

Reproducible complex analyses: Why Galaxy matters

Galaxy Australasia Workshop

March 2014

Ross Lazarus

Outline

- Science and reproducible results
- What Galaxy is & why it matters
- Tool Shed – VCS for tools
- What's in it for you?
- Commodity reproducible analysis

Reproducibility

- Experimental findings that cannot be reproduced either by the researcher or independently are *unlikely to be valid*
- Fundamental principle of the scientific method
- *Reproducible analyses* our focus
- Reproducibility gets harder as analyses get more complex

→ Challenges for scientists

- Commodity technologies
- Complex, large scale data
- Competing analysis tool ecology
- Intricate analysis pipelines needed
- Analysis → biological insight
- Manual steps not reproducible

Context: Sequence in research

- \$\$\$ → Human genome project
- Commodity molecular technologies
- MP short read sequencing
- DNA, RNA, miR, ChIP-seq

Very big genomic data

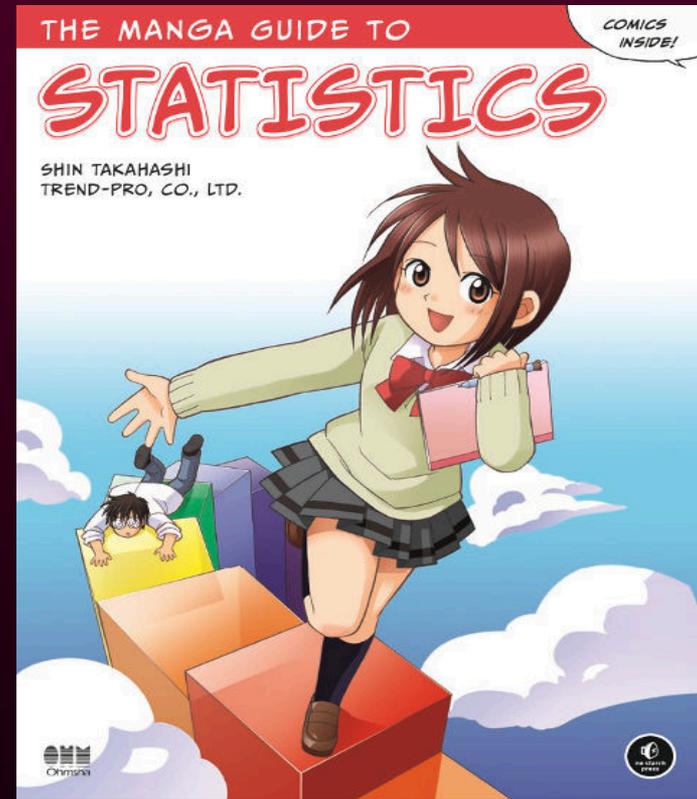
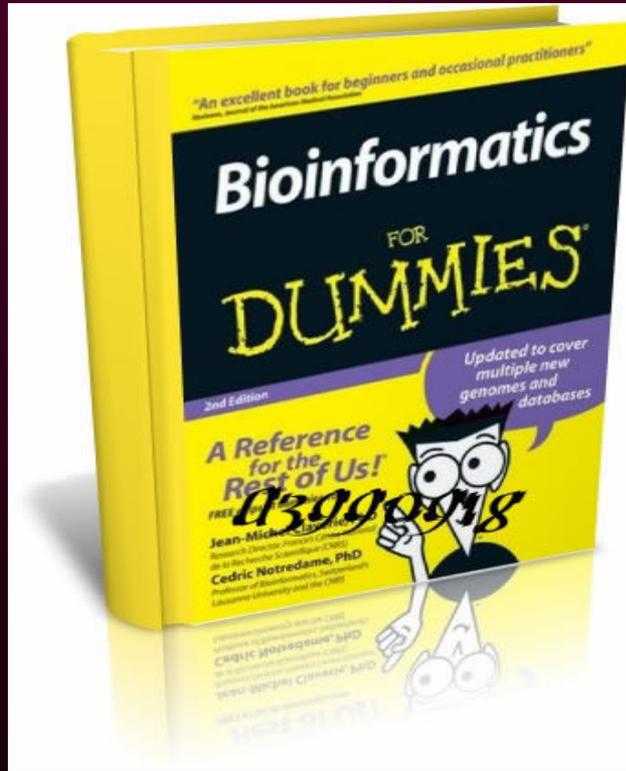
10@600 Gbase / day
60B 100nt reads



