2012/07/16

Running a Bioinformatics Help Desk

*from drawing colorful plasmid maps
to working with HiSeq data*

Solved and Unsolved Problems

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

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

Epigenetics (7 research groups)

Growth Control (8 research groups)

Neurobiology (8 research groups)

Technology Platforms

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

UNI BASEL

NOVARTIS

FMI
Friedrich Miescher Institute
for Biomedical Research

Computational Biology / Bioinformatics

- member of Swiss Institute of Bioinformatics

- 3 core funded and 2 third party funded FTE
- many interactions with Functional Genomics
- hardware is maintained by IT
- providing support for ~250 scientists
- all services are free
    - "collaborations" → papers
    - "helpdesk"

SIB Swiss Institute of Bioinformatics

UNI BASEL

NOVARTIS

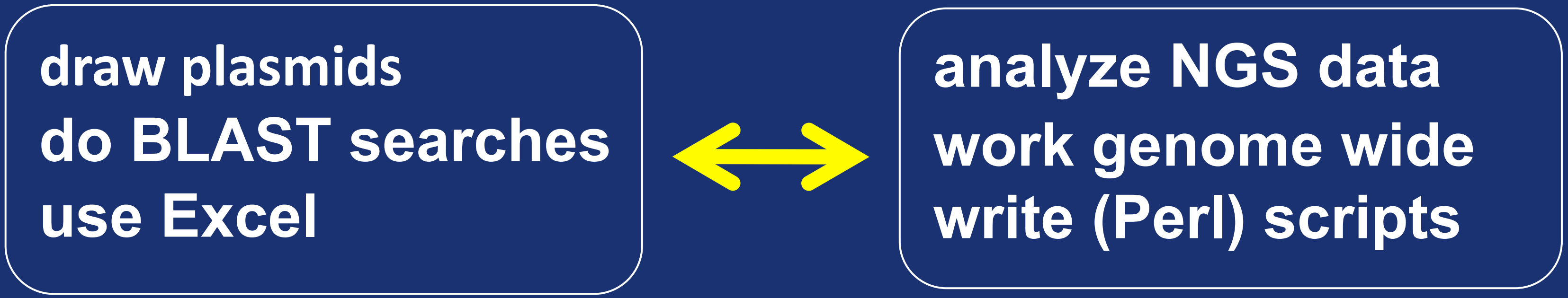
FMI
Friedrich Miescher Institute
for Biomedical Research

Bioinformatics Helpdesk

providing support for:

the “average” lab
scientist, who wants to:

the “modern” lab
scientist, who wants to:



draw plasmids
do BLAST searches
use Excel

analyze NGS data
work genome wide
write (Perl) scripts

....and how to bridge the gap?

FMI
Friedrich Miescher Institute
for Biomedical Research

the “average” lab scientist, struggles with

drawing plasmids

doing BLAST searches

using Excel

drawing plasmids:

the actual problem:
there is no good ‘desktop bioinformatics package’

our situation:

package A:
- 10 perpetual licenses bought in 2006
- windows only
- stuck on version X (does not run on windows 7)

package B:
- 20 perpetual licenses bought in 2008
- 3 year support and upgrades
- stuck on version Y
- windows/mac/linux

both packages are ridiculously expensive

Friedrich Miescher Institute
for Biomedical Research

drawing plasmids: open source/free alternatives

we have been looking at:

GENtle	http://gentle.magnusmanske.de/
Serial Cloner	http://serialbasics.free.fr/Serial_Cloner.html
pDRAW32	http://www.acaclone.com/
BioEdit	http://www.mbio.ncsu.edu/BioEdit/BioEdit.html
GeneCoder	http://www.algosome.com
Workbench	http://www.ncbi.nlm.nih.gov/tools/gbench/
Ape	http://biologylabs.utah.edu/jorgensen/wayned/ape/
UGene	http://ugene.unipro.ru/

Has anybody experience with these or other open source/free packages?

