Scaling Galaxy

Preparing for those next few orders of magnitude

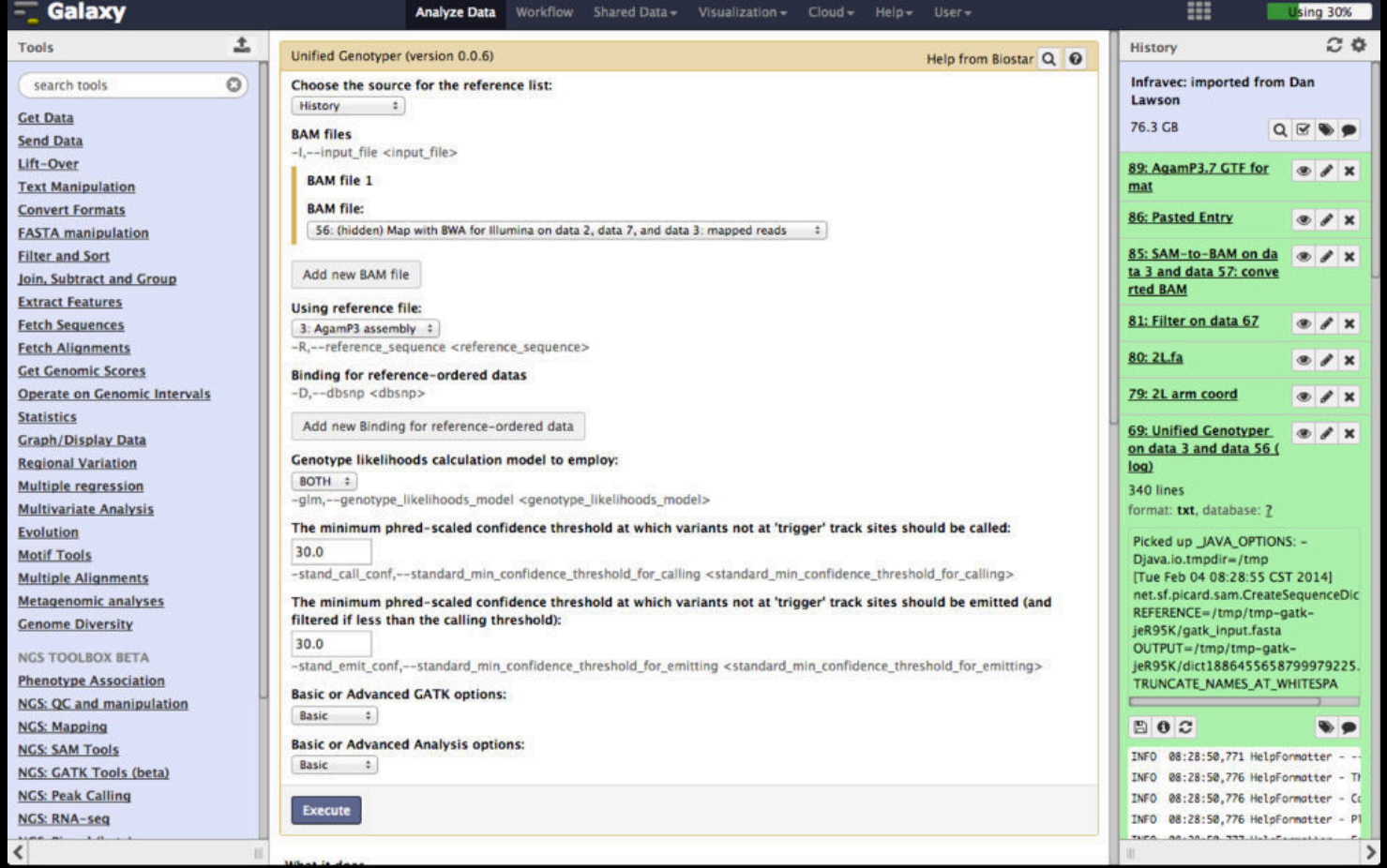
John Chilton, Dave Clements, and the Galaxy Team
galaxyproject.org