Raw data – A,C,G,T

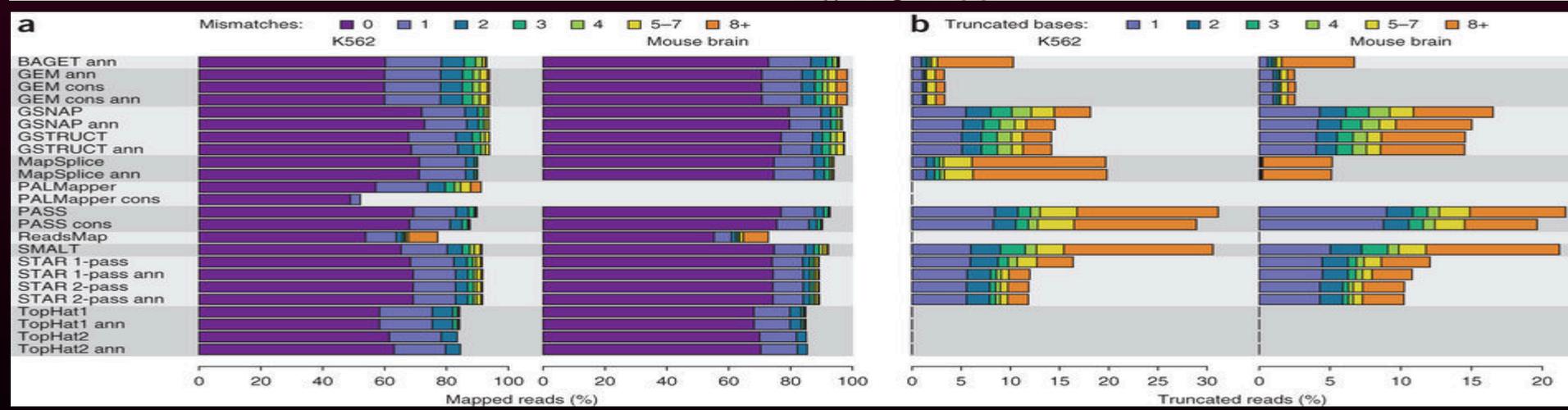
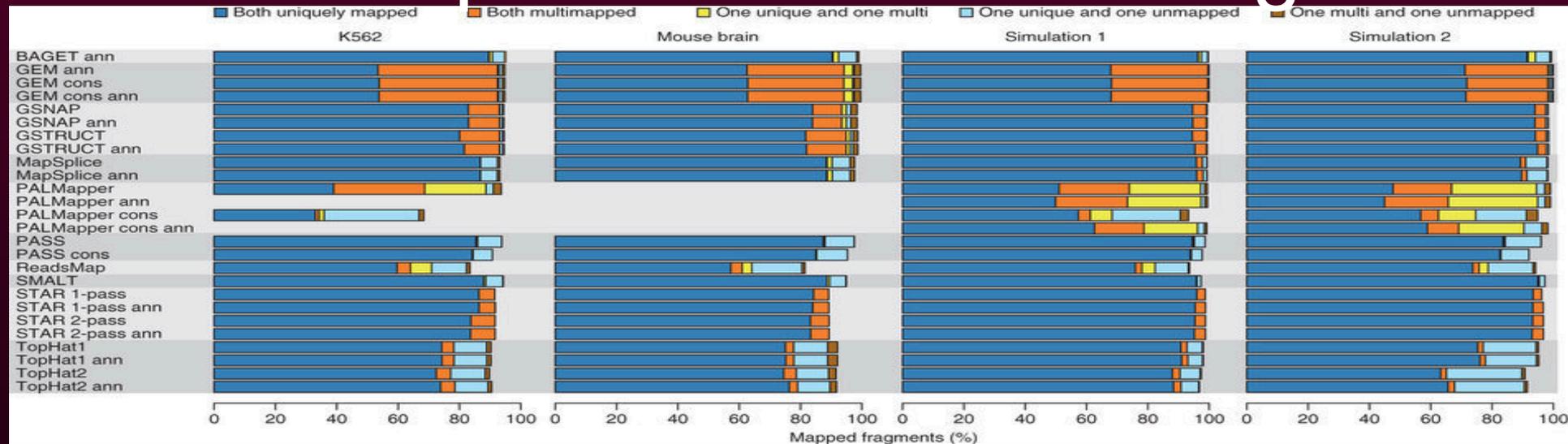


