State of the Galaxy

(Wie läuft’s?)

Dannon Baker, Johns Hopkins University, Galaxy Team
1. **Community** — GCC, public instances, code and tool contributions, publications

2. **Galaxy Main** — TACC move (more detail), Galaxy Main usage

3. **Recent Highlights** — pulsar, biostar, viz, dataset collections, toolshed, data managers, UI evolution

4. **Ongoing Efforts** — priority areas and things we’re excited about
State of the Community

GCC, public instances, code and tool contributions, papers
Galaxy Hackathon: Making Connections, Providing Footholds

- Developer Community Building
- Bandwidth is important
- Great opportunity for ‘fun’ projects
Hackathon Template Slide (copy me)

- Bullets
- Orange

Feel free to include screenshots/pictures. Please do include the names of people who hacked with you on the project. Pictures and logos too, if you have them!

(contributor 1*, contributor 2, contributor 3, docker)
wiki.galaxyproject.org/PublicGalaxyServers

14 General Purpose Servers
21 Domain Servers
30 Tool Publishing Servers
4 Public Toolsheds
Code Contributions

In a Nutshell, Galaxy Bioinformatics Platform...

... has had 14,841 commits made by 106 contributors
representing 429,567 lines of code

... is mostly written in JavaScript
with an average number of source code comments

... has a well established, mature codebase
maintained by a very large development team
with increasing Y-O-Y commits

... took an estimated 116 years of effort (COCOMO model)
starting with its first commit in November, 2006
ending with its most recent commit 4 days ago

Send a pull request!
The Galaxy Toolshed

897 repositories
222 unique owners
176 Tool dependency package installation recipes
2,330 valid tools
3,420 valid versions of tools
54 exported Galaxy workflows
455 custom datatypes
62,021 total repository installations
(as of July 2014)
Toolshed Contributions
Publications (this just got even easier)
The State of Galaxy Main
Galaxy Main (usegalaxy.org)
Architecture during ‘job numbers crash’

- Commodity Internet / Internet 2 / Lambda Rail
  - 1 GB
- Wartik 509
- XSEDE Network
- PSU Internal 10 GB
- Data Supercell Full Mirror
- CyberSTAR (NSF)
  - 128 Cores
  - 4 GB/core
- BioSTAR (NSF)
  - 128 Cores
  - 8 GB/core
- Pittsburgh Supercomputing Center
Galaxy Main (usegalaxy.org)

Quotas: (250/5 GB)
Concurrent Job Limiting: (6/1)
22 May 2013: Initial proposal to move Galaxy main to TACC

29 June: Galaxy Team visits TACC to plan and hack

mid August: Galaxy test running at TACC

October 7th: Galaxy main switched over to TACC completely

Continuing data migration in background...
Galaxy Main at TACC

Virtual machines

- Web Front-end 1
- Web Front-end 2
- Master Database
- Replicate Database

Corral
DDN Storage Appliance

Galaxy Cluster
256 Cores
16 GB/core

~1PB user data, may migrate to Stockyard
Job throughput post-move
usegalaxy.org userbase
Recent Highlights
Ansible for automated deployments

- Started as our private playbook for automating changes to main, galaxyproject.org, etc.

- github.com/galaxyproject/ansible-galaxy
Job Execution on External Resources

Pulsar is a distributed job execution application built for Galaxy: github.com/galaxyproject/pulsar

*Enabled On Main* for particular tools (for now): BWA, Bowtie, Bowtie2, Tophat, Tophat2, Lastz, Megablast, Cuff* with double the cores and permitted walltime per job.

XSEDE Allocation
Viz Framework
Viz Framework

Simon Gladman, Björn Grüning https://github.com/Slugger70/Vennt-galaxy-vis/
Data Libraries and other UI Evolution

It’s all about the API.

Shared Data -> “Data Libraries Beta” -- it’s FAST!

Histories (filters, multiple dataset operations, multiple history views), Tool Rendering, and more.
Dataset Collections

Run applications in parallel (one per input).

Input dataset collection
- output

Input dataset
- FASTQ Paired Dataset
  - output

Bowtie2
- Select the reference genome
  - output_unaligned_reads_1 (fastqanger)
  - output_unaligned_reads_2 (fastqanger)
  - output (bam)

FASTQ Groomer
- File to groom
  - output_file (fastqsanger, fastqcssanger, fastspolexa, fastqillumina)

Merge BAM Files
- Input Datasets
  - output1 (bam)
  - outputlog (txt)

flagstat
- BAM File to Convert
  - output1 (txt)

Merged output for subsequent processing.

Arbitrary # of Inputs (... paired).
Figure S1. Schematic overview of reference data, location files, data tables, and tools. A Galaxy Data Manager handles downloading, creating and installing each of the required facets necessary to provide built-in reference data through the use of a web-based GUI.

Reference Genome
(*.fasta)

BWA Index Files on disk

File listing paths, genome build, descriptions, etc. of available BWA indexes
(bwa_index.loc)

Galaxy BWA Tool
(bwa_wrapper.xml)

Tool Data Table
(tool_data_table_conf.xml)

abstraction layer

Galaxy's built-in reference data registry

Dynamically use built-in data in any number of Galaxy Tools
Community Interface / Help Desk
Toolshed

(particularly dependency management)
Ongoing Efforts

(see trello for details!)
Workflow Engine Overhaul

1) Stateful execution model
   - enables control structures

2) Make it pluggable
Niche-specific Tools, Guides, Workflows

Highly curated guides for particular analysis approaches - more than just Pages

Integrated ‘tour’ of Galaxy analysis
Federation / Distributed Data

*pretty much everything* should be portable

Especially important due to growing numbers of separate public, cloud, and local instances.
ToolShed, Tool Development Evolution

GitHub-backed repositories for more developer engagement
Submission process overhaul
Major UI overhaul
Tooling for local developers
Interface for truly large-scale analysis

The ‘history’ concept is fantastic for small exploratory analysis.

It does not scale for people needing to run thousands of samples.
Scalable Training

Cloud ‘workshop in a box’ and similar initiatives to distribute training materials and knowledge amongst the community.
Sample Tracking

App-based handling of samples
Thanks!

Keep the conversation going:

chat.freenode.net #galaxyproject

galaxy-dev@lists.bx.psu.edu

http://galaxyproject.org/trello

https://wiki.galaxyproject.org/GalaxyUpdates/2014_10#Galaxy_Needs_Your_Input.21