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

http://www.nature.com/nmeth/journal/v10/n12/fig_tab/nmeth.2722_F1.html
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
Sometimes like this
Depends on which bit you get

It’s a web/hpc application!

It’s a community of users and developers!

It’s an NIH supported project in bitbucket!

GALAXY
Where does Galaxy fit?

Developers,
Statisticians

edgeR
SNPEff
BWA

Biologists,
Clinicians

Galaxy

COMMAND LINE SKILLS BARRIER DO NOT CROSS

Bowtie
TopHat Cuffdiff
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