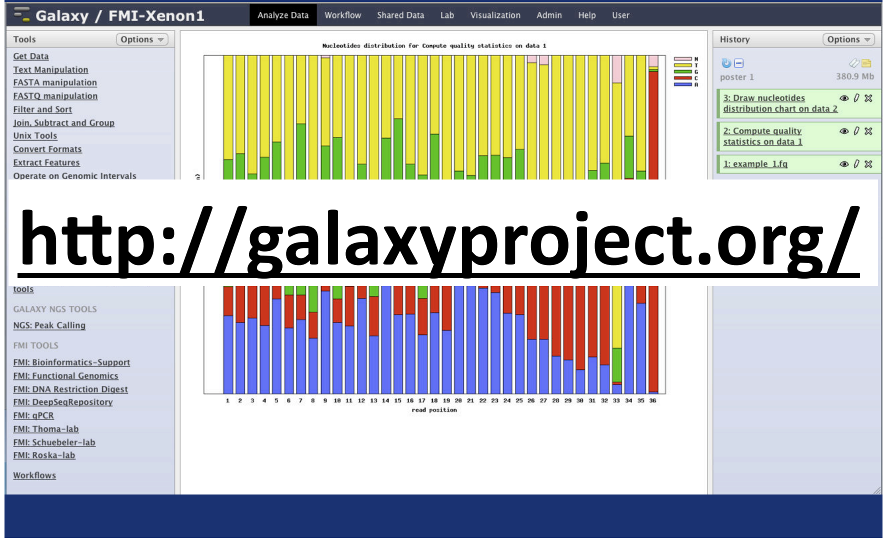
different experiments

Bioinformatics Helpdesk

new algorithms

limited IT-resources

the solution:



http://galaxyproject.org/

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.

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.

what is Galaxy?

Galaxy

- provides a GUI to Bioinformatics tools
- manages/stores your (raw) data and results
- allows you to create workflows
- allows sharing and reproducing your analysis

public Galaxy instance:

http://usegalaxy.org

FMI
Friedrich Miescher Institute
for Biomedical Research

why are we using Galaxy

Galaxy

- open source (http://wiki.g2.bx.psu.edu/Admin/License)
- we can modify the tools
- we can add our own tools
- it is part of a wider community: “GenomeSpace”, “GMOD”
- it is flexible and simple to install

FMI
Friedrich Miescher Institute
for Biomedical Research

it is really simple to install

Galaxy

requirements:

- Python (2.5 or 2.6)
- Mercurial

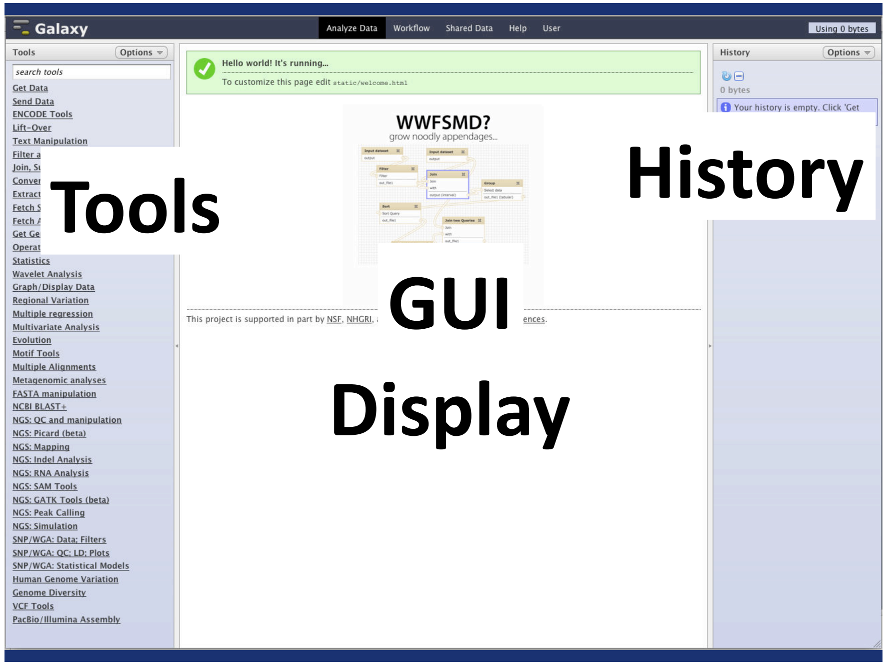
just 3 commands:

