Galaxy

Data intensive biology *for everyone.*

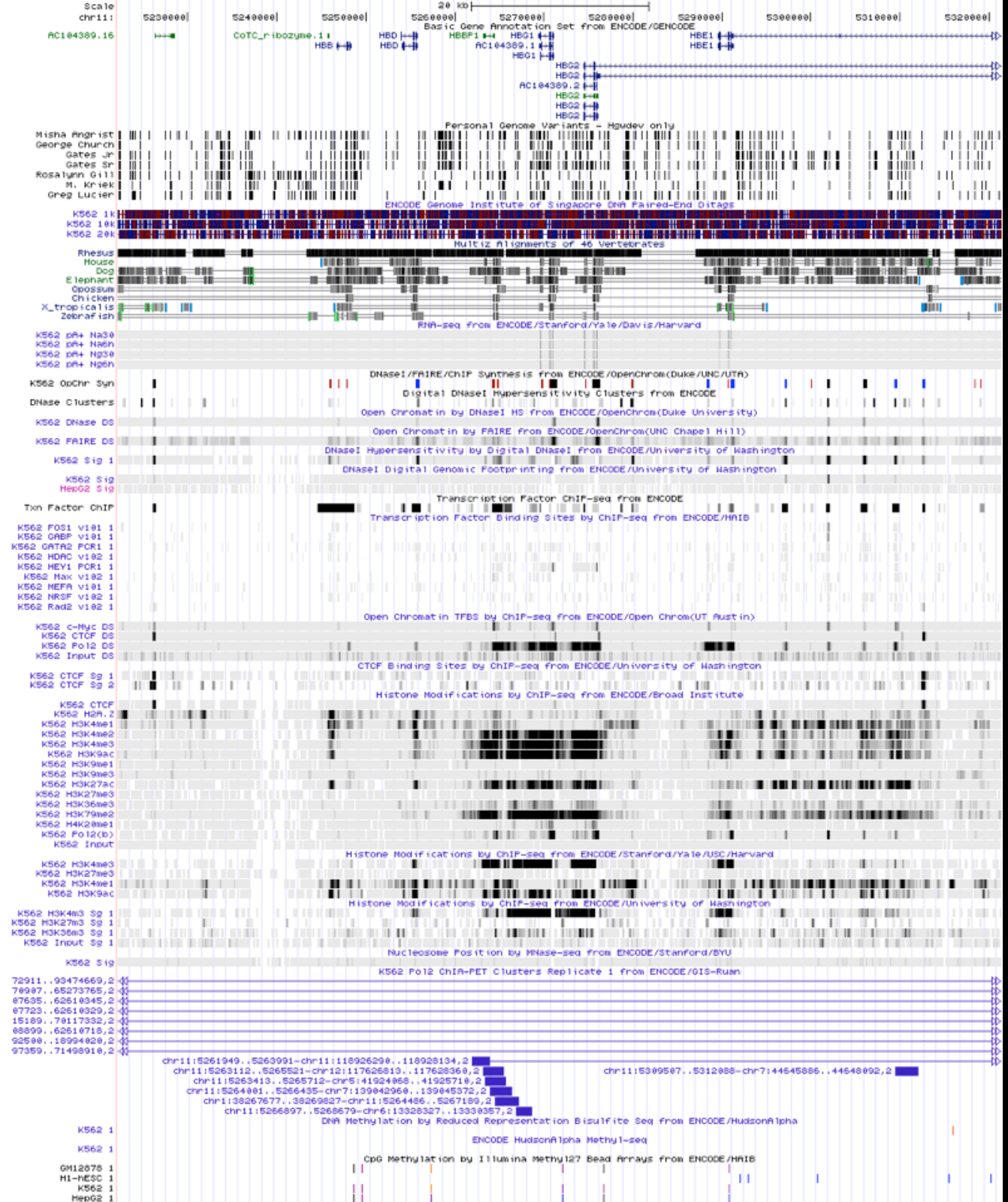
www.galaxyproject.org

@jxtx / #usegalaxy

High-Throughput

I ❤ SEQUENCING!