FMI

Friedrich Miescher Institute
for Biomedical Research

drawing plasmids:

what about EMBOSS ?
(we offer most EMBOSS tools in our Galaxy server)

The tools ‘cirdna’ and ‘lindna’ produce reasonable maps of DNA constructs......but the data needs to be in 'cirp' and 'linp' format, respectively.

How do I transform a genbank file to 'cirp’ format?

we have no satisfying solution

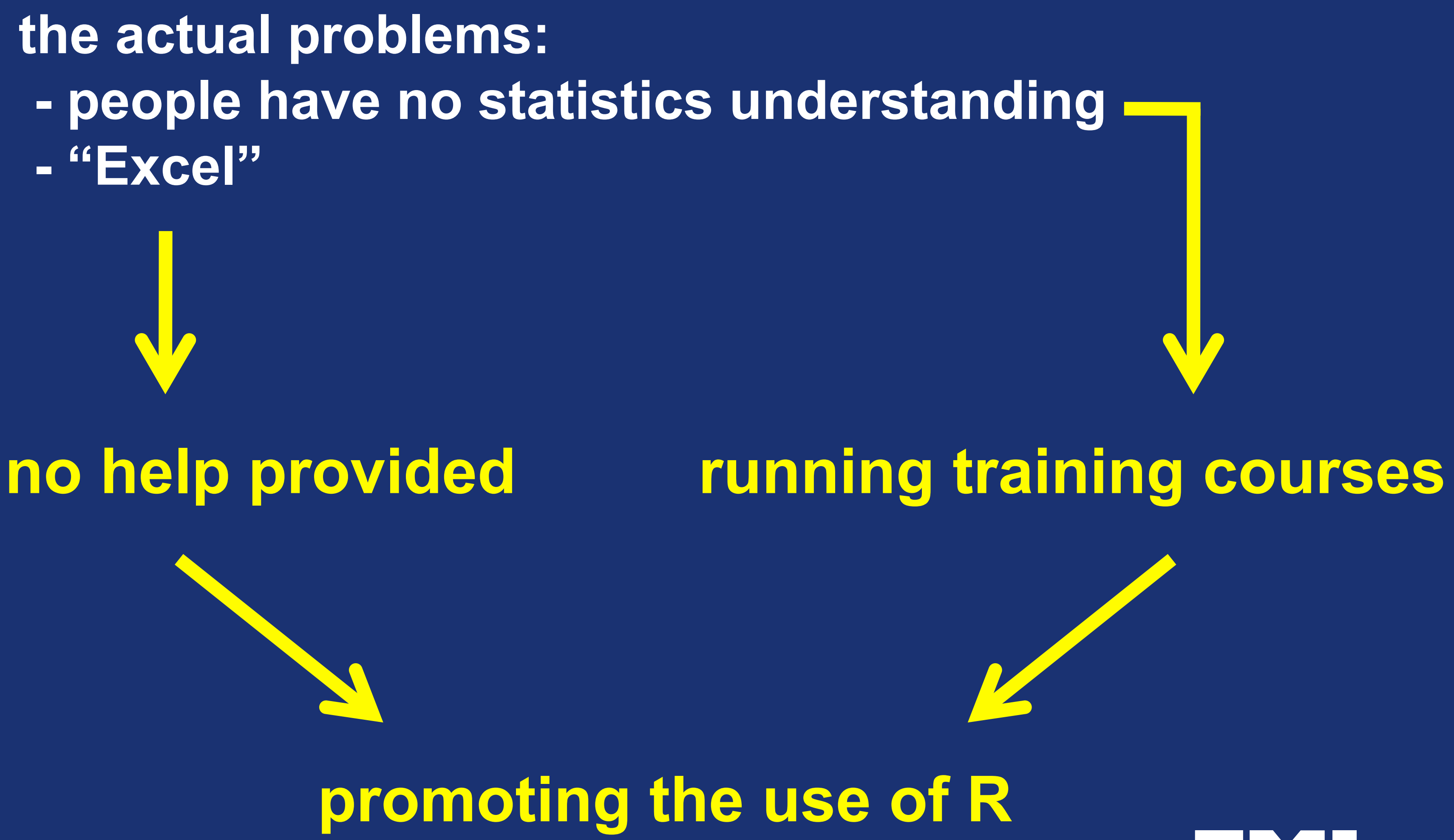
doing BLAST searches

the actual problem:
people are struggling using web resources

running training courses

internal wiki pages / FAQ

using Excel



the actual problems:

- people have no statistics understanding
- “Excel”

no help provided

running training courses

promoting the use of R

Friedrich Miescher Institute
for Biomedical Research

the “modern” lab scientist, struggles with

writing (Perl) scripts

analyzing NGS data

work genome wide

simple solution:
running introductory and advanced training courses in R

FMI

Friedrich Miescher Institute
for Biomedical Research

running R training courses

- R is a decent scripting language
- we can teach them statistics on the side
- they can start using Bioconductor

we currently re-implement our (perl based) NGS pipeline in a new Bioconductor package: “QuasR”

...but one problem remains: people want to display their data in a genome browsers

FMI

Friedrich Miescher Institute
for Biomedical Research

genome browsers

we use a combination of:

R/Bioconductor: *GenomeGraphs*, new package *Gviz*

web resources: *ensembl* - too slow
UCSC - *S. pombe* is missing
(we don't have the resources to run a local mirror)

local on desktop: *IGV and IGB*

Galaxy "Trackster"

Has anybody a perfect solution?

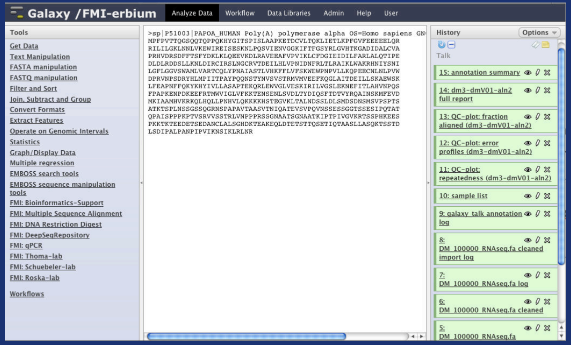
FMI
Friedrich Miescher Institute
for Biomedical Research

Bridging the Gap

the “average”
lab scientist

?

the “modern”
lab scientist


Galaxy /FMI-erbium
Analyze Data
Workflow
Data Libraries
Admin
Help
User
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
Workflows
History
Options
Talk
15: annotation summary
14: dm3-dmV01-aln2 full report
13: QC-plot: fraction aligned (dm3-dmV01-aln2)
12: QC-plot: error profiles (dm3-dmV01-aln2)
11: QC-plot: repeatedness (dm3-dmV01-aln2)
10: sample list
9: galaxy_talk annotation log
8: DM_100000_RNAseq.fa cleaned import log
7: DM_100000_RNAseq.fa log
6: DM_100000_RNAseq.fa cleaned
5: DM_100000_RNAseq.fa

```
>sp|P51003|PAPOA_HUMAN Poly(A) polymerase alpha OS=Homo sapiens GN
MPFPVTTQGSQQTQPPQKHYGITSPISLAAPKETDCVLTQKLIETLKPFGVFEEEEELQR
RILILGKLNNLVKEWIREISESKNLPQSVIENVGGKIFTFGSYRLGVHTKGADIDALCVA
PRHVDRSDFFTSFYDKLKLQEEVKDLRAVEEAFVPVIKLCFDGIEIDILFARLALQTIPE
DLDLRDDSLLKNLDIRCIRSLNGCRVTDEILHLVPNIDNFRLTLRAIKLWAKRHNIYSNI
LGFLGGVSWAMLVARTCQLYPNAIASTLVHKFFLVFSKWEWPNPVLLKQPEECNLNLPVW
DPRVNPSDRYHLMPIITPAYPQQNSTYNVSVSTRMVMVEEFKQGLAITDEILLSKAEWSK
LFEAPNFFQKYKHYIVLLASAPTEKQRLEWVGLVESKIRILVGSLEKNEFITLAHVNPQS
FPAPKENPDKEEFRTMWVIGLVFKKTENSENLSVDLTYDIQSFTDTVYRQAINSKMFEVD
MKIAAMHVKRKQLHQLLPNHVLQKKKKHSTEGVKLTALNDSSLDLSMDSDNSMSVPSPTS
ATKTSPLNSSGSSQGRNSPAPAVTAASVTNIQATEVSVPQVNSSESSGGTSSESIPQTAT
QPAISPPPKPTVSRVVSSTRLVNPPPRSSGNAATSGNAATKIPTPIVGVKRTSSPHKEES
PKKTKTEEDETSEDANCLALSGHDKTEAKEQLDTETSTTQSETIQTAASLLASQKTSSTD
LSDIPALPANPIPVIKNSIKLRLNR
```