Slides @ bit.ly/ismb2014
Tweet w/ #usegalaxy #ismb

Big…

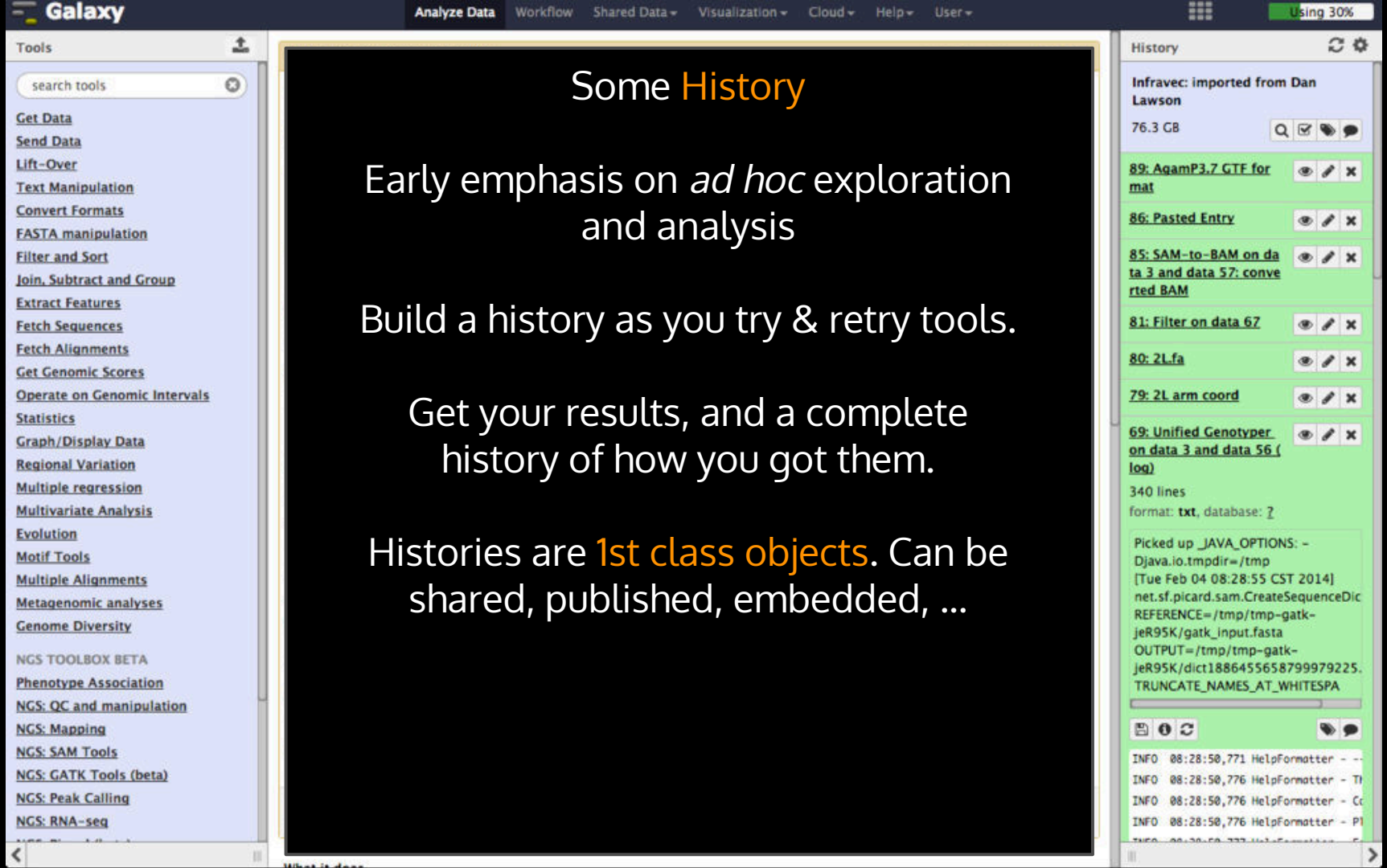
Data generation is cheap and will stay cheap.
Scale & complexity of analysis will continue to grow.
More researchers are running bioinformatics analyses of all scales and complexities.