High-throughput sequencing is
transformative



Resequencing

De novo genome sequencing

Direct RNA sequencing

Open Chromatin assays (DNase, FAIRE)

Transcription factors (ChIP-seq)

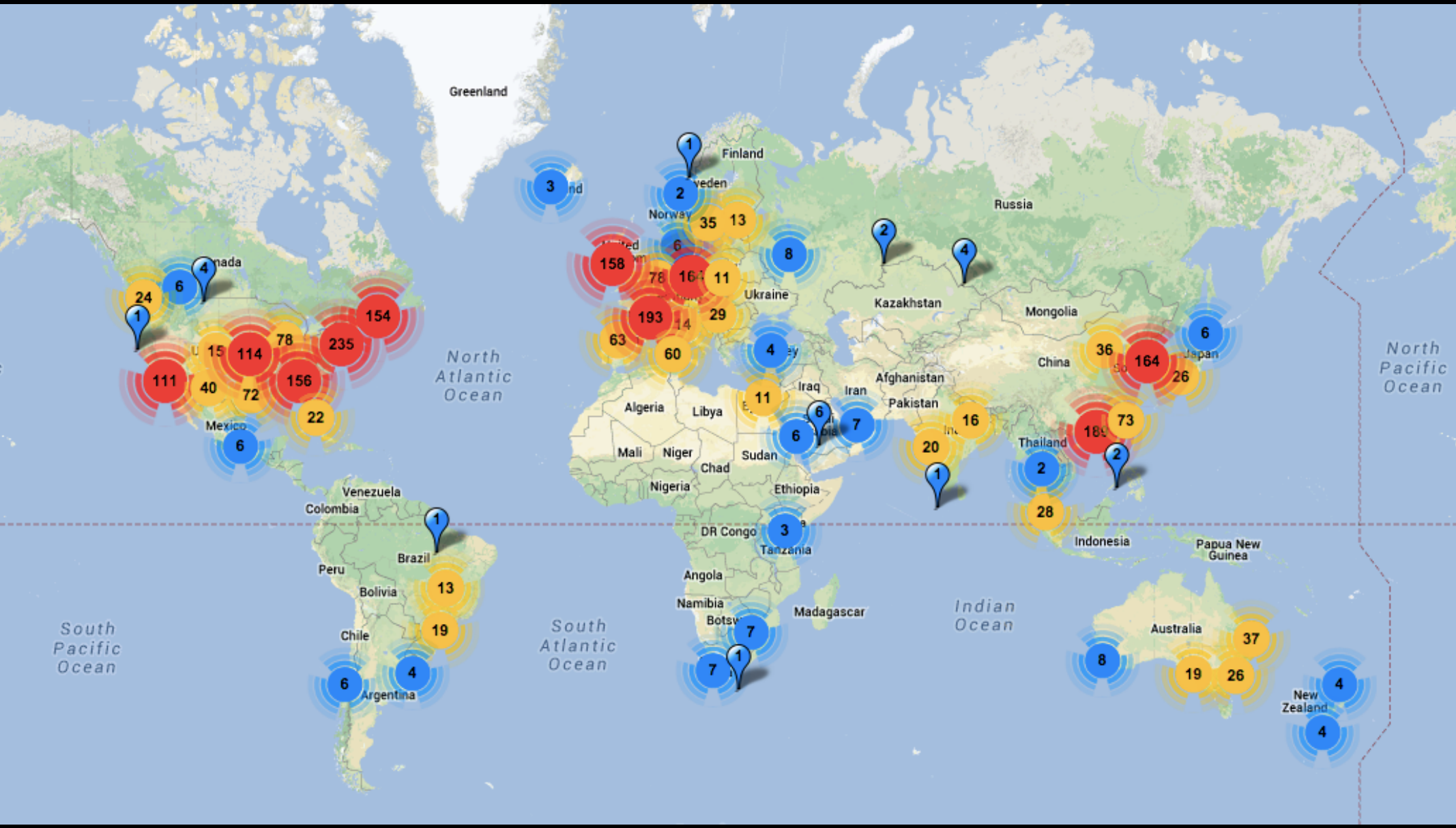
Histones variants (ChIP-seq, MNase-seq)

Long range interactions (5C, Hi-C, ChIA-PET

Methylation (Bisulfite-seq)

High-throughput sequencing is
democratizing

It is widely available...



