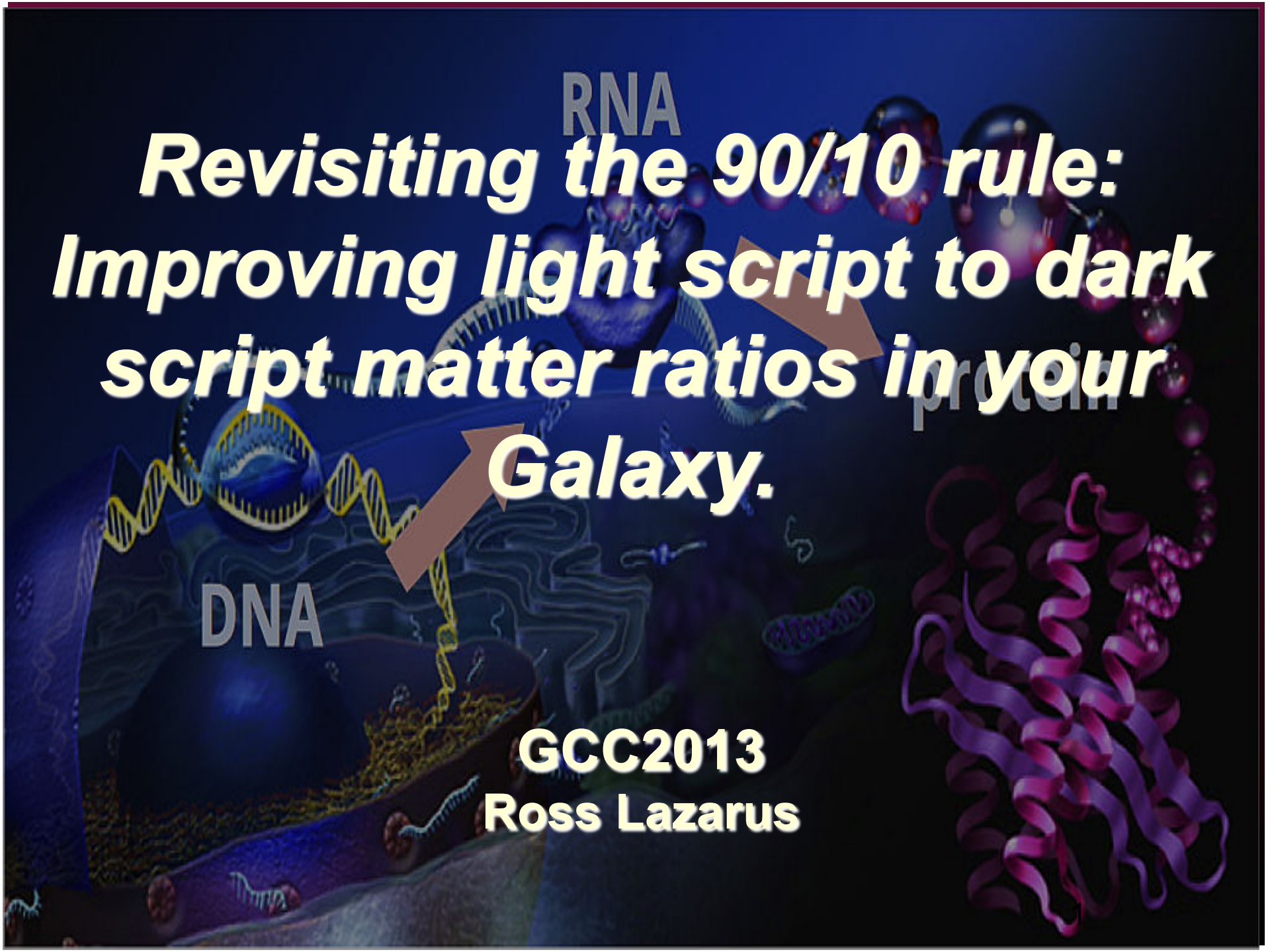


Revisiting the 90/10 rule: Improving light script to dark script matter ratios in your Galaxy.