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

...and it is ready (on linux and Mac) at:

http://localhost:8080

FMI
Friedrich Miescher Institute
for Biomedical Research



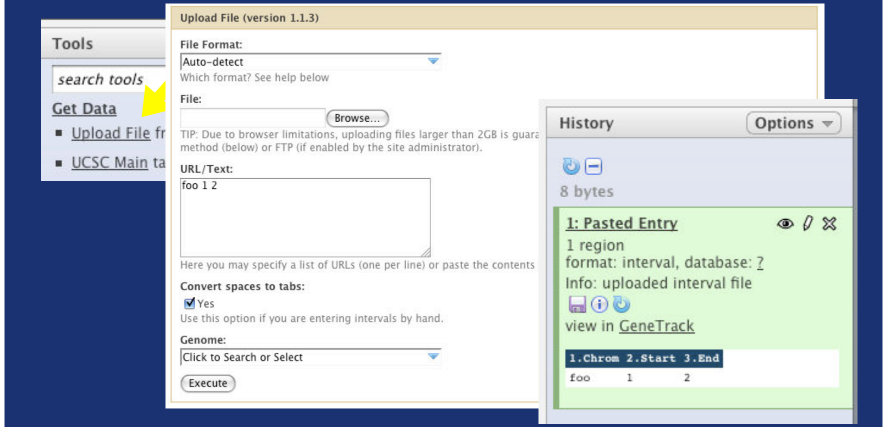
Hello world! It's running...

To customize this page edit `static/welcome.html`

This project is supported in part by NSF, NHGRI, ... ences.