(http://omicsmaps.com/)

...and practically free!

(NHGRI / *Nature* 497:546–547)

Making sense of this data requires sophisticated methods

How can we ensure that these methods are **accessible** to researchers?

...while also ensuring that scientific results remain **reproducible**?

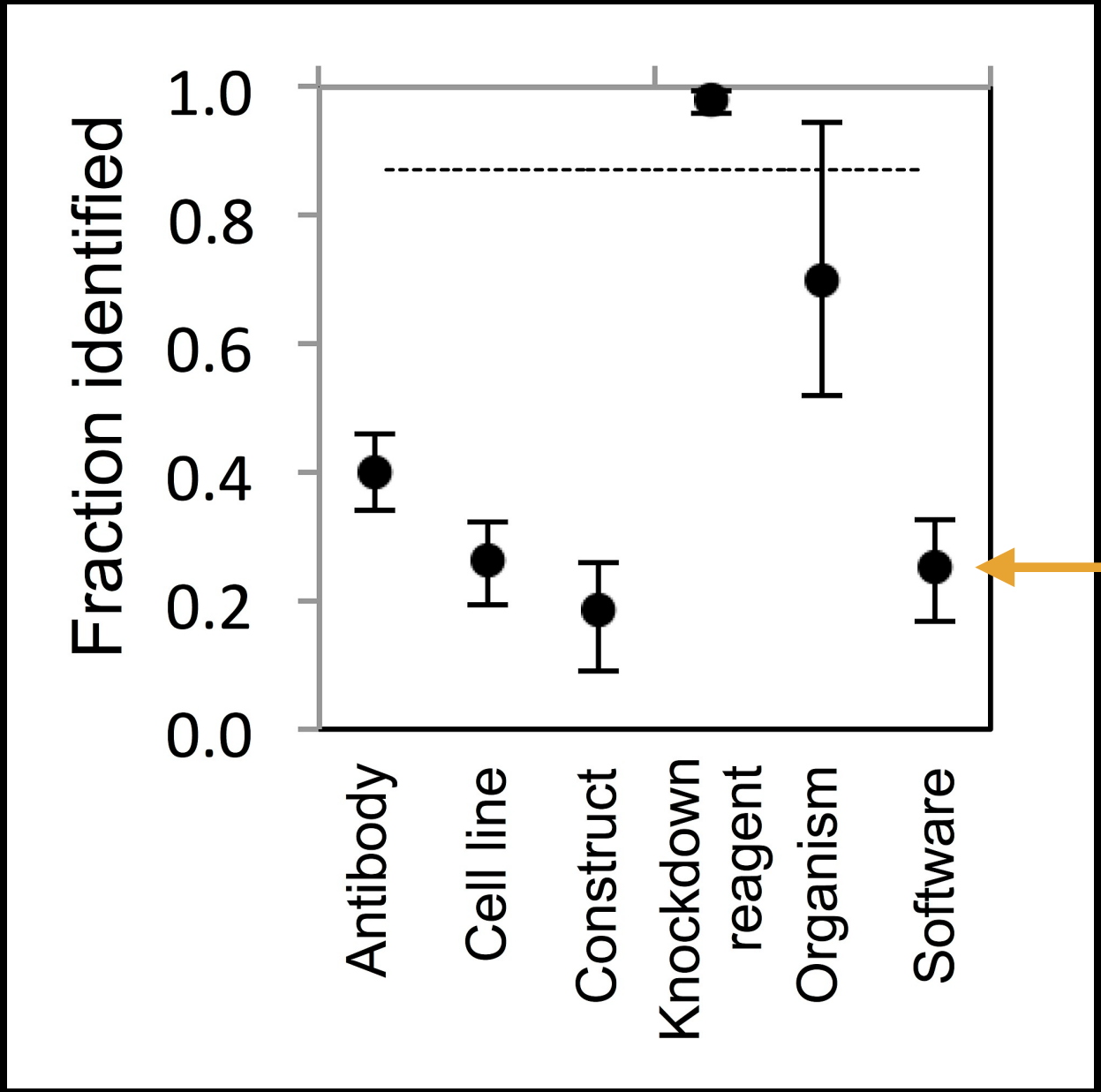
A crisis in genomics research:

reproducibility

Reproducibility Project: Cancer Biology

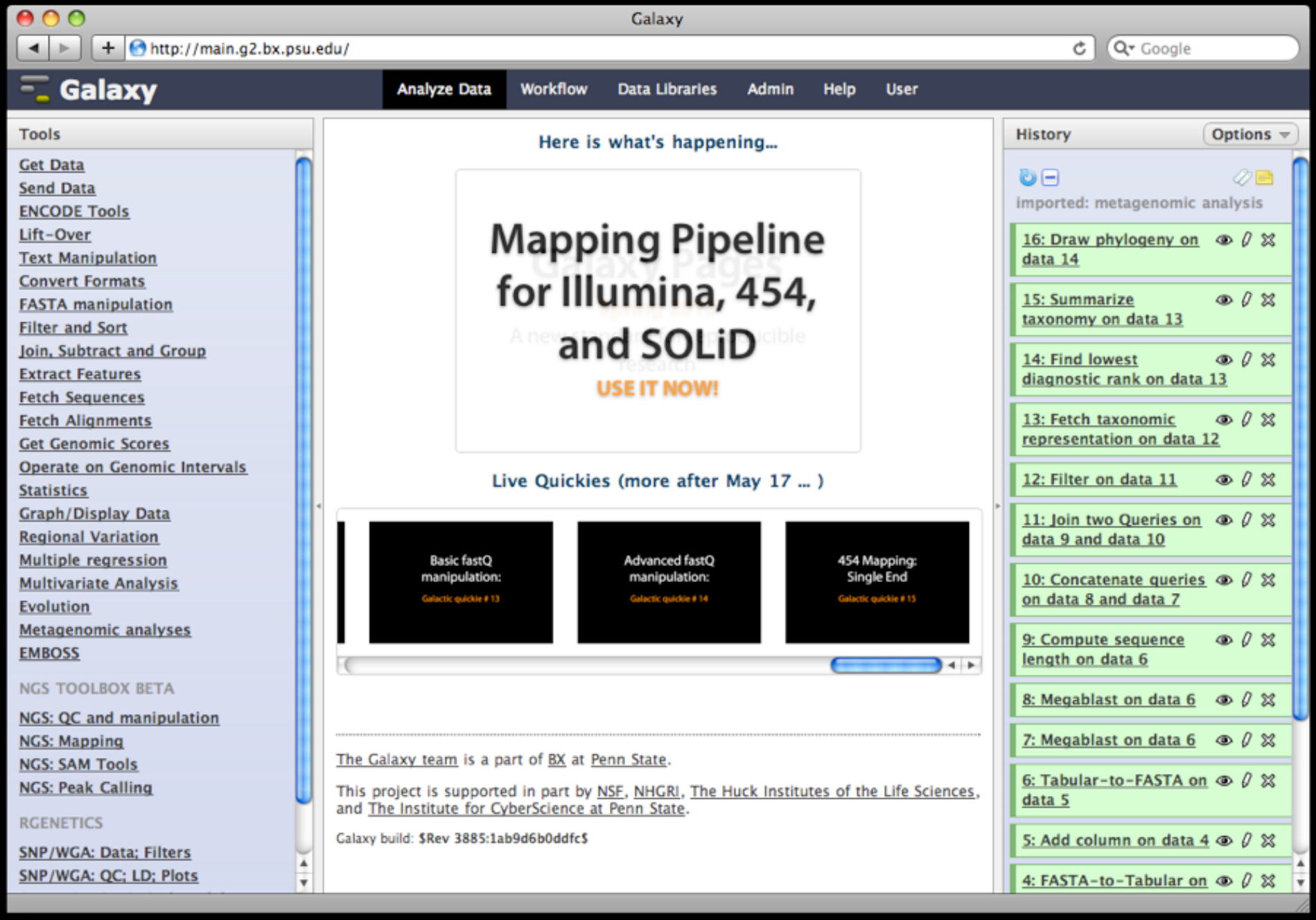
Independently replicating 50 "high-impact" cancer studies from 2010-2012