GCC2013
Ross Lazarus

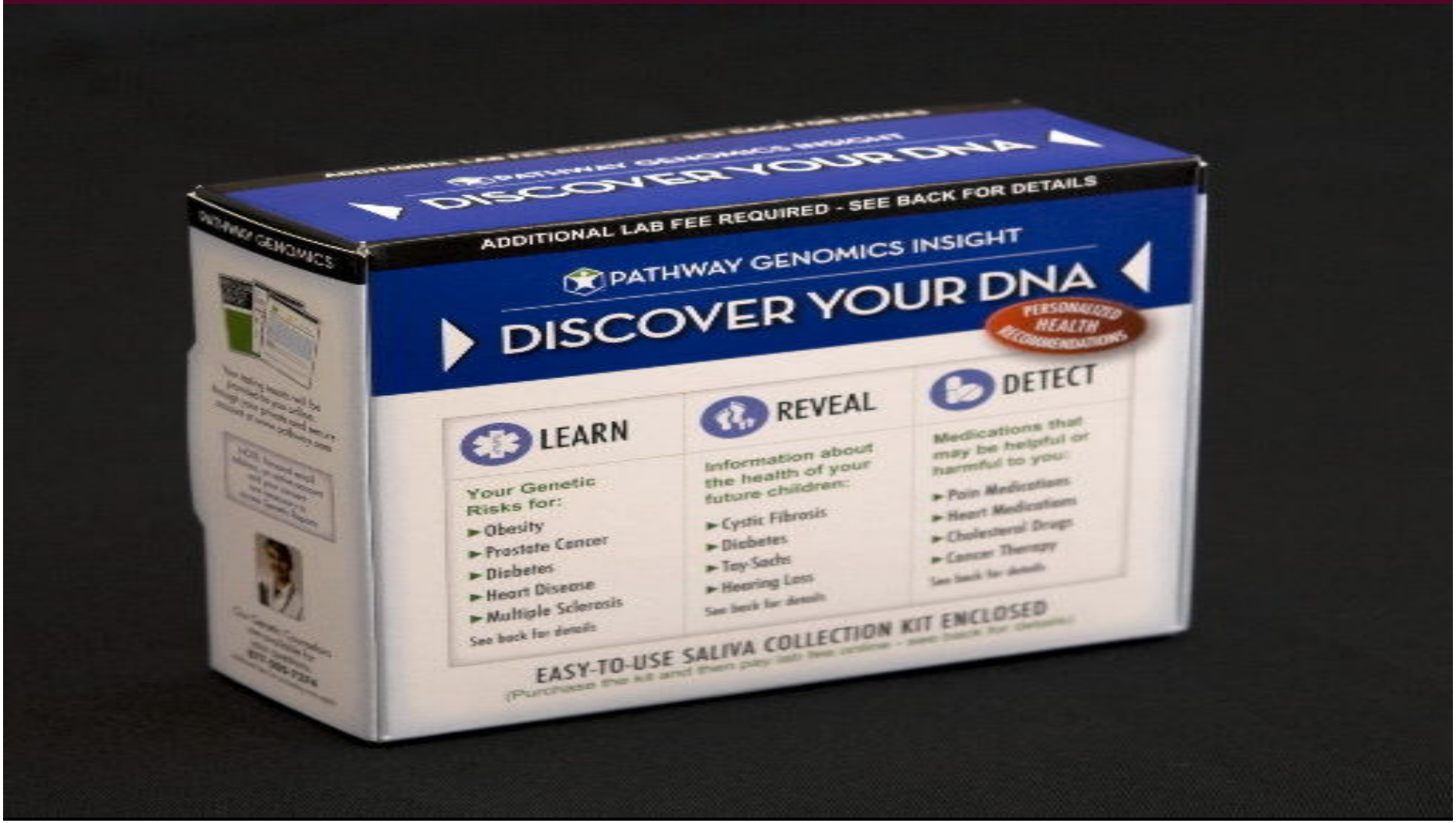
Outline

- Context: Genomic sequence research
- Reproducible analysis
- Dark script matter origins
- Making scripts reproducible
- Conclusions

Sequencing in biomedical research

- $$$ → Human genome project
- Commodity molecular technologies
- MP short read sequencing
- DNA, RNA, miR, …. ChIP-seq
- Unprecedented opportunity
- Unprecedented challenges

The Promise: Personalised Medicine



For Bioinformaticians and Biologists, things are a little more complex
down among the weeds.

Rooms full of big machines

Producing ***very*** big genomic data

600 Gbases / week
8B 75nt reads

public personal genomics

Home About Projects Members Resources Data Tools Subscribe

« Sequencing DNA with Nanopores

Cluster Sequencing with Oxford Nanopore's GridION System »

HiSeq doubles its output, a next-gen sequencing primer, and return of genetic data to patients