Your history is empty. Click 'Get

how does it work



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

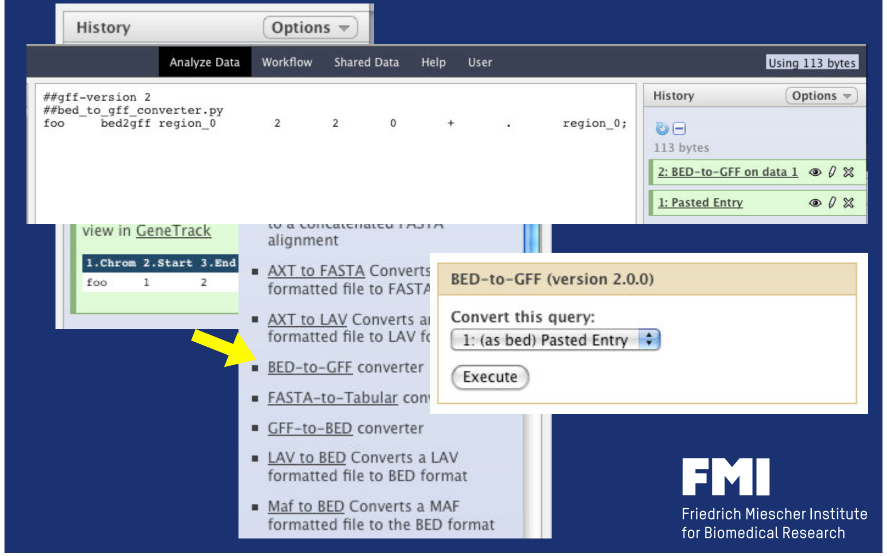
1: Pasted Entry

1 region
format: interval, database: ?
Info: uploaded interval file

view in GeneTrack

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

how does it work



History Options

```
##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 conv
- 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

what kind of tools do you get

input tools:

- text box / upload file / url
- access to a local file system   (“Data Libraries”)
- access to UCSC table browser and ensembl biomart

text manipulation tools:

- file conversion
- table calculation
- operation on genomic intervals

wrappers (and GUIs):

EMBOSS, NCBI BLAST+, NGS: QC and manipulation, Picard, NGS: Mapping, NGS: Indel Analysis, NGS: RNA Analysis, SAM Tools, GATK Tools, NGS: Peak Calling, and much more...

what needs to be added

EMBOSS, NCBI BLAST+, bowtie, BWA, macs, cufflinks, samtools, etc

plus the corresponding databases, indices, etc

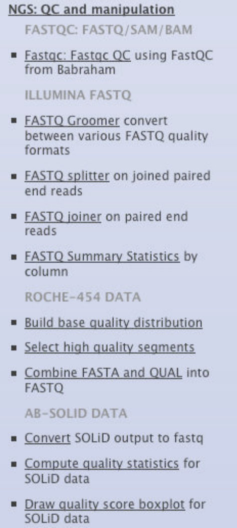
yes, this is annoying.....

but, provides a lot of flexibility.....

and if you don’t want to use eg BLAST, remove it from the list and don’t care about the binary.

do you really need all the tools

Galaxy



NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

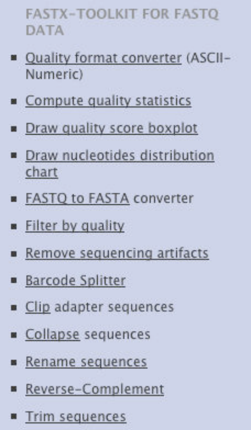
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

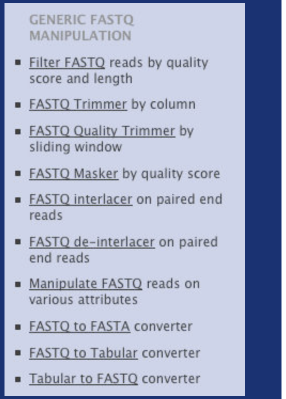
AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data



FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement
- Trim sequences

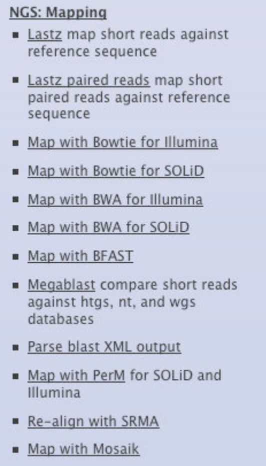


GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- FASTQ interlacer on paired end reads
- FASTQ de-interlacer on paired end reads
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

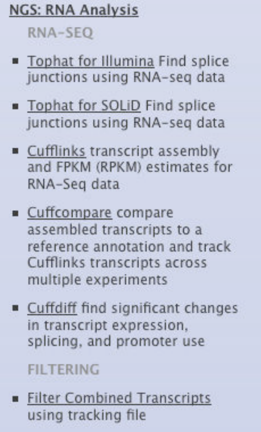
do you really need all the tools

Galaxy



NGS: Mapping

- Lastz map short reads against reference sequence
- Lastz paired reads map short paired reads against reference sequence
- Map with Bowtie for Illumina
- Map with Bowtie for SOLiD
- Map with BWA for Illumina
- Map with BWA for SOLiD
- Map with BFAST
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output
- Map with PerM for SOLiD and Illumina
- Re-align with SRMA
- Map with Mosaik



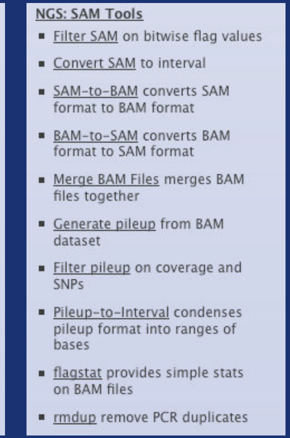
NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat for SOLiD Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file



NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files
- rmdup remove PCR duplicates

from the laptop to a production environment

- remove the tools you don't want

- switch from SQLite to PostgreSQL or MySQL

- use a proxy server

- authenticate users externally via Kerberos or LDAP

- use a 'big' server

- use a compute cluster (TORQUE PBS, PBS Pro, Platform LSF, and Sun Grid Engine)

or use the "cloud"

Galaxy

http://usegalaxy.org/cloud
(using "Amazon Elastic Compute Cloud")

Enis Afgan, et al.
Harnessing cloud computing with Galaxy Cloud
Nature Biotechnology 29, 972–974,
Published online 08 November 2011
www.nature.com/nbt/journal/v29/n11/full/nbt.2028.html

adding your own tools

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

FMI
Friedrich Miescher Institute
for Biomedical Research

tool definition file

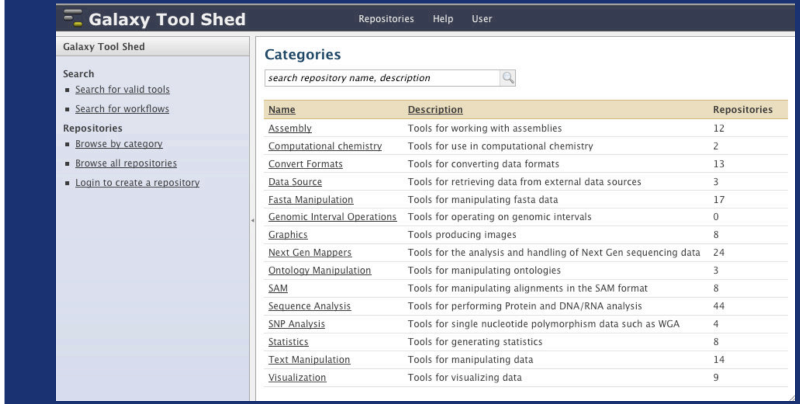
```
<tool id="bed2gff1" name="BED-to-GFF" version="2.0.0">
 <description>converter</description>