Data generation never sleeps



usegalaxy.org

Showing the Infravec history published by Dan Lawson


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 30%
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity
NGS TOOLBOX BETA
Phenotype Association
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: RNA-seq
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, ...
History
Infravec: imported from Dan Lawson
76.3 GB
89: AgamP3.7 GTF for mat
86: Pasted Entry
85: SAM-to-BAM on da ta 3 and data 57: conve rted BAM
81: Filter on data 67
80: 2L.fa
79: 2L arm coord
69: Unified Genotyper on data 3 and data 56 (log)
340 lines
format: txt, database: ?
Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/tmp
[Tue Feb 04 08:28:55 CST 2014] net.sf.picard.sam.CreateSequenceDic REFERENCE=/tmp/tmp-gatk-jeR95K/gatk_input.fasta OUTPUT=/tmp/tmp-gatk-jeR95K/dict1886455658799979225. TRUNCATE_NAMES_AT_WHITESPA
INFO 08:28:50,771 HelpFormatter -
INFO 08:28:50,776 HelpFormatter - T
INFO 08:28:50,776 HelpFormatter - C
INFO 08:28:50,776 HelpFormatter - P

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

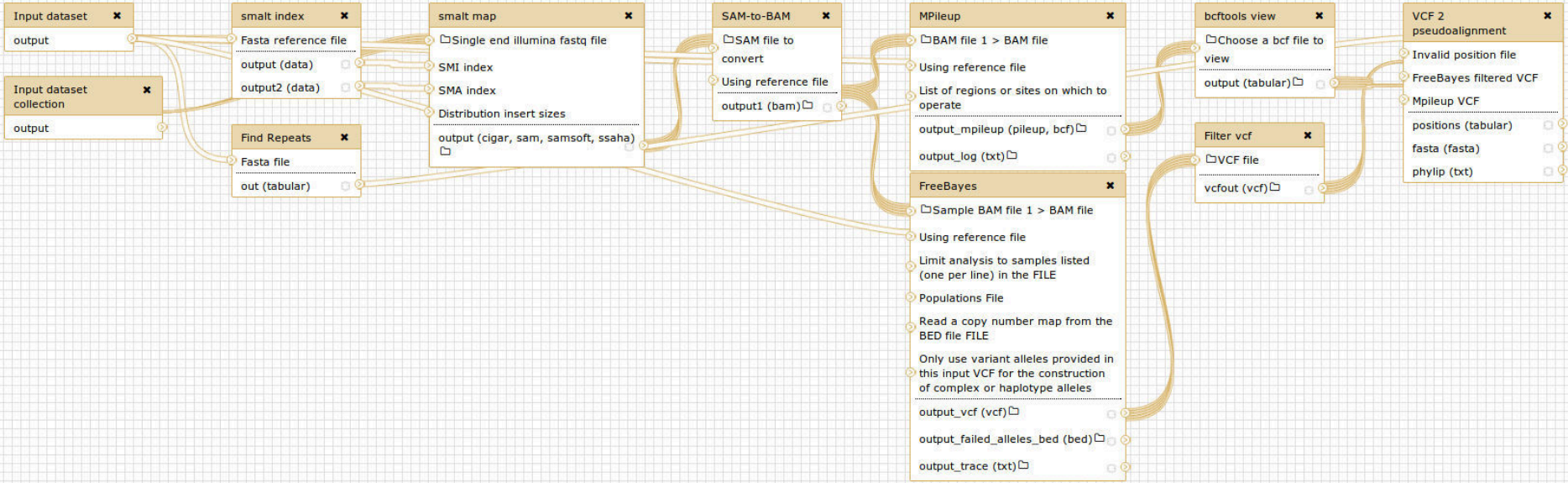
Collections as 1st class objects.