28/01/2011

Categories: Friday Links

Written by Katherine Morley, Daniel MacArthur, Jeff Barrett and Dan Vorhaus

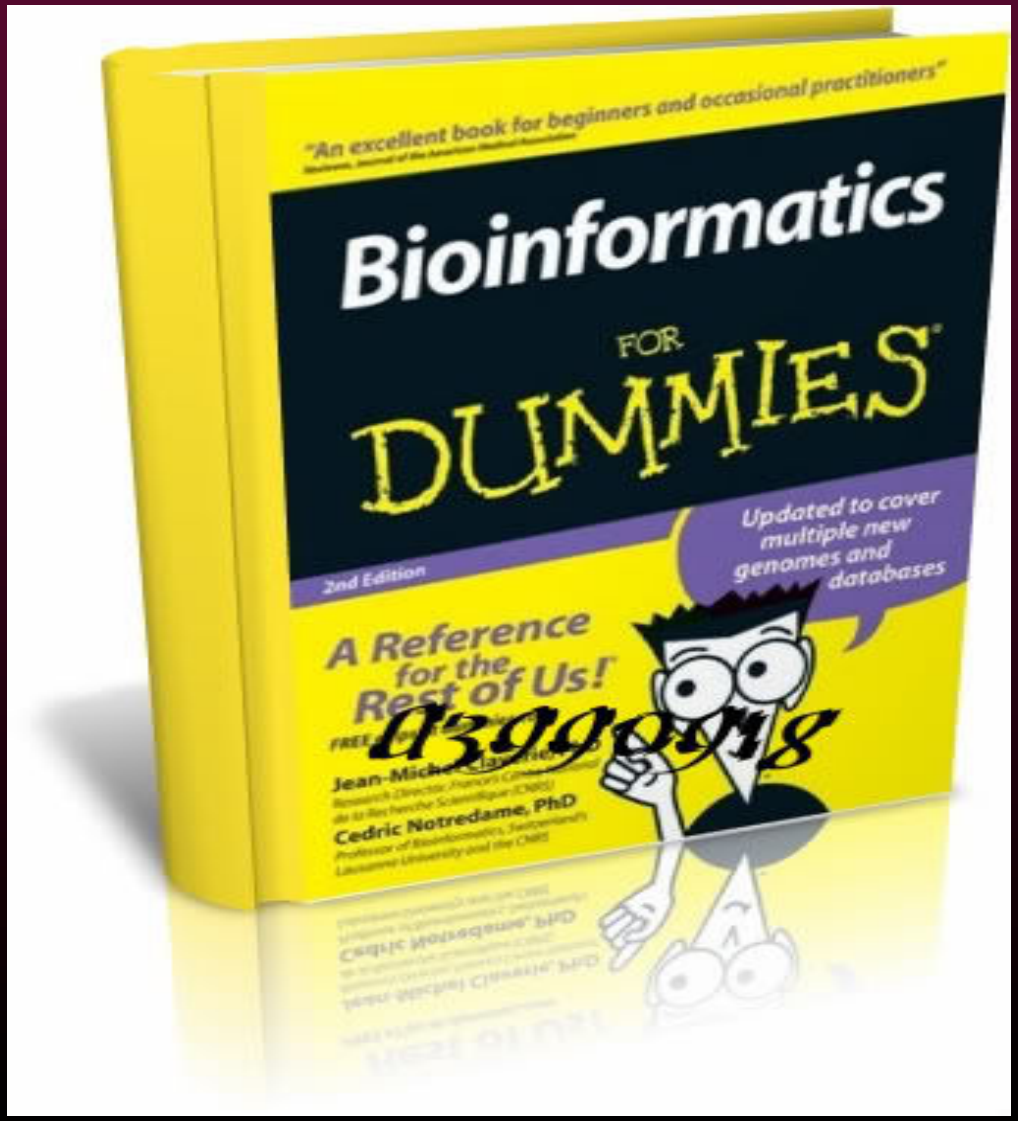
Illumina CEO Jay Flatley announced that an upgrade to their HiSeq 2000 platform expected this spring will allow users to generate 600 gigabases of sequence (the equivalent of 5 high quality human genomes) per one-week run of the machine. This would essentially double the current throughput of the platform and propel Illumina even further ahead in the arms race of delivering vast quantities of low cost sequence data. [JCB]

