Scaling Galaxy for Big Data
Prepping for those next few orders of magnitude

NGS Data after the Gold Rush
The Genome Analysis Centre (TGAC)
Norwich, United Kingdom

7 May 2014

Dave Clements (and the Galaxy Team)
Johns Hopkins University

http://galaxyproject.org/
Big Data generation is cheap and will stay cheap. Larger / more complex analysis are being done. More and more people are running bioinformatics analysis of all complexities! Scalability haunts us. Data generation never sleeps.
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
Unified Genotyper (version 0.0.6)

Choose the source for the reference list:

- History

BAM files

- $1==input_file <input_file>

BAM file 1

BAM file:

- 56: (hidden) Map with BWA for Illumina on data 2, data 7, and data 3: mapped reads

Using reference file:

- 3: AgamP3 assembly
- $1==reference_sequence <reference_sequence>

Binding for reference-ordered datas

- D==dbsnp <dbsnp>

Genotype likelihoods calculation model to employ:

- BOTH:

  - $1==genotype_likelihoods_model <genotype_likelihoods_model>

The minimum phred-scaled confidence threshold at which variants not at ‘trigger’ track sites should be called:

- 30.0

- $1==standard_min_confidence_threshold_for_calling <standard_min_confidence_threshold_for_calling>

The minimum phred-scaled confidence threshold at which variants not at ‘trigger’ track sites should be emitted (and filtered if less than the calling threshold):

- 30.0

- $1==standard_min_confidence_threshold_for_emitting <standard_min_confidence_threshold_for_emitting>

Basic or Advanced GATK options:

- Basic

Basic or Advanced Analysis options:

- Basic

Execute

usegalaxy.org
Showing the Infravec history published by Dan Lawson
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 linking datasets
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)
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 static HTML to Javascript
User Interface

Visualization

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
Charts

Bar diagram:

X axis:

Axis label: X-axis
Provide a label for the axis.

Axis value type: Float
Select the value type of the axis.

Axis tick format: 0.1
Select the tick format for the axis.

Y axis:

Axis label: Y-axis
Provide a label for the axis.

Axis value type: Float
Select the value type of the axis.

Axis tick format: 0.1
Select the tick format for the axis.

Others:

Show legend: Yes
Would you like to add a legend?
Charts
Trackster

Genome browser embedded in Galaxy.
Why do all this?

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.
Support dataset collections as 1st class objects. Run tools once on each dataset in the collection. Run tools on the collection as a whole. Tools become much more dynamic, flexible and responsive to input. Support map/reduce paradigm.

Makes it possible to build workflows that can reason about paired datasets, technical replicates, multiple biological samples, ...
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 API:** 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.
Back End Support: **Workflow System**

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

**Define a workflow engine API** so that the workflow engine is interchangeable.
Better support heterogeneous back ends: Local cluster, institutional cloud, public cloud, national supercomputing centres, ...

Enable Galaxy instances to be setup to exploit these resources.
Back-End: Scaling Resources: **Storage**

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

NFS, iRODS, S3, ....
ObjectStore in action: usegalaxy.org move to TACC

Galaxy Server Processes

Read Data

- In Corral? (Yes or No)
  - Yes: Corral
  - No: In Staging? (Yes or No)
    - Yes: Corral Staging
    - No: In PSU? (Yes or No)
      - Yes: Penn State
      - No: Object Not Found

Write Data
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.

However, *a centralized solution cannot support the different analysis needs of the entire world.*
It's good to be popular, isn't it?

Leveraging the national cyberinfrastructure for biomedical research
Scaling for Big Demand: Open Source Software

http://getgalaxy.org

Galaxy is installed at organizations around the world

Working hard to make installing a local Galaxy easy

Galaxy ToolShed

Data Managers
Scaling for Big Demand: **Public Galaxy Instances**


**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

Over 60 public Galaxy servers
Scaling for Big Demand: Galaxy on the Cloud

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

A ready-to-use appliance
(BioTeam)

Cloud-based solutions
(ABgenomica, AIS, Appistry, GenomeCloud)

Consulting & Customization
(Arctix, BioTeam, Deena Bioinformatics)
Scaling for Big Demand: **Support**

Tens of thousands of users leads to a lot of questions.

Absolutely have to **encourage community support**.

Project traditionally uses mailing list.

Just moved the **user support list** to Galaxy Biostar, an **online forum**, that uses the Biostar platform.

![Galaxy-User]

*is now Biostars*
Scaling for Big Demand: **Mailing Lists**

wiki.galaxyproject.org/MailingLists

**Galaxy-Dev**
Questions about developing for and deploying Galaxy
High volume (5200 posts in 2013, 900+ members)

**Galaxy-Announce**
Project announcements, low volume, moderated
Low volume (47 posts in 2013, 3400+ members)

**Galaxy-User (deprecated)**
Questions about using Galaxy and usegalaxy.org
High volume (1328 posts in 2013, 2600+ members)
Scaling for Big Demand: **Screencasts**

“How to” screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject
Scaling for Big Demand: Feedback and guidance

http://wiki.galaxyproject.org

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

- **Accessible**: Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible**: Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent**: Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

This is the Galaxy Community Wiki. It describes all things Galaxy.

Use Galaxy

Galaxy's public service web site makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive user documentation (applicable to any public or local Galaxy instance) is available on this wiki and elsewhere.

[usegalaxy.org](http://usegalaxy.org)

Deploy Galaxy

Galaxy is open source for all organizations. Local Galaxy servers can be set up by downloading and customizing the Galaxy application.

- Admin
- Cloud
- Galaxy Appliance

[getgalaxy.org](http://getgalaxy.org)

Community & Project

Galaxy has a large and active user community and many ways to Get Involved.

- Community
- News
- Events
- Support
- Galaxy Project

Contribute

- **Users**: Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.
- **Deployers and Developers**: Contribute tool definitions to the Galaxy Tool Shed (making it easy for others to use those tools on their installations), and code to the core release.
- **Everyone**: Get Involved!
Scaling for Big Demand: **Gather**

**Galaxy Australasia Workshop 2014**

Support community organized efforts and events.

- **Swiss German Galaxy Tour**
  - Bern: 30 Sep - 1 Oct
  - Freiburg: 2 Oct
Scaling for Big Demand: **Training**

Workshops in Norwich (this Friday) and Edinburgh (next week)

https://wiki.galaxyproject.org/Events
Galaxy is hiring post-docs and software engineers

Please help.
http://wiki.galaxyproject.org/GalaxyIsHiring
Also Thanks To

Matt Drew
Vicky Schneider-Gricar
Helen Tunney
Thank you

Dave Clements
Galaxy Project
Johns Hopkins University
outreach@galaxyproject.org