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.

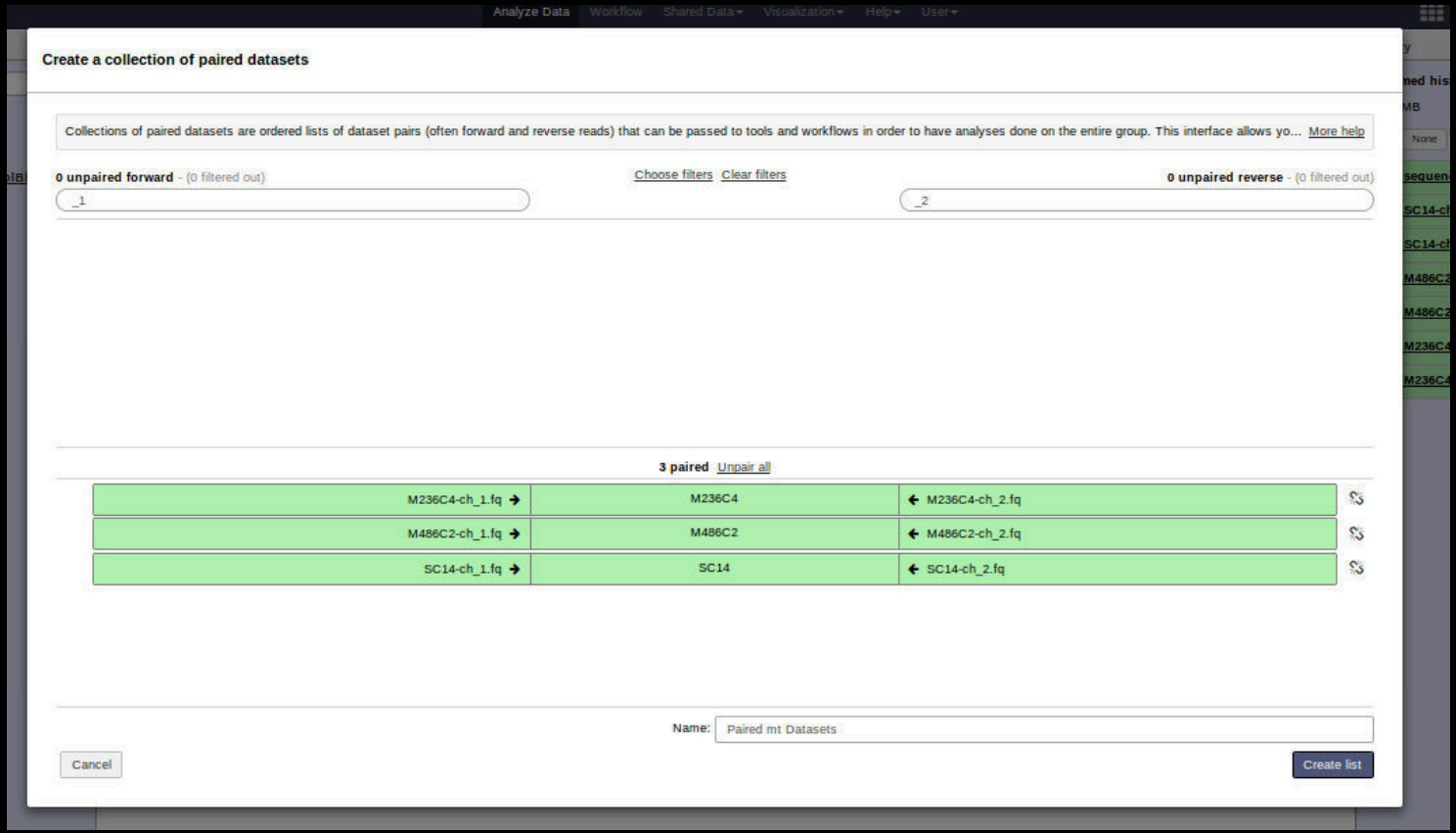
Makes it possible to build workflows that can reason about paired datasets, technical replicates, multiple biological samples, ...