 <command>bed_to_gff_converter.py $input $out_file1</command>

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

 <outputs>
  <data format="gff" name="out_file1" />
 </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

Galaxy Tool Shed

enables sharing of tools across the Galaxy community.



Galaxy Tool Shed — Repositories · Help · User

Galaxy Tool Shed

Search
- Search for valid tools
- Search for workflows

Repositories
- Browse by category
- Browse all repositories
- Login to create a repository

Categories

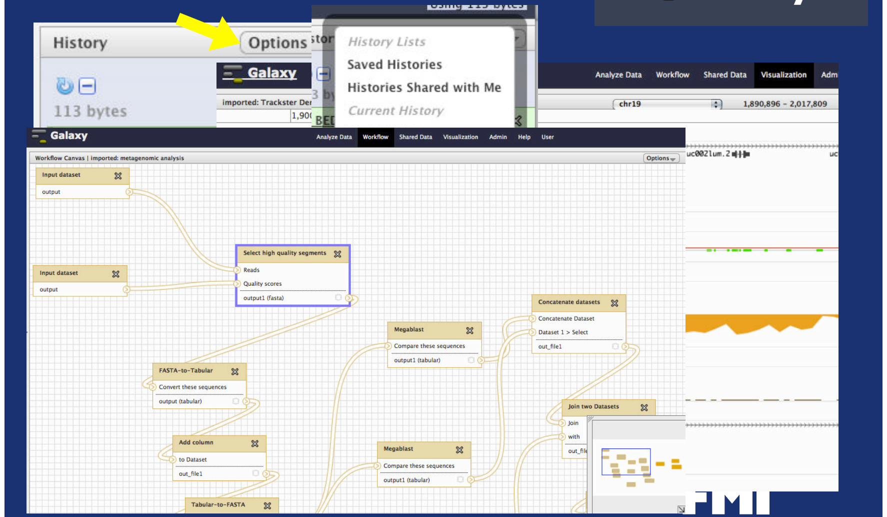
search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	12
Computational chemistry	Tools for use in computational chemistry	2
Convert Formats	Tools for converting data formats	13
Data Source	Tools for retrieving data from external data sources	3
Fasta Manipulation	Tools for manipulating fasta data	17
Genomic Interval Operations	Tools for operating on genomic intervals	0
Graphics	Tools producing images	8
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	24
Ontology Manipulation	Tools for manipulating ontologies	3
SAM	Tools for manipulating alignments in the SAM format	8
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	44
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	4
Statistics	Tools for generating statistics	8
Text Manipulation	Tools for manipulating data	14
Visualization	Tools for visualizing data	9