http://galaxyproject.org/

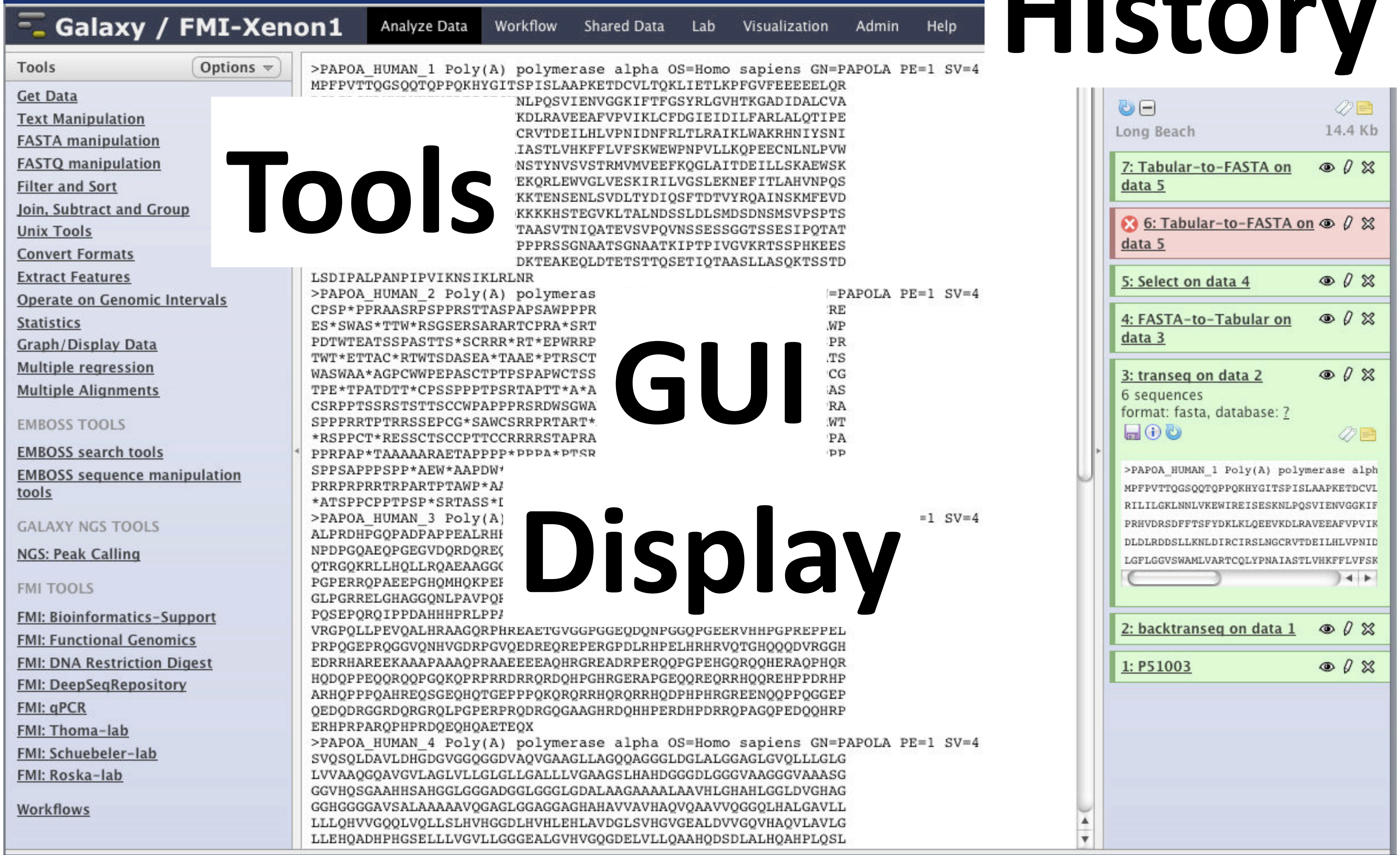
FMI
Friedrich Miescher Institute
for Biomedical Research