More Powerful Workflows...



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

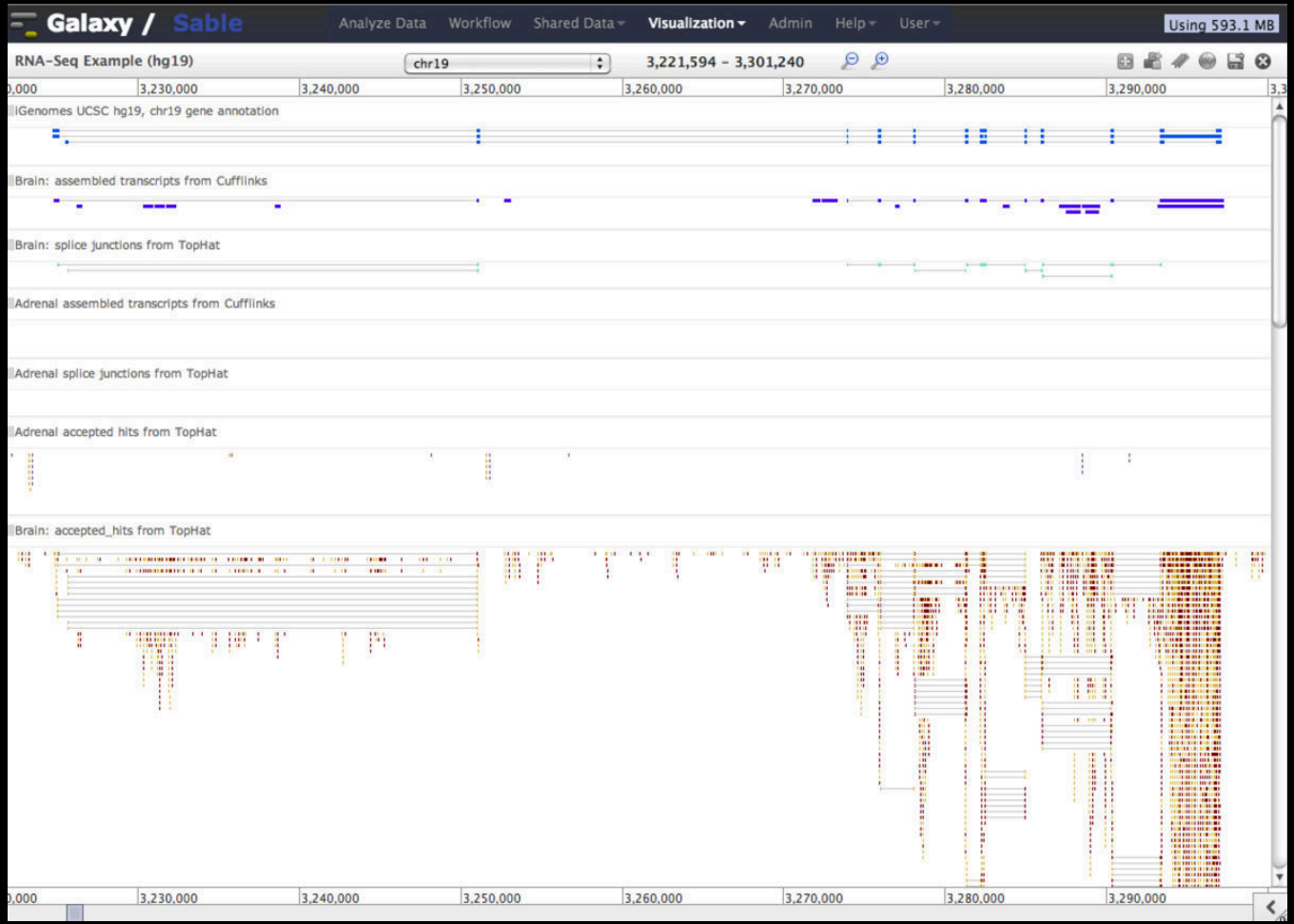
Why? *Visual Analytics*

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


Charts - New Chart
List of created charts:
Delete
New
New Chart
Bar diagram
Last change: 9/3/2014, 14:06
New Chart
Customize
Grouped
Stacked
1:Data label
2:Data label
3:Data label
4:Data label
5:Data label
362.0
300.0
250.0
200.0
150.0
100.0
50.0
0.0
Y-axis
12.0
25.0
38.0
51.0
64.0
77.0
90.0
X-axis
Delete
New
Last change: 9/3/2014, 14:07
Last change: 9/3/2014, 14:06
New Chart
Customize
Stacked
Stream
Expanded
3984.6
0.0
Y-axis
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
99.0
X-axis

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.