http://toolshed.g2.bx.psu.edu/

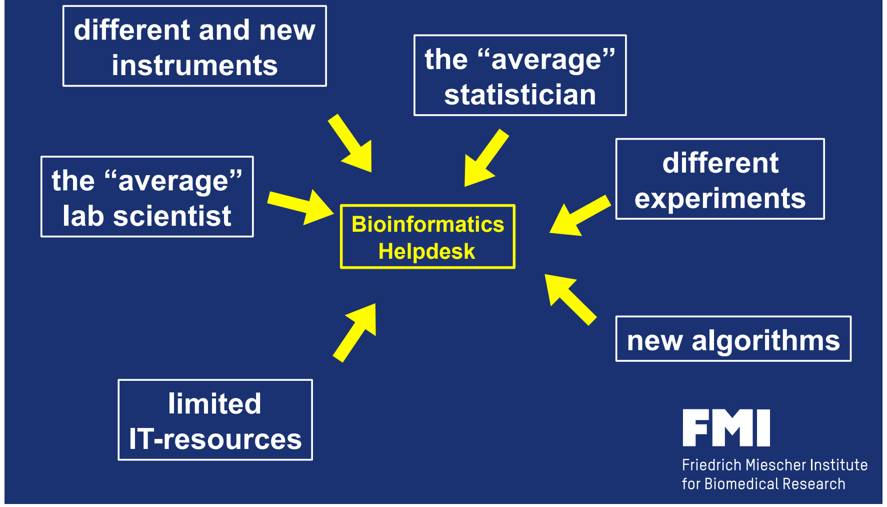
FMI
Friedrich Miescher Institute
for Biomedical Research

a few more highlights



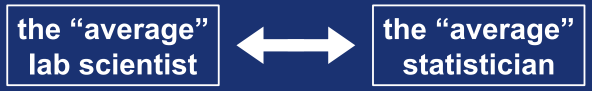
Friedrich Miescher Institute
for Biomedical Research

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


different and new instruments
the “average” statistician
the “average” lab scientist
Bioinformatics Helpdesk
different experiments
new algorithms
limited IT-resources

He is a Galaxy administrator

Galaxy



...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

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

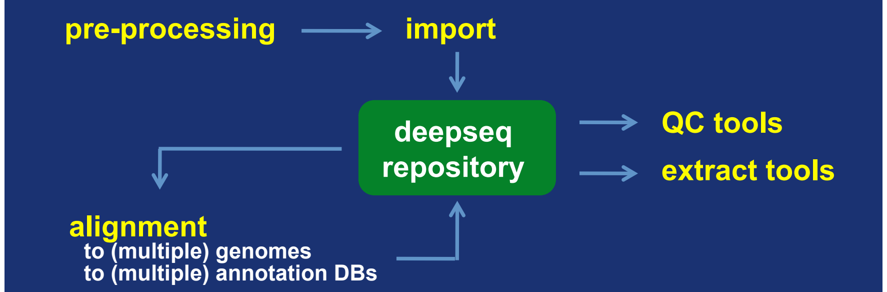
He is a Galaxy administrator

Galaxy

limited IT-resources

- shared infrastructure
- Galaxy provides ‘report’ tools

NGS analysis at the FMI



The NGS pipeline at the FMI is Galaxy

....just a bunch of Perl scripts *(currently)*

→ which can be easily added to Galaxy

....just a simple file system *(currently)*

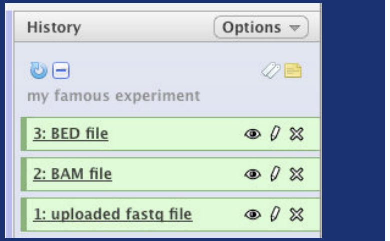
....which cannot be added to Galaxy.
(Galaxy uses its own data directory)