History

Tools

GUI

Display



```
>PAPOA_HUMAN_1 Poly(A) polymerase alpha OS=Homo sapiens GN=PAPOLA PE=1 SV=4
MPFPVTTQGSQQTQPPQKHYGITSPISLAAPKETDCVLTQKLIETLKPFGVFEEEEELQR
                              NLPQSVIENVGGKIFTFGSYRLGVHTKGADIDALCVA
                              KDLRAVEEAFVPVIKLCFDGIEIDILFARLALQTIPE
                              CRVTDEILHLVPNIDNFRLTLRAIKLWAKRHNIYSNI
                              IASTLVHKFFLVFSKWEWPNPVLLKQPEECNLNLPVW
                              NSTYNVSVSTRMVMVEEFKQGLAITDEILLSKAEWSK
                              EKQRLEWVGLVESKIRILVGSLEKNEFITLAHVNPQS
                              KKTENSENLSVDLTYDIQSFTDTVYRQAINSKMFEVD
                              KKKKHSTEGVKLTALNDSSLDLSMDSDNSMSVPSPTS
                              TAASVTNIQATEVSVPQVNSSESSGGTSSESIPQTAT
                              PPPRSSGNAATSGNAATKIPTPIVGVKRTSSPHKEES
                              DKTEAKEQLDTETSTTQSETIQTAASLLASQKTSSTD
LSDIPALPANPIPVIKNSIKLRLNR
>PAPOA_HUMAN_2 Poly(A) polymeras                              =PAPOLA PE=1 SV=4
CPSP*PPRAASRPSPPRSTTASPAPSAWPPPR                              RE
ES*SWAS*TTW*RSGSERSARARTCPRA*SRT                              WP
PDTWTEATSSPASTTS*SCRRR*RT*EPWRRP                              PR
TWT*ETTAC*RTWTSDASEA*TAAE*PTRSCT                              TS
WASWAA*AGPCWWPEPASCTPTPSPAPWCTSS                              CG
TPE*TPATDTT*CPSSPPPTPSRTAPTT*A*A                              AS
CSRPPTSSRSTSTTSCCWPAPPPRSRDWSGWA                              RA
SPPPRRTPTRRSSEPCG*SAWCSRRPRTART*                              WT
*RSPPCT*RESSCTSCCPTTCCRRRRSTAPRA                              PA
PPRPAP*TAAAAARAETAPPPP*PPPA*PTSR                              PP
SPPSAPPPSPP*AEW*AAPDW
PRRPRPRRTRPARTPTAWP*A
*ATSPPCPPTPSP*SRTASS*
>PAPOA_HUMAN_3 Poly(A)                                            =1 SV=4
ALPRDHPGQPADPAPPEALRH
NPDPGQAEQPGEGVDQRDQRE
QTRGQKRLLHQLLRQAEAAGG
PGPERRQPAEEPGHQMHQKPE
GLPGRRELGHAGGQNLPAVPQ
PQSEPQRQIPPDAHHHPRLPP
VRGPQLLPEVQALHRAAGQRPHREAETGVGGPGGEQDQNPGGQPGEERVHHPGPREPPEL
PRPQGEPRQGGVQNHVGDRPGVQEDREQREPERGPDLRHPELHRHRVQTGHQQQDVRGGH
EDRRHAREEKAAAPAAAQPRAAEEEEAQHRGREADRPERQQPGPEHGQRQQHERAQPHQR
HQDQPPEQQRQQPGQKQPRPRRDRRQRDQHPGHRGERAPGEQQREQRRHQQREHPPDRHP
ARHQPPPQAHREQSGEQHQTGEPPPQKQRQRRHQRQRRHQDPHPHRGREENQQPPQGGEP
QEDQDRGGRDQRGRQLPGPERPRQDRGQGAAGHRDQHHPERDHPDRRQPAGQPEDQQHRP
ERHPRPARQPHPRDQEQHQAETEQX
>PAPOA_HUMAN_4 Poly(A) polymerase alpha OS=Homo sapiens GN=PAPOLA PE=1 SV=4
SVQSQLDAVLDHGDGVGGQGGDVAQVGAAGLLAGQQAGGGLDGLALGGAGLGVQLLLGLG
LVVAAQGQAVGVLAGLVLLGLGLLGALLLVGAAGSLHAHDGGGDLGGGVAAGGGVAAASG
GGVHQSGAAHHSAHGGLGGGADGGLGGGLGDALAAGAAAALAAVHLGHAHLGGLDVGHAG
GGHGGGGAVSALAAAAAVQGAGLGGAGGAGHAHAVVAVHAQVQAAVVQGGQLHALGAVLL
LLLQHVVGQQLVQLLSLHVHGGDLHVHLEHLAVDGLSVHGVGEALDVVGQVHAQVLAVLG
LLEHQADHPHGSELLLVGVLLGGGEALGVHVGQGDELVLLQAAHQDSDLALHQAHPLQSL
```