BOSC Talk: http://bit.ly/bosc2014

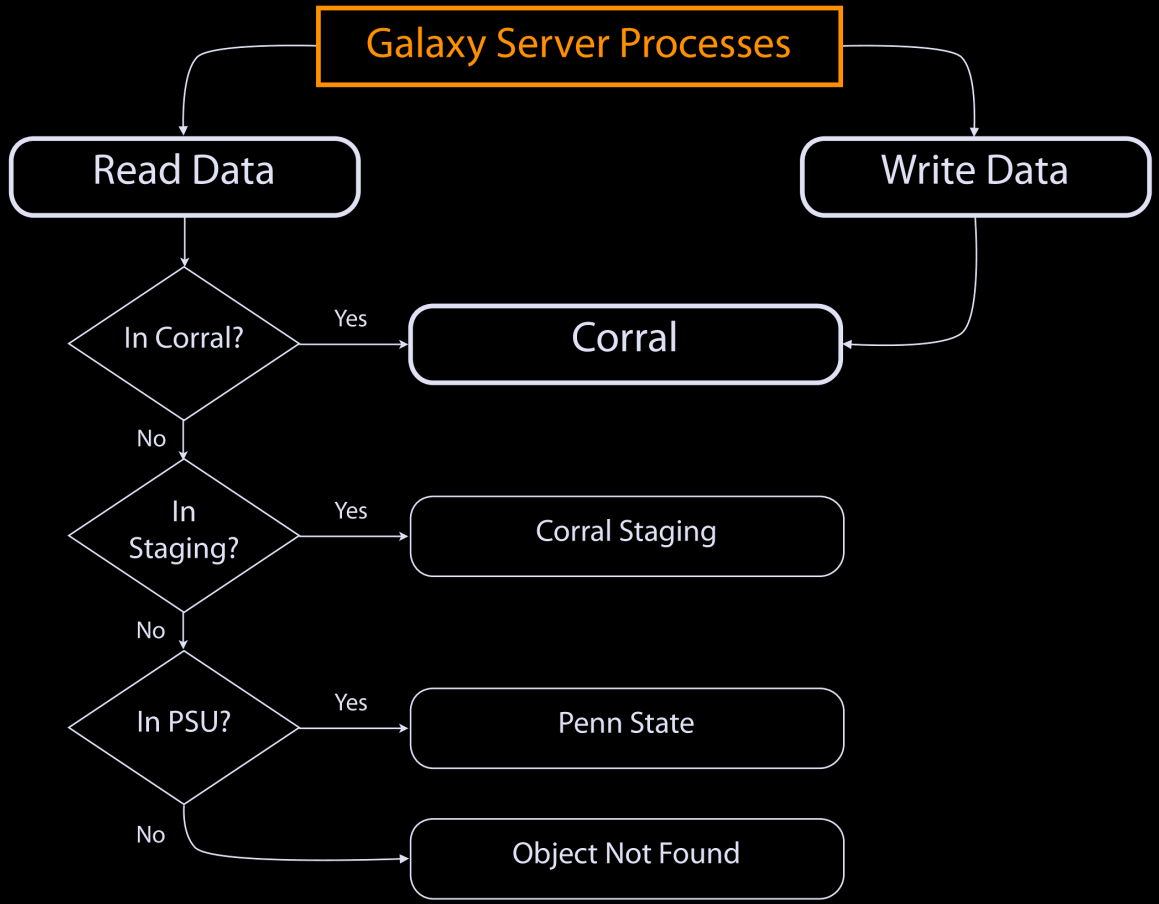
BOSC Video http://bit.ly/bosc2014video

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 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