Raw data – A,C,G,T

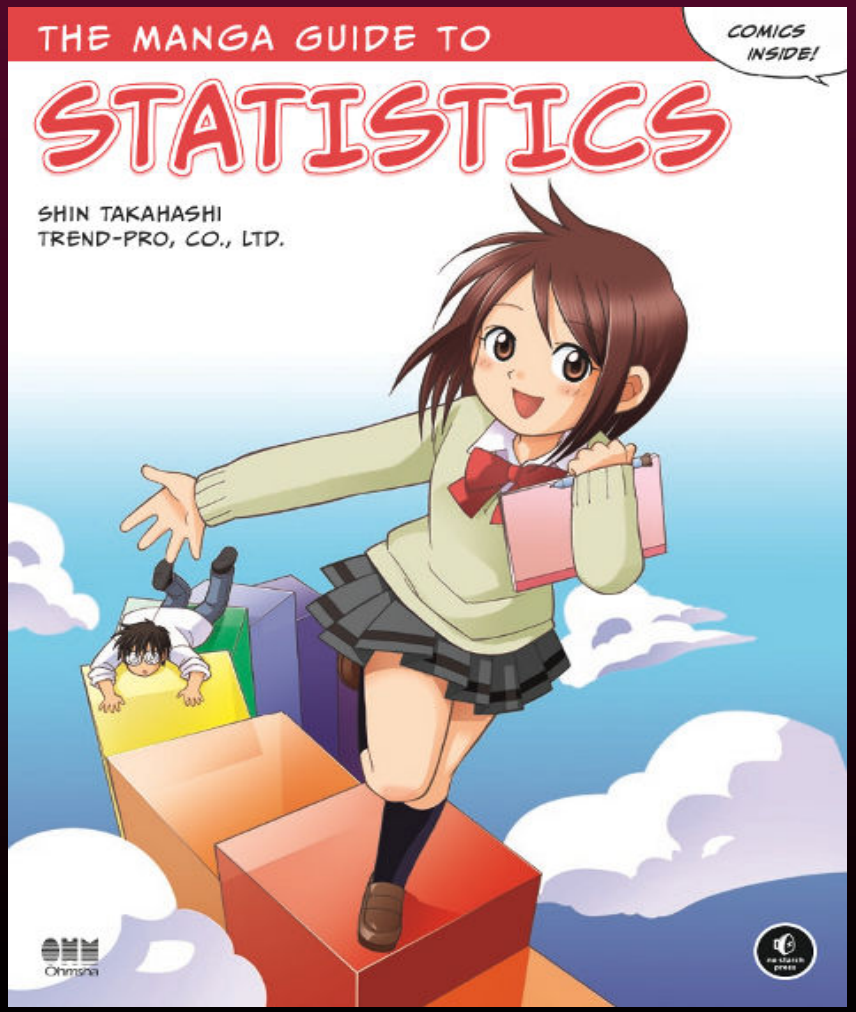
From sequence to biological insight

- Complex pipelines
- Multiple, rapidly evolving tools
- Far easier to get wrong than right
- Routinely need to fix/repeat analyses
- Reproducibility is a fundamental virtue

Rapidly changing: bioinformatics



Rapidly evolving: statistical methods and implementations



Outline

- Context: Genomic sequence in biomedicine
- **Reproducible analysis**
- Dark script matter origins
- Making scripts reproducible
- Conclusions

Goal: Reproducible Analysis

- Good science is reproducible
- Requires analyses to be reproducible
- Genomics analyses are complicated
- Tools and methods evolve rapidly
- Manual steps not reliably repeatable.
- RA depends on *automated complexity*

Reproducible code = ***light script matter***

- Replicable automated analysis code
- Source scripts, doc in VCS
- Tool Shed dependencies
- Secure - back up, access control..

Outline

- Context: Genomic sequence research
- Reproducible analysis
- **Dark script matter**
- Making scripts reproducible
- Conclusions