(https://osf.io/e81xl/wiki/home/)



Vasilevsky, Nicole; Kavanagh, David J; Deusen, Amy Van; Haendel, Melissa; Iorns, Elizabeth (2014): Unique Identification of research resources in studies in Reproducibility Project: Cancer Biology. figshare. http://dx.doi.org/10.6084/m9.figshare.987130

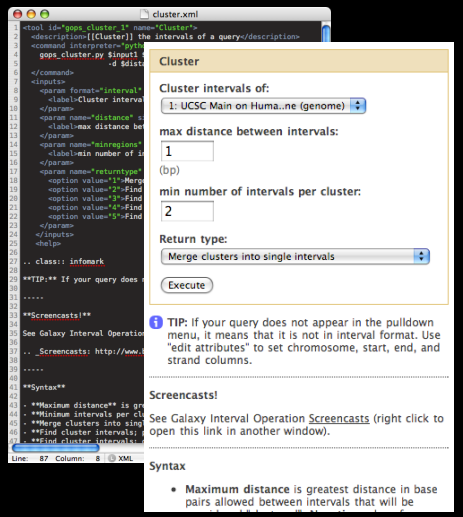
Galaxy: accessible analysis system



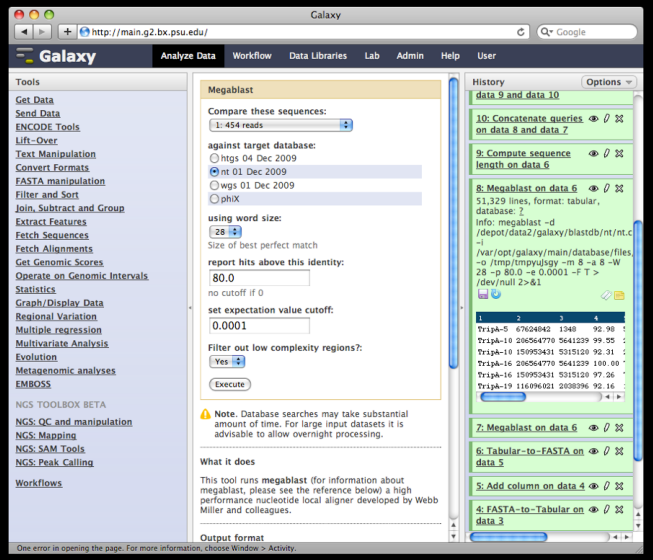
A free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

Open source software that makes integrating your own tools and data and customizing for your own site simple

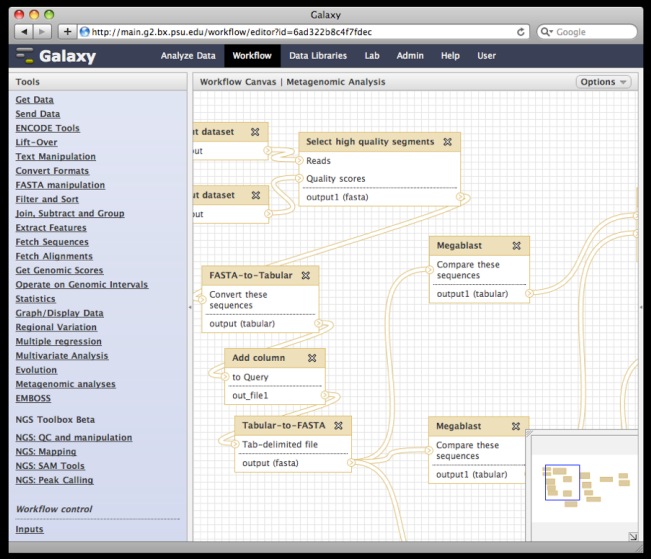
An open extensible platform for sharing tools, datatypes, workflows, ...