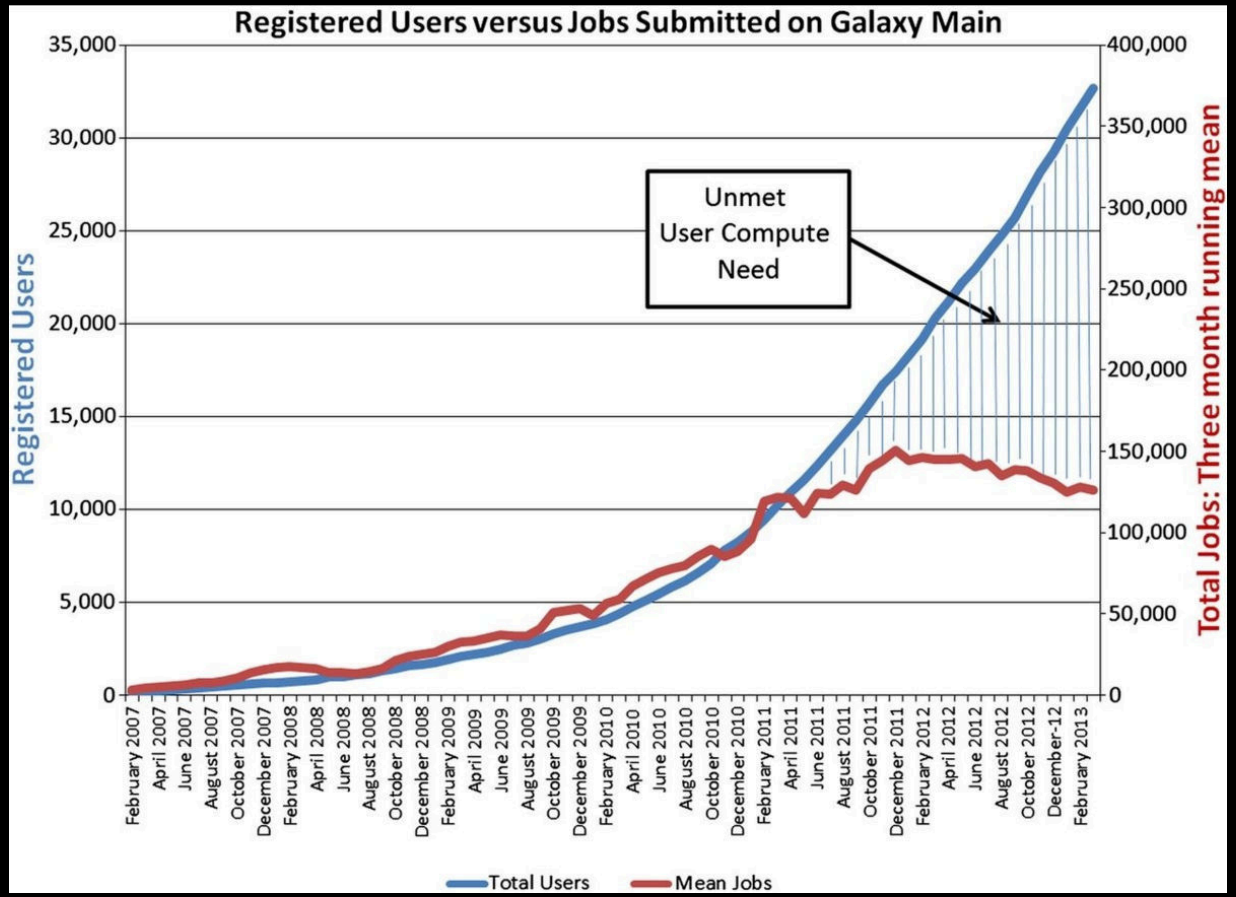
Scaling for Big Demand: usegalaxy.org

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

LeDuc, et al. J Am Med Inform Assoc doi:10.1136/amiajnl-2013-002059



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

bit.ly/gxyServers

Scaling for Big Demand: Cloud

https://wiki.galaxyproject.org/Cloud

Scaling for Big Demand: Support

Scaling for Big Demand: Gather

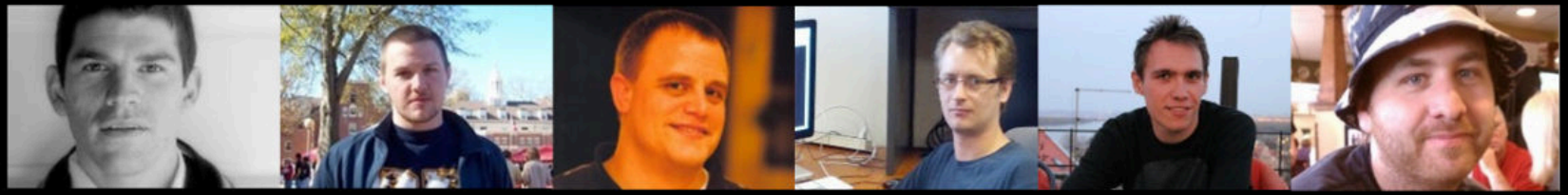