The 90/10 rule

- Automated systems: 80% - 90%
- At best !
- Change is the only constant

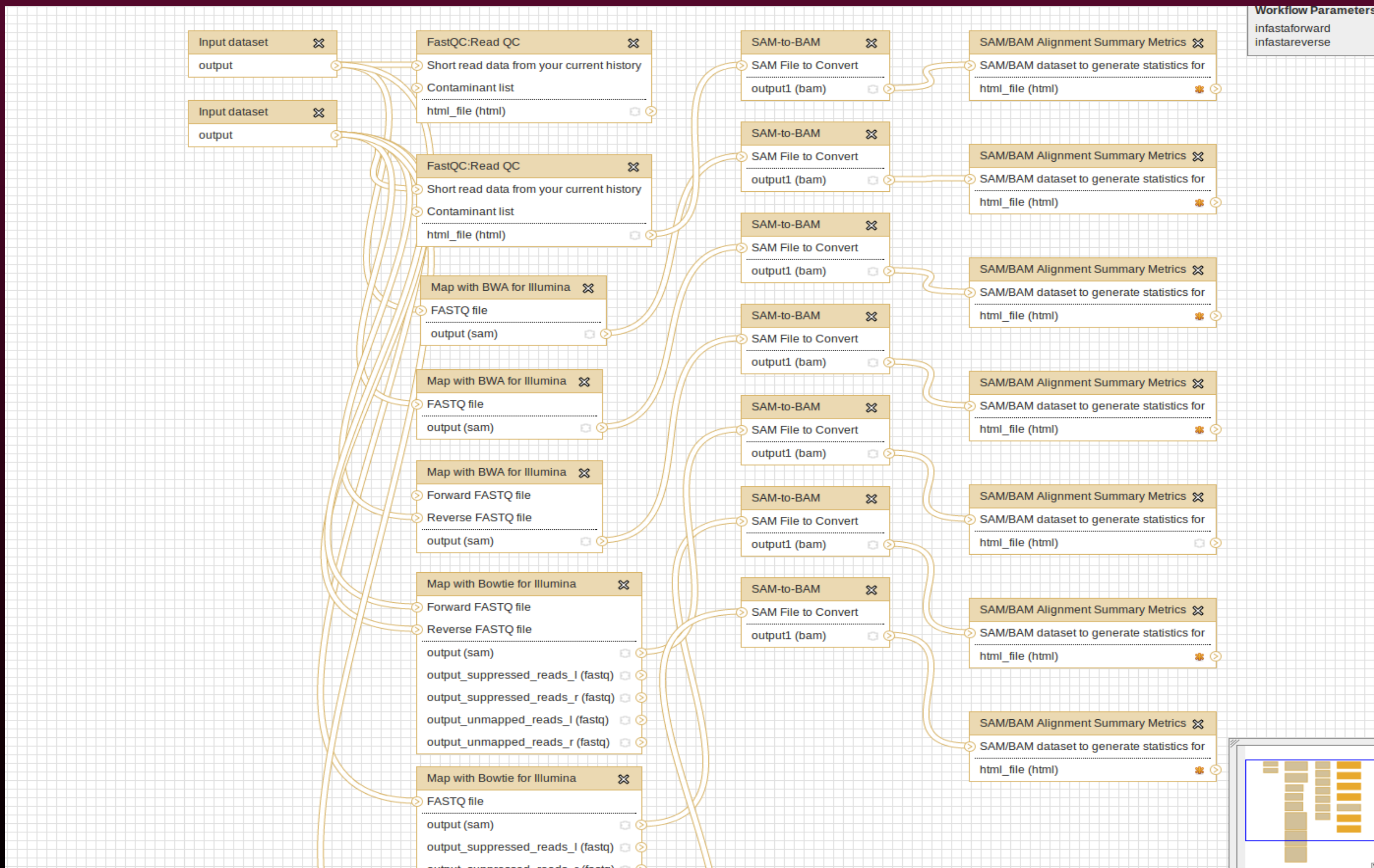
Origins of dark script matter

- Pipeline breaks – needs tweaking
- Quick script gets run
- Fixed data reinserted downstream
- Code probably not in VCS
- Reproducibility diminished

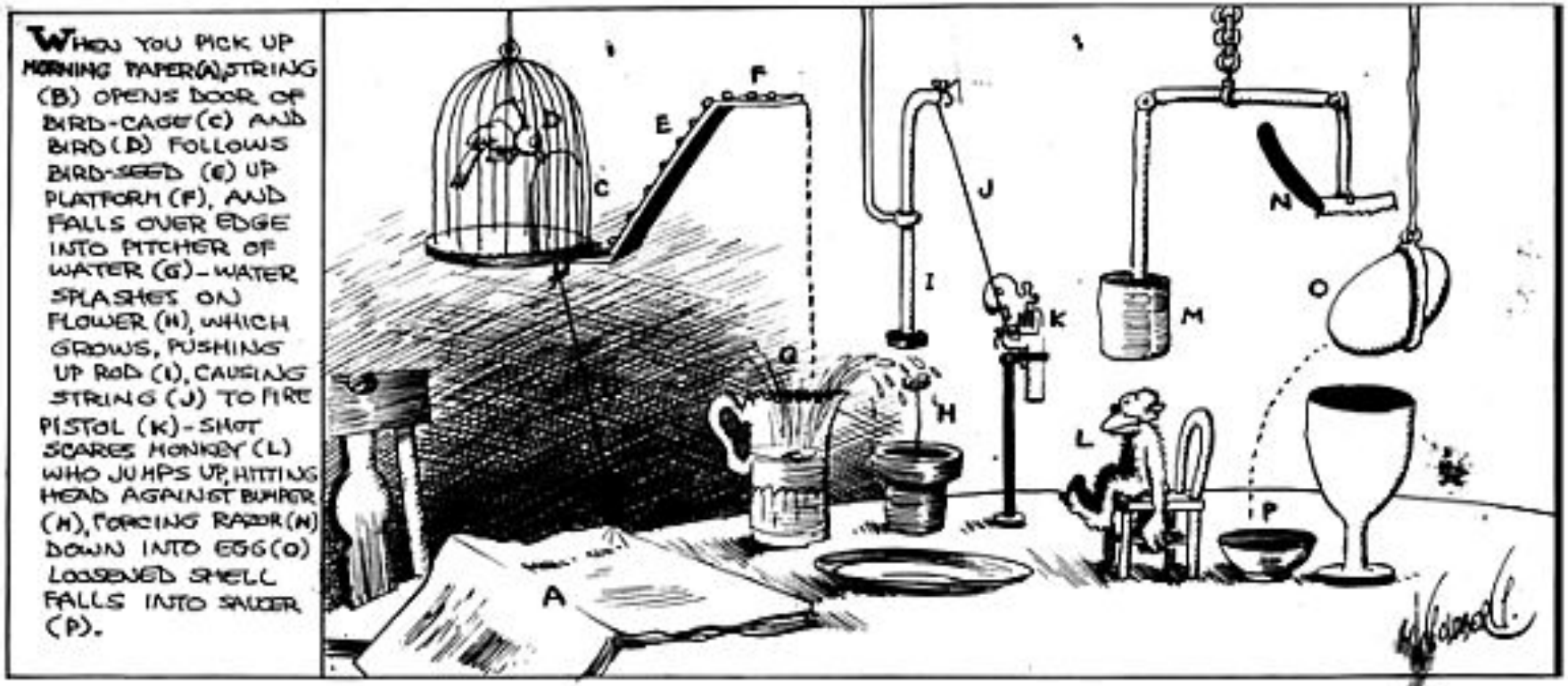
Detecting dark script matter

- Only discover it when you look for it
- Turns out to be missing – invisible
- Needs to be rewritten to rerun
- Impossible to quantify