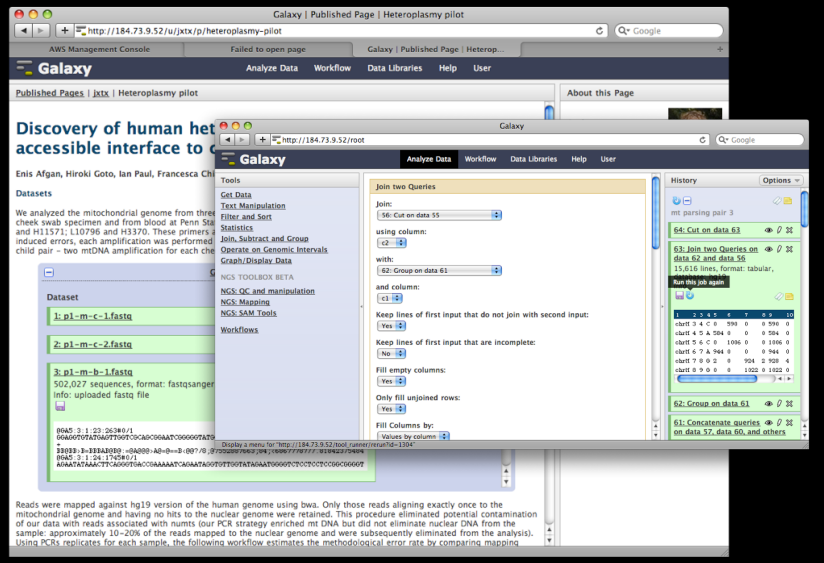
Describe analysis tool behavior abstractly



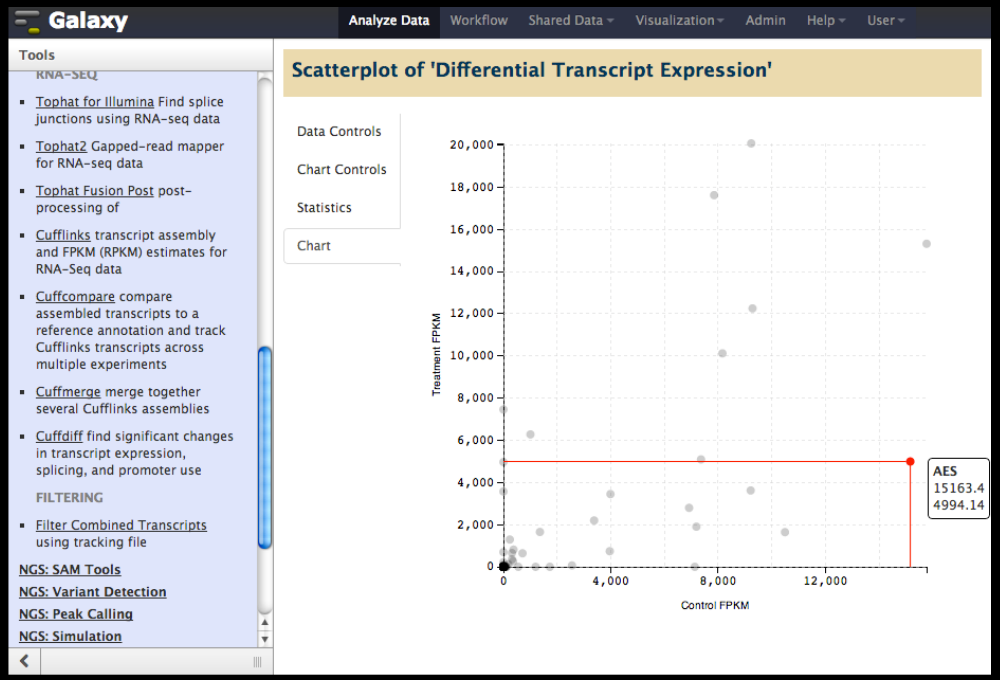
Analysis environment automatically and transparently tracks details