Long Beach 14.4 Kb

7: Tabular-to-FASTA on data 5

6: Tabular-to-FASTA on data 5

5: Select on data 4

4: FASTA-to-Tabular on data 3

3: transeq on data 2
6 sequences
format: fasta, database: ?

```
>PAPOA_HUMAN_1 Poly(A) polymerase alph
MPFPVTTQGSQQTQPPQKHYGITSPISLAAPKETDCVL
RILILGKLNNLVKEWIREISESKNLPQSVIENVGGKIF
PRHVDRSDFFTSFYDKLKLQEEVKDLRAVEEAFVPVIK
DLDLRDDSLLKNLDIRCIRSLNGCRVTDEILHLVPNID
LGFLGGVSWAMLVARTCQLYPNAIASTLVHKFFLVFSK
```

2: backtranseq on data 1

1: P51003

why are we using Galaxy

- open source
- we can modify the tools
- we can add our own tools
  (we offer our own NGS pipeline tools, and have disabled the provided Galaxy NGS tools/wrapper)

- the “Galaxy” community is big and part of a wider community: “GenomeSpace”, “GMOD”

- it is simple to install and maintain
- you can adjust the set-up according your needs
- it is easy to track what people are doing

we are using Galaxy for:

- microarray analysis (wrapped R/Bioconductor scripts)
- NGS analysis (wrapped perl scripts)
- EMBOSS
- file format conversion
- genomic interval operations
- providing a GUI for ‘helpdesk’ scripts

Galaxy is a stepping stone

- people learn how to built workflows instead of pressing red buttons

FMI

Friedrich Miescher Institute
for Biomedical Research

Galaxy does not solve all your problems

- there are no plasmid drawing tools
- built in genome browser (“Trackster”) is Beta
- it does not replace the ‘Bioinformatician’

  → do not offer tools you don’t understand

- it does not replace the ‘sys-admin’

  → if your tool does not run on the command line, it won’t run in Galaxy

  → ‘big data’ needs ‘big toys’

- it is simple to install and maintain....but
  it does need maintenance!

FMI

Friedrich Miescher Institute
for Biomedical Research

Acknowledgment

Michael Stadler Lukas Burger

Anita Lerch Dimos Gaidatzis

Tim Roloff Stefan Grzybek

Galaxy

SIB
Swiss Institute of
Bioinformatics

UNI
BASEL

NOVARTIS

FMI
Friedrich Miescher Institute
for Biomedical Research