Biological insight requires:



Independent projects, all rapidly evolving

EG: 26 splice aware RNA aligners



Outline

- Science builds on reproducible results
- **What Galaxy is & why it matters**
- Tool Shed repositories – VCS for tools
- What's in it for you?
- Commodity reproducible analysis

Galaxy looks like this

Galaxy / @BakerIDI

Analyze Data Workflow Shared Data Visualization Admin Help User Using 44%

Tools

search tools

Galaxy Tool "activedgeRpaired" run at 29/08/2012 18:25:24

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

History

gregorevic.activina.results.with.tool.factory.script
133.0 MB

58: activinaPairedGSEA.html

53: activedgeRpaired.html
43.2 KB
format: html, database: mm9

HTML f11e

52: activedgeRpaired.tabular

44: activedgeRpaired.html

43: activedgeRpaired.tabular

35: pairedSPIA.html

34: pairedSPIA.xls

29: paired_gsea_activin_GSEA.html

27: activedgeRpaired_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: DESeqgenesovertime rankings.xls

13: DESeq_56_days_DESeq.html

12: DESeq_56_days_DESeq.xls

11: DESeq_14

MDS Plot for Treatment Vs Control

Before: Filter below 0.2 quantile

After: Filter below 0.2 quantile

Smear Plot for Treatment Vs Control (FDR@0.05 N = 31)