What Biologists Need



What they sometimes end up with



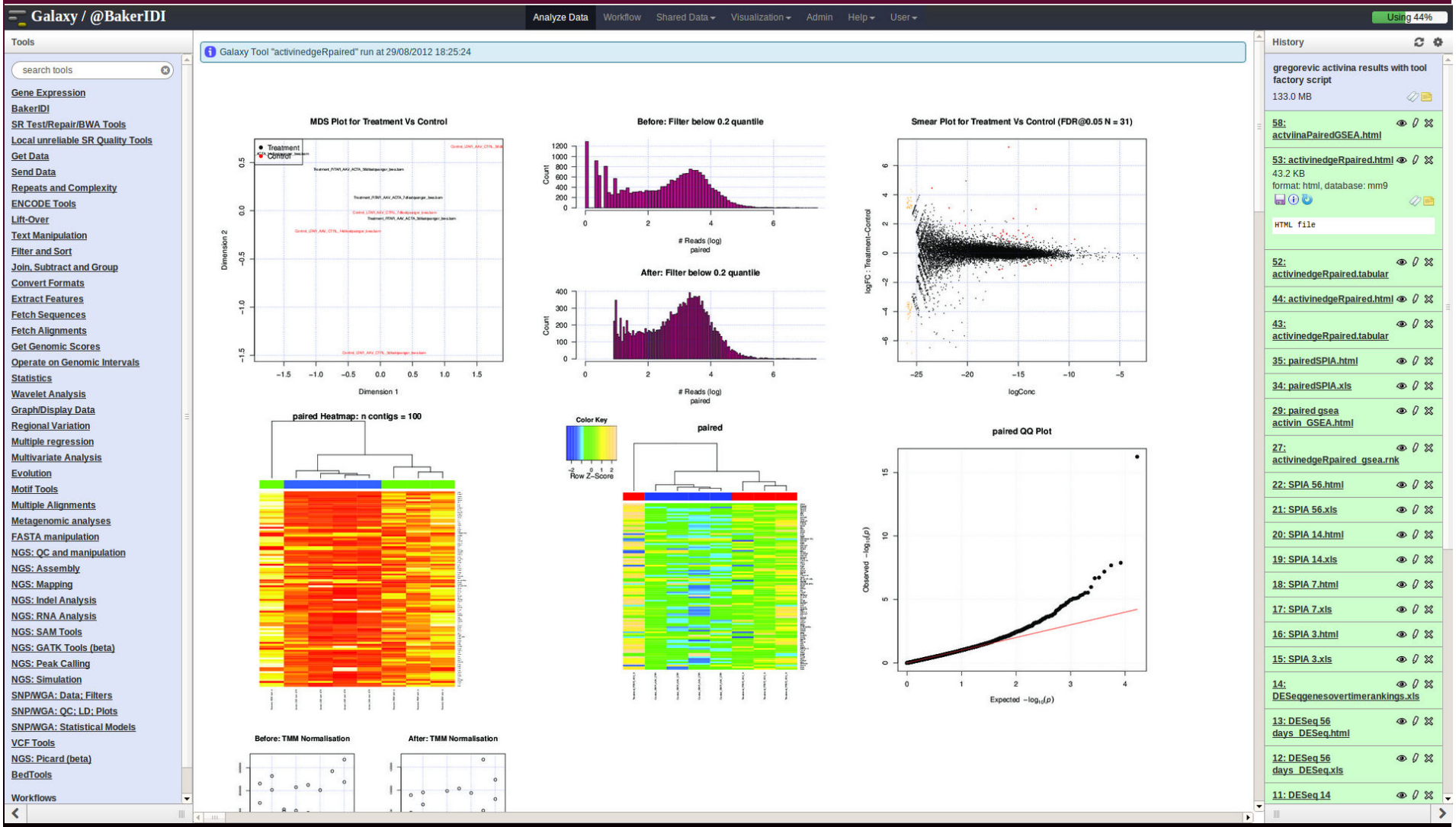
Simple way to open an egg without dropping it in your