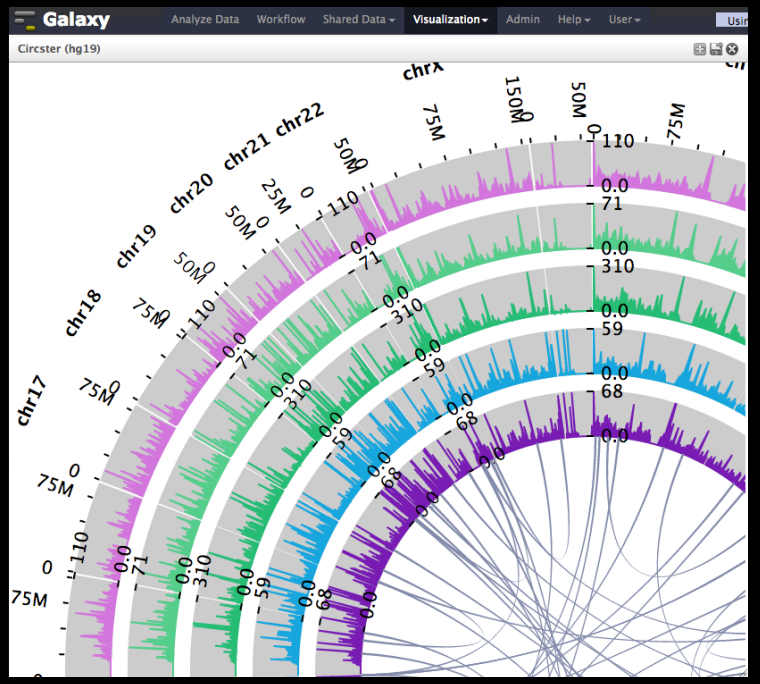


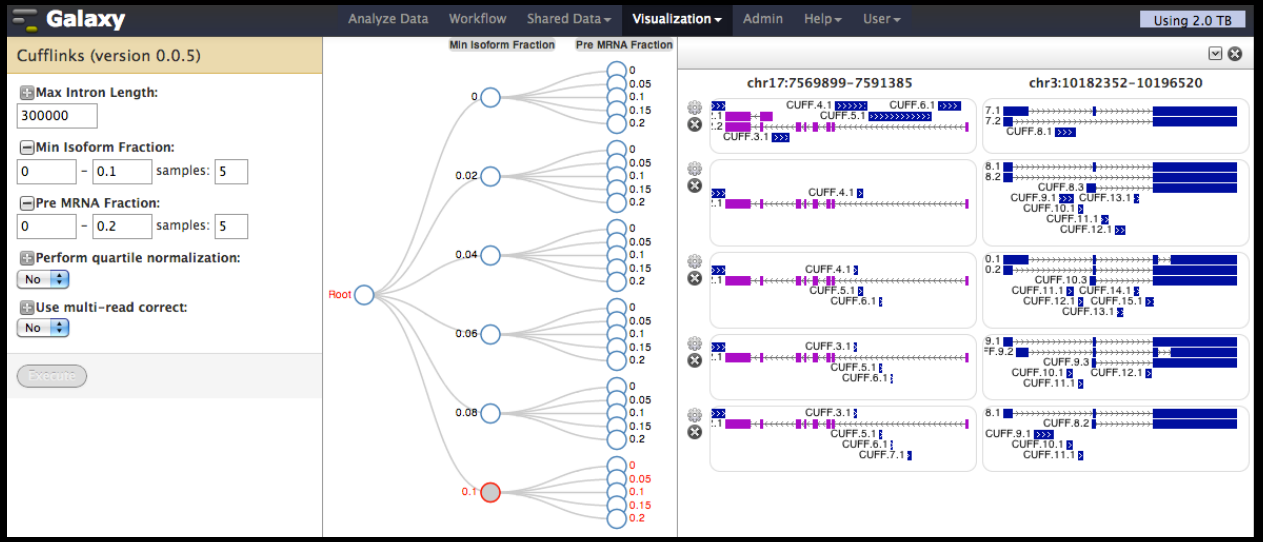
Workflow system for complex analysis, constructed explicitly or automatically