paired Heatmap: n contigs = 100

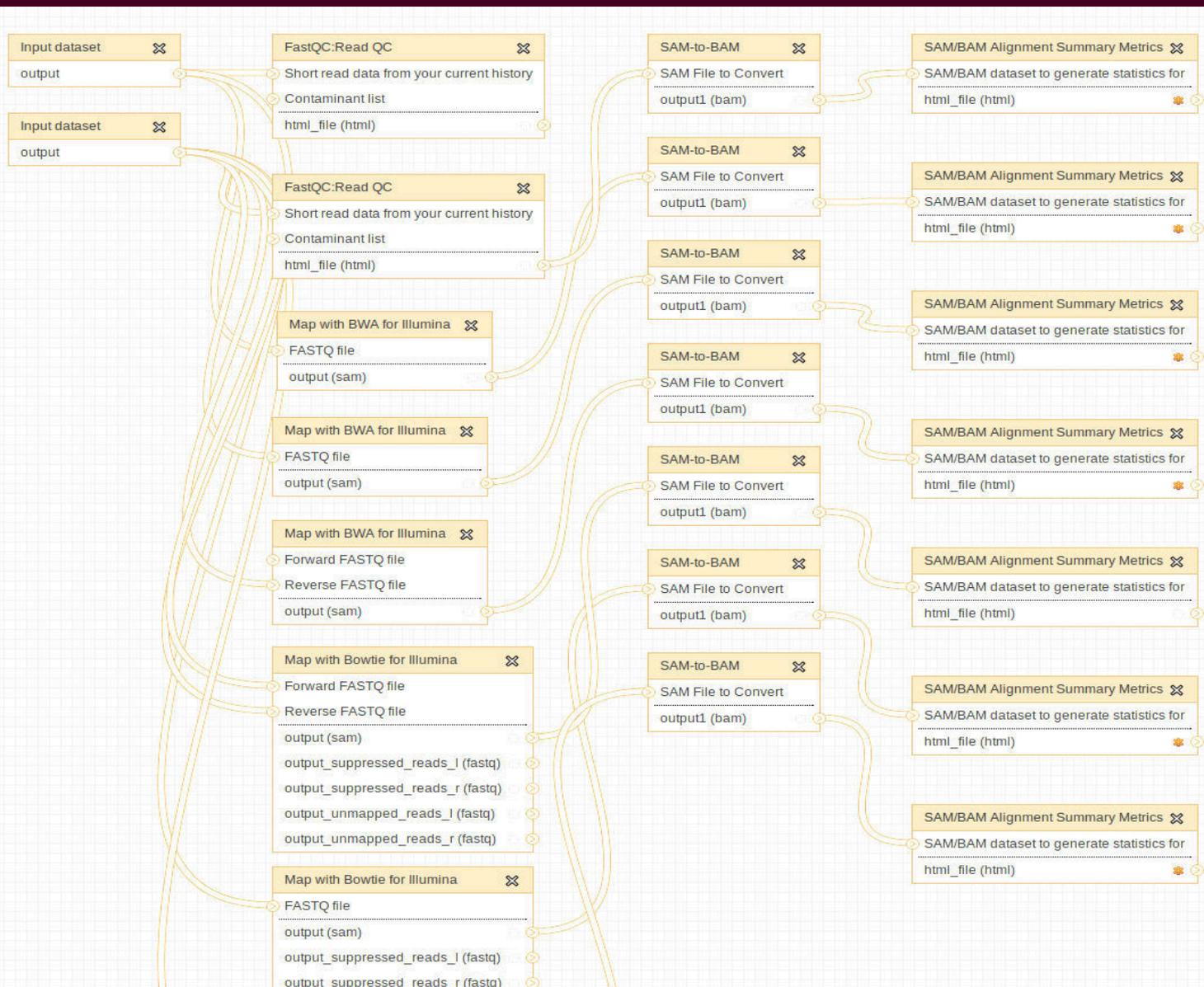
paired

paired QQ Plot

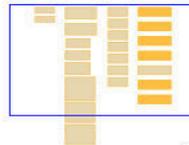
Before: TMM Normalisation

After: TMM Normalisation

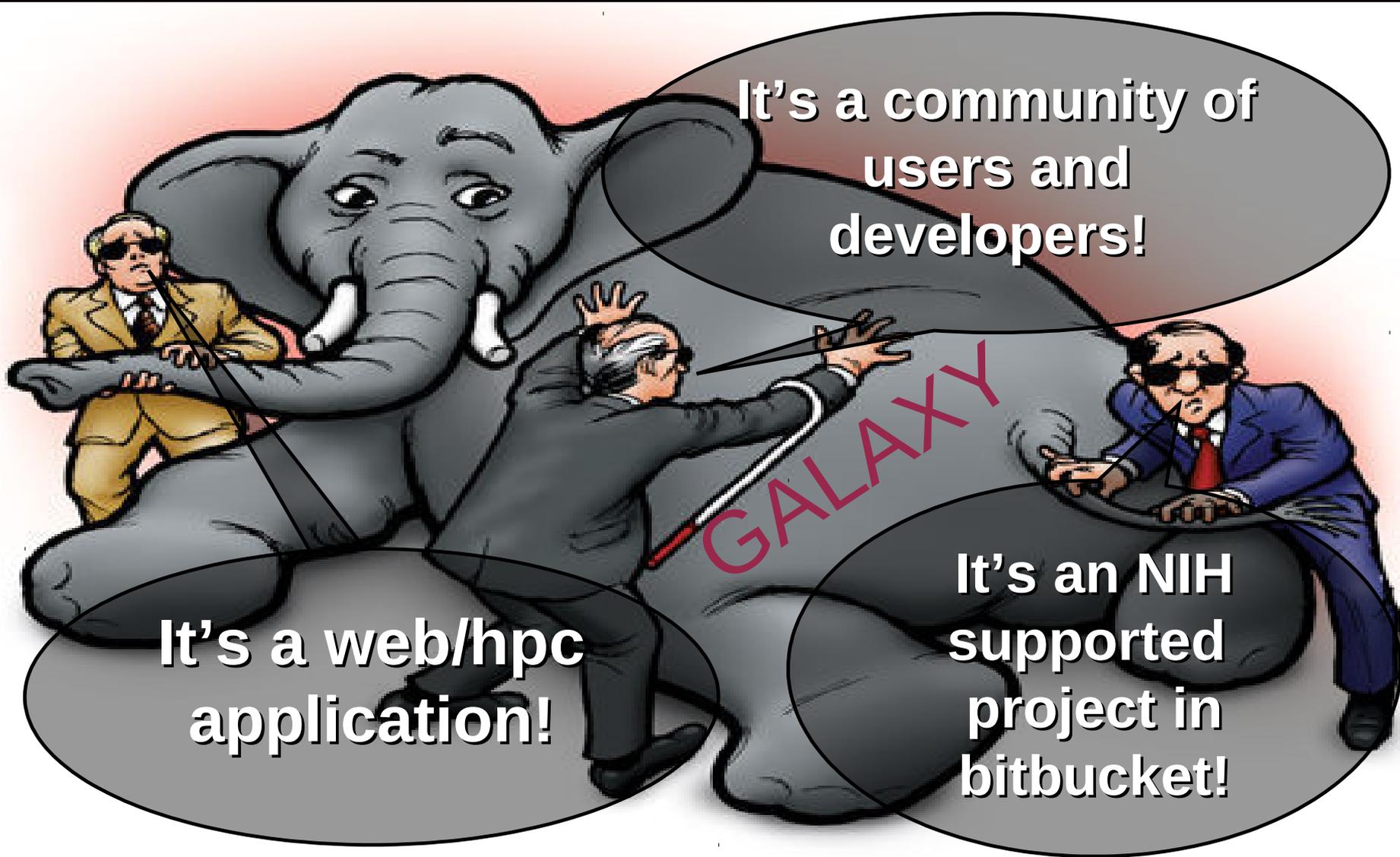
Sometimes like this



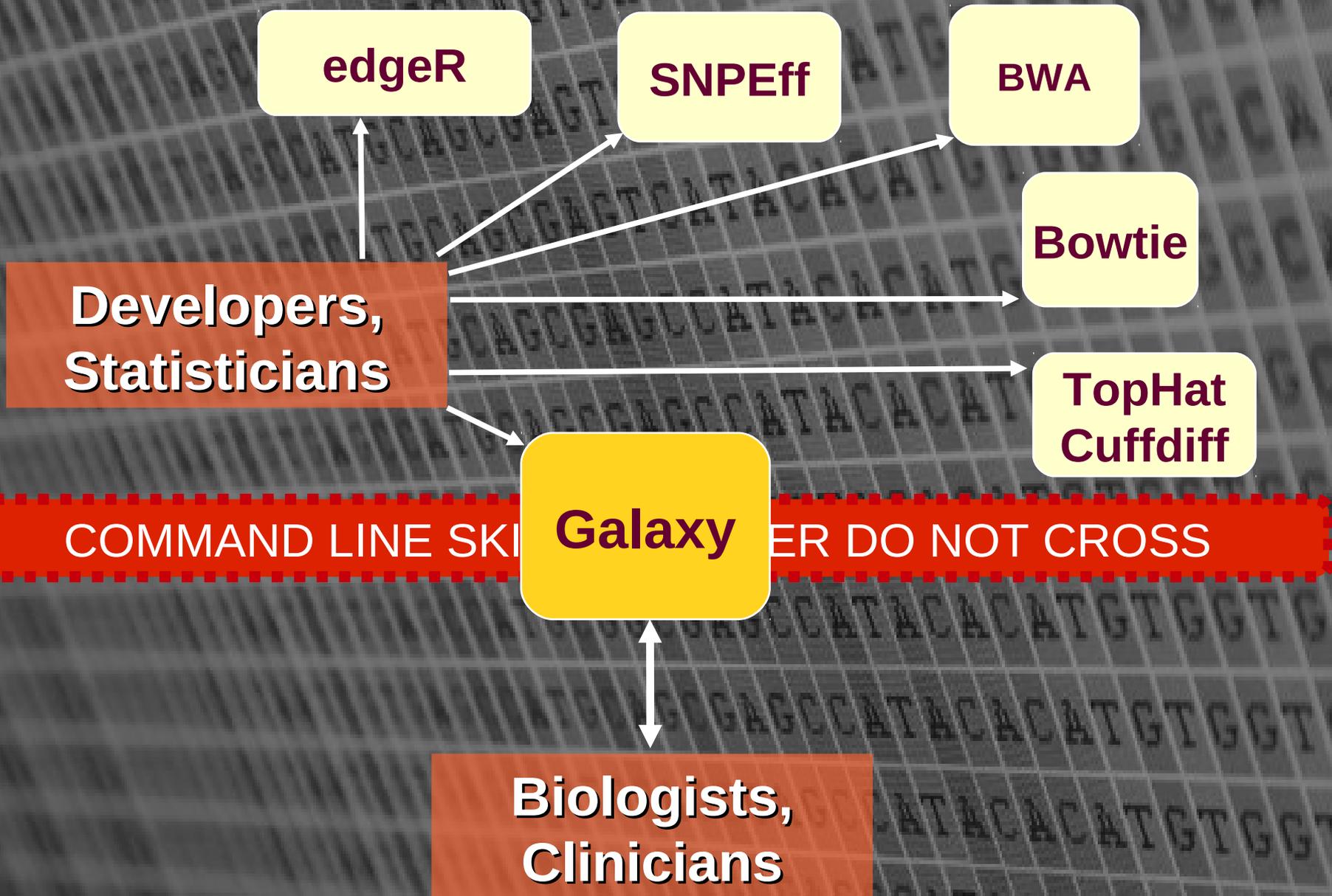
Workflow Parameters
infastaforward
infastareverse



Depends on which bit you get



Where does Galaxy fit?



Why does Galaxy matter?

- Supports complex reproducible analyses
- Low entry barriers – uniform web UI
- Exposes popular tools
- O.S.: transparent, shareable research
- Engaged user & dev communities
- Responsive support and training
- NIH/NSF/etc supported dev team

Reproducibility claim in Galaxy