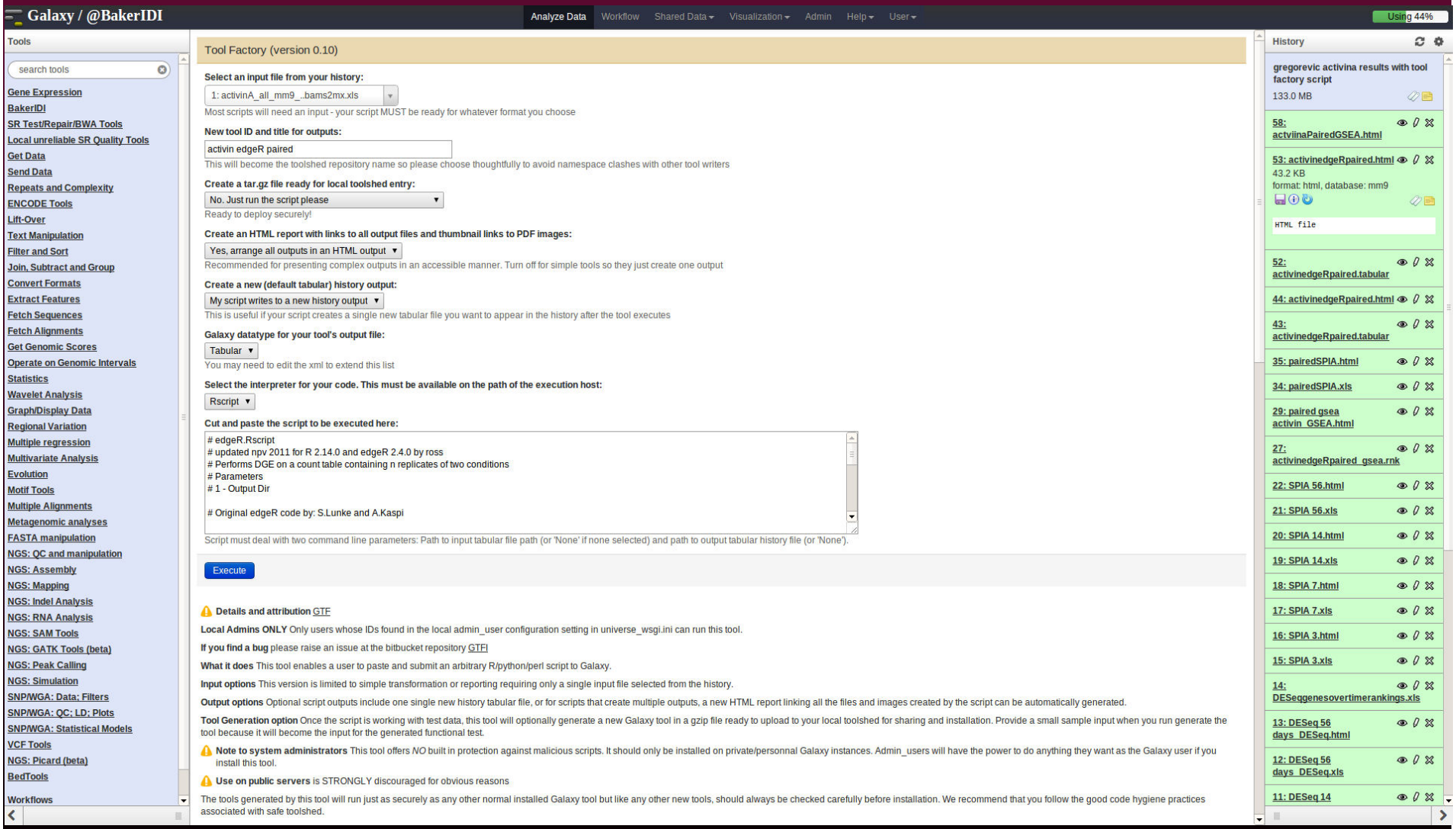
Outline

- Context: Genomic sequence research
- Reproducible analysis
- Dark script matter origins
- **Making scripts reproducible**
- Conclusions

Reproducible Quick Scripts

- Need fully automated analysis
- Shareable, transparent results
- Supported in a framework
- New Galaxy tools ideal
- Require expert resources
- Automate script → tool?


Galaxy / @BakerIDI
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 44%
Tools
search tools
Gene Expression
BakerIDI
SR Test/Repair/BWA Tools
Local unreliable SR Quality Tools
Get Data
Send Data
Repeats and Complexity
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
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: Assembly
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
VCF Tools
NGS: Picard (beta)
BedTools
Workflows
Galaxy Tool "activinedgeRpaired" run at 29/08/2012 18:25:24
MDS Plot for Treatment Vs Control
Treatment
Control
Dimension 1
Dimension 2
Before: Filter below 0.2 quantile
After: Filter below 0.2 quantile
Count
Reads (log) paired
Smear Plot for Treatment Vs Control (FDR@0.05 N = 31)
logFC : Treatment-Control
logConc
paired Heatmap: n contigs = 100
Color Key
Row Z-Score
paired
paired QQ Plot
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
Before: TMM Normalisation
After: TMM Normalisation
History
gregorevic activina results with tool factory script
133.0 MB
58: actvilinaPairedGSEA.html
53: activinedgeRpaired.html
43.2 KB
format: html, database: mm9
HTML file
52: activinedgeRpaired.tabular
44: activinedgeRpaired.html
43: activinedgeRpaired.tabular
35: pairedSPIA.html
34: pairedSPIA.xls
29: paired gsea activin_GSEA.html
27: activinedgeRpaired_gsea.rnk
22: SPIA 56.html
21: SPIA 56.xls
20: SPIA 14.html
19: SPIA 14.xls
18: SPIA 7.html
17: SPIA 7.xls
16: SPIA 3.html
15: SPIA 3.xls
14: DESeqgenesovertimerankings.xls
13: DESeq 56 days_DESeq.html
12: DESeq 56 days_DESeq.xls
11: DESeq 14