Stop by on your way
to ISMB/ECCB 2015

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

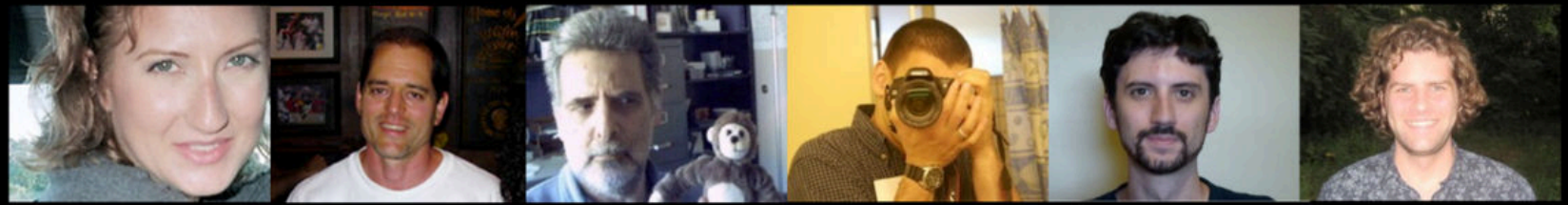
Reply Retweeted Favorite

Galaxy Team



Enis Afgan | Dannon Baker | Dan Blankenberg | Dave Bouvier | Marten Cech | John Chilton

Dave Clements | Nate Coraor | Carl Eberhard | Dorine Francheteau | Jeremy Goecks | Sam Guerler



Jen Jackson | Greg von Kuster | Ross Lazarus | Anton Nekrutenko | Nick Stoler | James Taylor