→ we don't have to, we just have to give Galaxy access to the directory
(without using "Data Libraries")

a simple NGS workflow Galaxy

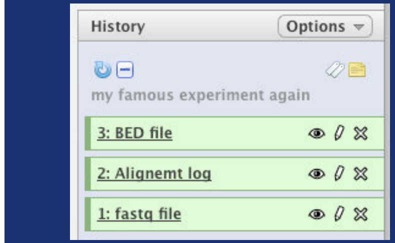
- your famous aligner

- your famous extract tool



a simple NGS workflow Galaxy

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

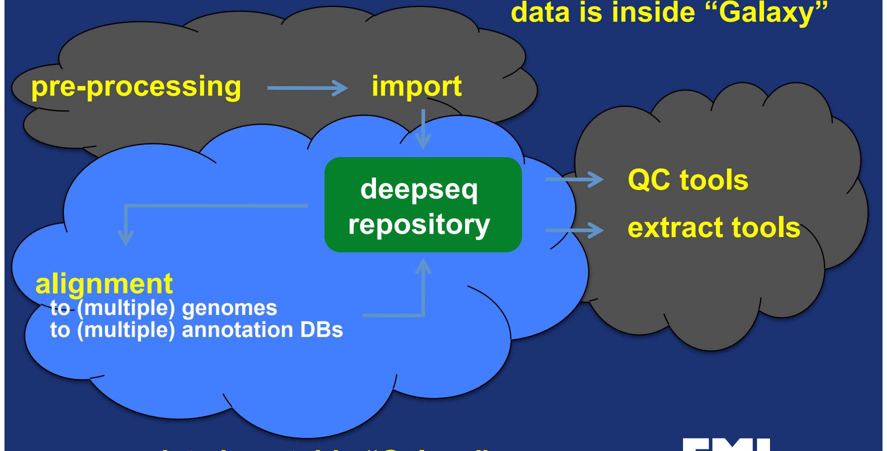


- the 'famous aligner' has a wrapper storing the BAM file in the central NGS repository and creating just a log file for Galaxy
- your 'famous extract tool' knows the location of the NGS repository

NGS analysis at the FMI

Galaxy

data is inside “Galaxy”



pre-processing → import

deepseq repository

QC tools

extract tools

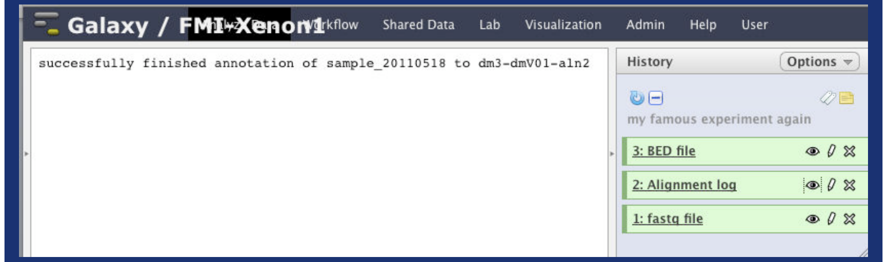
alignment
to (multiple) genomes
to (multiple) annotation DBs

data is outside “Galaxy”

storing data outside of Galaxy

Galaxy

makes it easier to share with non-Galaxy users



```
successfully finished annotation of sample_20110518 to dm3-dmV01-aln2
```

my famous experiment again

3: BED file

2: Alignment log

1: fastq file

makes it easier to share with non-Galaxy users

```
successfully finished annotation of
sampleId_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 ~]$
```