- Tool Shed - VCS for Galaxy tools
- Integrated into Galaxy admin panel
- Tool/deps/jobs all version controlled
- Same revision/deps when redo old
- Latest default for new jobs
- Unaffected by system binary updates

For scientists

- Ugly technical complexities hidden
- Tools and methods evolving rapidly
- Version controlled tools and deps
- Best of breed tools – uniform UI
- Complex, repeatable workflows
- Sharable data, analyses and results

For tool developers

- Easy to package up new tools
- Self installing VCS repositories
- As repositories in Tool Sheds
- Easy way to make your work available

For Everyone

- Growing, supportive community
- Responsive support mailing lists
- wiki.galaxyproject.org for help
- Active developer community
- Tool sheds ← 100s of contributions
- Growing community source patches

Outline

- Science and reproducible results
- What Galaxy is & why it matters
- Tool Shed – VCS for tools
- What's in it for you?
- **Commodity reproducible analysis**

Reproducible analysis for all

- Efficient X of bioinformatician time
- Commodification of complex analysis
- Democratisation of reproducible analysis

Conclusion

- Welcome to the Galaxy community!
- ToolShed == App store for Galaxy
- Admin can click to install/update
- Tight tool/dependency versioning
- Enhanced reproducibility
- Enhanced sharing of tools
- (technical detail this afternoon)



VIVA LA EVOLUCIÓN

GALAXY

<http://usegalaxy.org>