Galaxy / @BakerIDI
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 44%
Tools
search tools
Gene Expression
BakerIDI
SR Test/Repair/BWA Tools
Local unreliable SR Quality Tools
Get Data
Send Data
Repeats and Complexity
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
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: Assembly
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
VCF Tools
NGS: Picard (beta)
BedTools
Workflows
Tool Factory (version 0.10)
Select an input file from your history:
1: activinA_all_mm9_..bams2mx.xls
Most scripts will need an input - your script MUST be ready for whatever format you choose
New tool ID and title for outputs:
activin edgeR paired
This will become the toolshed repository name so please choose thoughtfully to avoid namespace clashes with other tool writers
Create a tar.gz file ready for local toolshed entry:
No. Just run the script please
Ready to deploy securely!
Create an HTML report with links to all output files and thumbnail links to PDF images:
Yes, arrange all outputs in an HTML output
Recommended for presenting complex outputs in an accessible manner. Turn off for simple tools so they just create one output
Create a new (default tabular) history output:
My script writes to a new history output
This is useful if your script creates a single new tabular file you want to appear in the history after the tool executes
Galaxy datatype for your tool's output file:
Tabular
You may need to edit the xml to extend this list
Select the interpreter for your code. This must be available on the path of the execution host:
Rscript
Cut and paste the script to be executed here:
edgeR.Rscript
updated npv 2011 for R 2.14.0 and edgeR 2.4.0 by ross
Performs DGE on a count table containing n replicates of two conditions
Parameters
1 - Output Dir
Original edgeR code by: S.Lunke and A.Kaspi
Script must deal with two command line parameters: Path to input tabular file path (or 'None' if none selected) and path to output tabular history file (or 'None').
Execute
Details and attribution GTF
Local Admins ONLY Only users whose IDs found in the local admin_user configuration setting in universe_wsgi.ini can run this tool.
If you find a bug please raise an issue at the bitbucket repository GTFI
What it does This tool enables a user to paste and submit an arbitrary R/python/perl script to Galaxy.
Input options This version is limited to simple transformation or reporting requiring only a single input file selected from the history.
Output options Optional script outputs include one single new history tabular file, or for scripts that create multiple outputs, a new HTML report linking all the files and images created by the script can be automatically generated.
Tool Generation option Once the script is working with test data, this tool will optionally generate a new Galaxy tool in a gzip file ready to upload to your local toolshed for sharing and installation. Provide a small sample input when you run generate the tool because it will become the input for the generated functional test.
Note to system administrators This tool offers NO built in protection against malicious scripts. It should only be installed on private/personnal Galaxy instances. Admin_users will have the power to do anything they want as the Galaxy user if you install this tool.
Use on public servers is STRONGLY discouraged for obvious reasons
The tools generated by this tool will run just as securely as any other normal installed Galaxy tool but like any other new tools, should always be checked carefully before installation. We recommend that you follow the good code hygiene practices associated with safe toolshed.
History
gregorevic activina results with tool factory script
133.0 MB
58: actviinaPairedGSEA.html
53: activinedgeRpaired.html
43.2 KB
format: html, database: mm9
HTML file
52: activinedgeRpaired.tabular
44: activinedgeRpaired.html
43: activinedgeRpaired.tabular
35: pairedSPIA.html
34: pairedSPIA.xls
29: paired gsea activin_GSEA.html
27: activinedgeRpaired_gsea.rnk
22: SPIA 56.html
21: SPIA 56.xls
20: SPIA 14.html
19: SPIA 14.xls
18: SPIA 7.html
17: SPIA 7.xls
16: SPIA 3.html
15: SPIA 3.xls
14: DESeqgenesovertimerankings.xls
13: DESeq 56 days_DESeq.html
12: DESeq 56 days_DESeq.xls
11: DESeq 14

Galaxy Tool Factory

- Installable Galaxy tool
- Admins can run scripts !
- Reproducible outputs
- Optional new tool generation
- App store (toolshed) compatible
- Ordinary Galaxy tools - WF ready

Tool Factory Operation

- Paste script
- Choose interpreter
- Select (→test) data input
- Optional Html: autoshow all outputs
- Execute
- Optionally generate a new tool

A tool that makes tools

- Generates a new tool XML file
- Pasted script ***inside*** XML
- As a <configfile>
- Test, test data, test output
- Downloadable Tool Shed archive

Outline

- Context: Genomic sequence research
- Reproducible analysis
- Dark script matter origins
- Making scripts reproducible
- **Conclusions**

Tool Factory

- Python, Perl, R, Bash scripts
- Paste into Tool Factory text box
- One input, one history output
- Suits workflow transformations
- (Html for any number of outputs!)
- (Can manually add complexity)

Tool Shed Compatible

- App store for Galaxy tools
- Admin can click to install
- Explicit tool versioning
- Remove DSM from your Galaxy
- Enhanced reproducibility
- Enhanced sharing of tools

Acknowledgements

- The Galaxy Team
- Antony Kaspi and Mark Ziemann
- Google “galaxy toolfactory youtube”

VIVA LA EVOLUCIÓN

GALAXY

http://usegalaxy.org