makes it easier to share with non-Galaxy users

command line

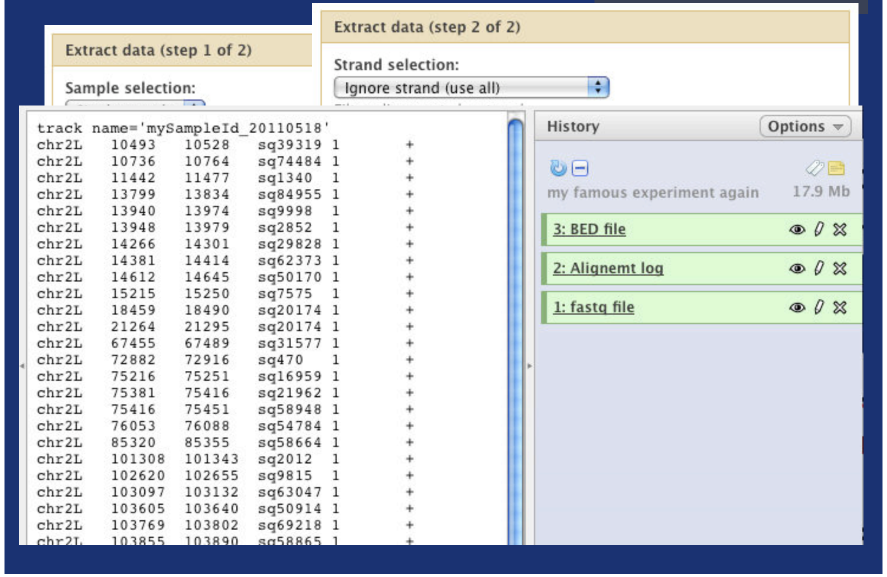
```
extractData.pl -f -s p -m 100 -i
mySampleId_20110518 dm3-dmV01-aln2 genome |
frag2bed.pl -t -q -U -
```

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
```

and doing the same in Galaxy

Galaxy



Extract data (step 1 of 2)

Sample selection:

Extract data (step 2 of 2)

Strand selection: Ignore strand (use all)

```
track name='mySampleId_20110518'
chr2L   10493   10528   sq39319 1       +
chr2L   10736   10764   sq74484 1       +
chr2L   11442   11477   sq1340  1       +
chr2L   13799   13834   sq84955 1       +
chr2L   13940   13974   sq9998  1       +
chr2L   13948   13979   sq2852  1       +
chr2L   14266   14301   sq29828 1       +
chr2L   14381   14414   sq62373 1       +
chr2L   14612   14645   sq50170 1       +
chr2L   15215   15250   sq7575  1       +
chr2L   18459   18490   sq20174 1       +
chr2L   21264   21295   sq20174 1       +
chr2L   67455   67489   sq31577 1       +
chr2L   72882   72916   sq470   1       +
chr2L   75216   75251   sq16959 1       +
chr2L   75381   75416   sq21962 1       +
chr2L   75416   75451   sq58948 1       +
chr2L   76053   76088   sq54784 1       +
chr2L   85320   85355   sq58664 1       +
chr2L   101308  101343  sq2012  1       +
chr2L   102620  102655  sq9815  1       +
chr2L   103097  103132  sq63047 1       +
chr2L   103605  103640  sq50914 1       +
chr2L   103769  103802  sq69218 1       +
chr2L   103855  103890  sq58865 1       +
```

my famous experiment again 17.9 Mb

3: BED file

2: Alignemt log

1: fastq file

Summary

Mission running a Bioinformatics Helpdesk

Vision I don't have to do anything

Strategy Galaxy

FMI
Friedrich Miescher Institute
for Biomedical Research

Acknowledgment

Galaxy

Michael Stadler Lukas Burger
Dimos Gaidatzis
Tim Roloff Adam Feigin
Anita Lerch
Thomas Übermeier Jan Welker

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

SIB Swiss Institute of Bioinformatics UNI BASEL NOVARTIS FMI Friedrich Miescher Institute for Biomedical Research

a few web sites

Galaxy

http://galaxyproject.org

http://usegalaxy.org
http://usegalaxy.org/galaxy101
http://usegalaxy.org/cloud
http://toolshed.g2.bx.psu.edu/
http://wiki.g2.bx.psu.edu/Admin/License

SIB Swiss Institute of Bioinformatics UNI BASEL NOVARTIS FMI Friedrich Miescher Institute for Biomedical Research