Scaling Galaxy

Preparing for those next few orders of magnitude

John Chilton, Dave Clements, and the Galaxy Team
galaxyproject.org

Slides @ bit.ly/ismb2014
Tweet w/ #usegalaxy #ismb
Data generation is cheap and will stay cheap. Scale & complexity of analysis will continue to grow. More researchers are running bioinformatics analyses of all scales and complexities.

Data generation never sleeps
Some **History**

Early emphasis on *ad hoc* exploration and analysis

Build a history as you try & retry tools.

Get your results, and a complete history of how you got them.

Histories are *1st class objects*. Can be shared, published, embedded, ...
Traditional Strengths

ad hoc learning and exploration
Protect bench scientists from command line interfaces, programming, Unix/Linux system administration
Sharing and reproducibility
Consistent and easy to use web interface
Extensible tool interface to incorporate tools
Along came **Workflows**

- **Workflow as recipe**
  - A series of steps that can be run to repeat an analysis on different data.
- **Create workflows in a couple of ways**
  - de novo using the Workflow Editor
  - Extract a workflow from a current history
- **Workflows are 1st class objects in Galaxy**
Some Workflow Extensions

- Enable hiding of intermediate datasets
  - Imagine running a 25 step workflow on 20 samples.
- Support for batch submission
  - Avoids having to start a workflow 20 times, to process 20 samples

Still, a simple concept of workflow
The Challenge

Solutions for ad hoc learning and experimenting solve different issues than do solutions that make very large analyses understandable and manageable.

Can these scalability challenges be addressed without sacrificing existing strengths?
Approaches

The problem needs to be attacked from both the client side interface (the front end) and the server side implementation (the back end).

Will also talk about social problems associated with scaling up.
Dataset Collections

Collections as 1st class objects.
Run tools in parallel over collection or on the collection as a whole. Support map/reduce paradigm.
Tools become much more dynamic, flexible and responsive to input.
Makes it possible to build workflows that can reason about paired datasets, technical replicates, multiple biological samples, ...
Core phylogenomics SNP pipeline by Aaron Petkau, Gary Van Domselaar, Philip Mabon, and Lee Katz. Worked 208 single end reads producing 1469 datasets. Galaxy took 10 minutes to schedule workflow.
Build Lists of Paired Datasets...
Dataset Collections

Much more at my Galaxy Community Conference talk

- bit.ly/gcc2014workflows
- bit.ly/gcc2014workflowsvideo
User Interface - *Basics*

Dynamic smooth, user interface
Provide data progressively on demand
Many fewer page loads
Better search mechanisms and scalable interfaces

Often implemented by switching from generated HTML to Javascript
User Interface - Visualizations

Web based visualization for high-throughput biology is a challenge. Requires client side, modular, scalable components.

General visualization framework implemented. Visualizations are 1st class objects.
Why? *Visual Analytics*

Researchers often do an analyze-summarize-visualize-repeat loop.

By bringing visualization into Galaxy we hope to tighten that loop as much as possible.
Trackster Genome browser embedded in Galaxy.
Charts
Back End Support: **Workflow System**

Replace current workflow system with, well, a workflow system. Current system could be, umm..., more robust.

Define a workflow engine plugin interface so that the workflow engine is interchangeable.
Back-End: Scaling Resources: **Compute**

**Better support heterogeneous clusters**
Galaxy is constantly providing better support for local clusters, public/private clouds, national resources - and mixes thereof.

Back-End: Scaling Resources: **Storage**

**ObjectStore**: Galaxy plugins for connecting to different storage backends

- Traditional file systems (NFS), iRODS, S3, ....
Moving usegalaxy.org to TACC

Galaxy Server Processes

Read Data

In Corral? Yes -> Corral

In Staging? Yes -> Corral Staging

In PSU? Yes -> Penn State

Write Data

Object Not Found

No ->

No ->

No ->
Galaxy API - Galaxy for the Bioinformatician

Scaling up also requires support for bioinformaticians and core staff.

Graphical user interfaces are great way to learn and explore tools...

...but running analysis from a GUI can kinda irritate a bioinformatician who is adept at scripting and command line interfaces.
Galaxy for the Bioinformatician

But if you go to the command line, you give up on Galaxy's user management, sharing, persistence, reproducibility, publishing, visualization, ... capabilities

The Galaxy API: full programmatic access to Galaxy, without going the a GUI. Allows bioinformaticians to get the best of both worlds.
Scaling for **Big Demand**

So far all about big data
That's part of the challenge

An orthogonal challenge is the **sheer number of researchers** now interested in doing bioinformatics analysis
Scaling for Big Demand: usegalaxy.org

When people think of Galaxy they often think of usegalaxy.org, the project's free (for everyone) web server. This integrates a wealth of tools, compute resources, terabytes of reference data and permanent storage.
It's good to be popular, isn't it?

Leveraging the national cyberinfrastructure for biomedical research

Scaling for Big Demand: usegalaxy.org

A centralized solution cannot support all of the users or all of the different types of needs.
Scaling for Big Demand: Open Source

http://getgalaxy.org

Galaxy installed at hundreds of organizations around the world

Working hard to ease Galaxy installation

Galaxy ToolShed

Data Managers
Scaling for Big Demand: **Public Galaxies** over 60!

Interested in:

- ChIP-seq?
  - ✓ Cistrome, Nebula

- Statistical Analysis?
  - ✓ Genomic Hyperbrowser

- Protein Synthesis?
  - ✓ GWIPS-viz

- *de novo* assembly?
  - ✓ GigaGalaxy

- Reasoning with ontologies?
  - ✓ GO Galaxy

- Repeats?
  - ✓ RepeatExplorer

Scaling for Big Demand: Cloud

https://wiki.galaxyproject.org/Cloud
Scaling for Big Demand: **Support**

[Image of Biostars forum and Galaxy interface]

- biostar.usegalaxy.org/
Scaling for Big Demand: **Gather**

Stop by on your way to **ISMB/ECCB 2015**

---

Peter Cock @pjacock • Jul 2

Next year's Galaxy Community Conference will be 6-8 July 2015 in Norwich, England

#usegalaxy [pic.twitter.com/ETCsCbb5md](https://twitter.com/ETCsCbb5md)