Pervasive sharing, and publication of documents with integrated analysis

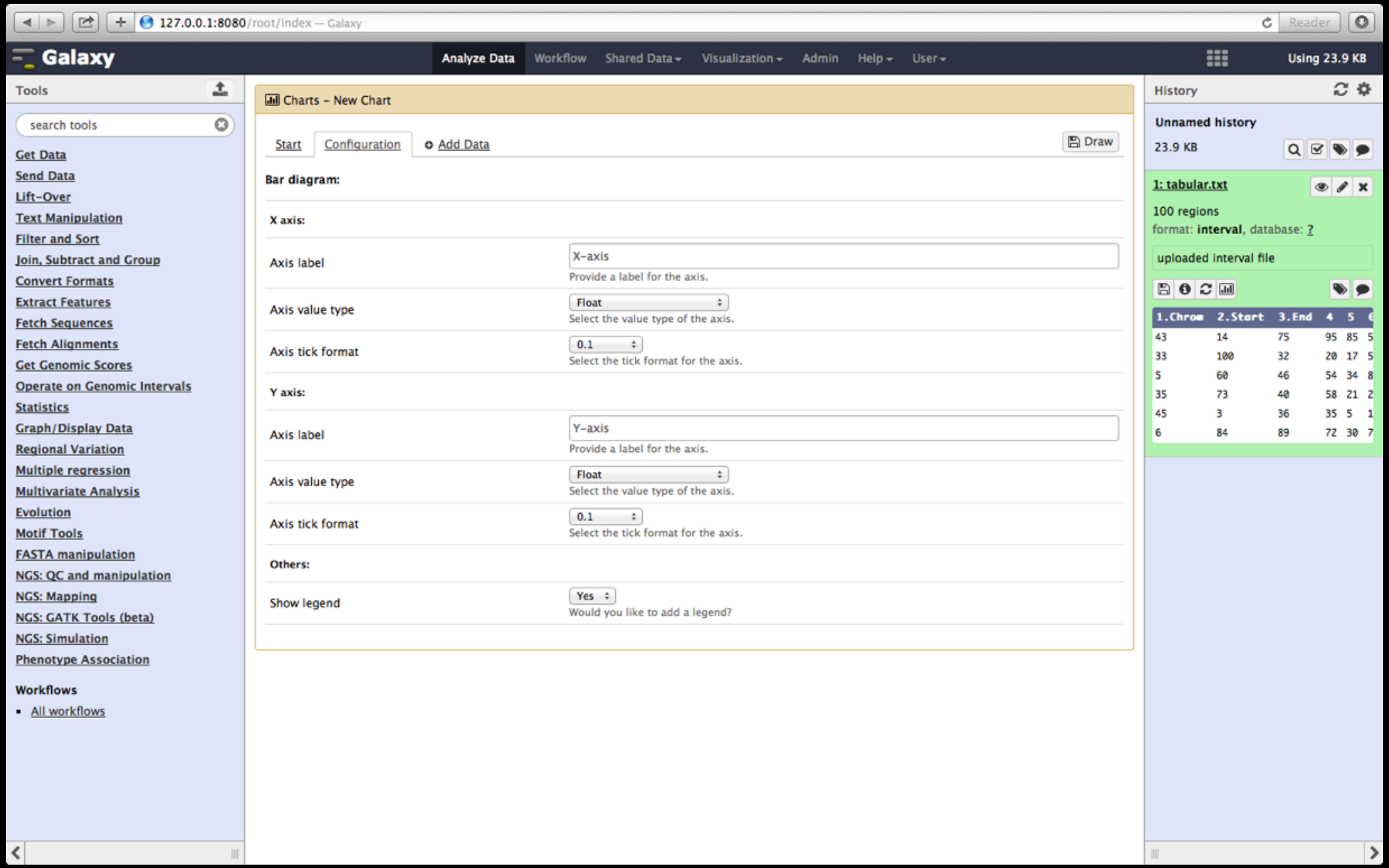




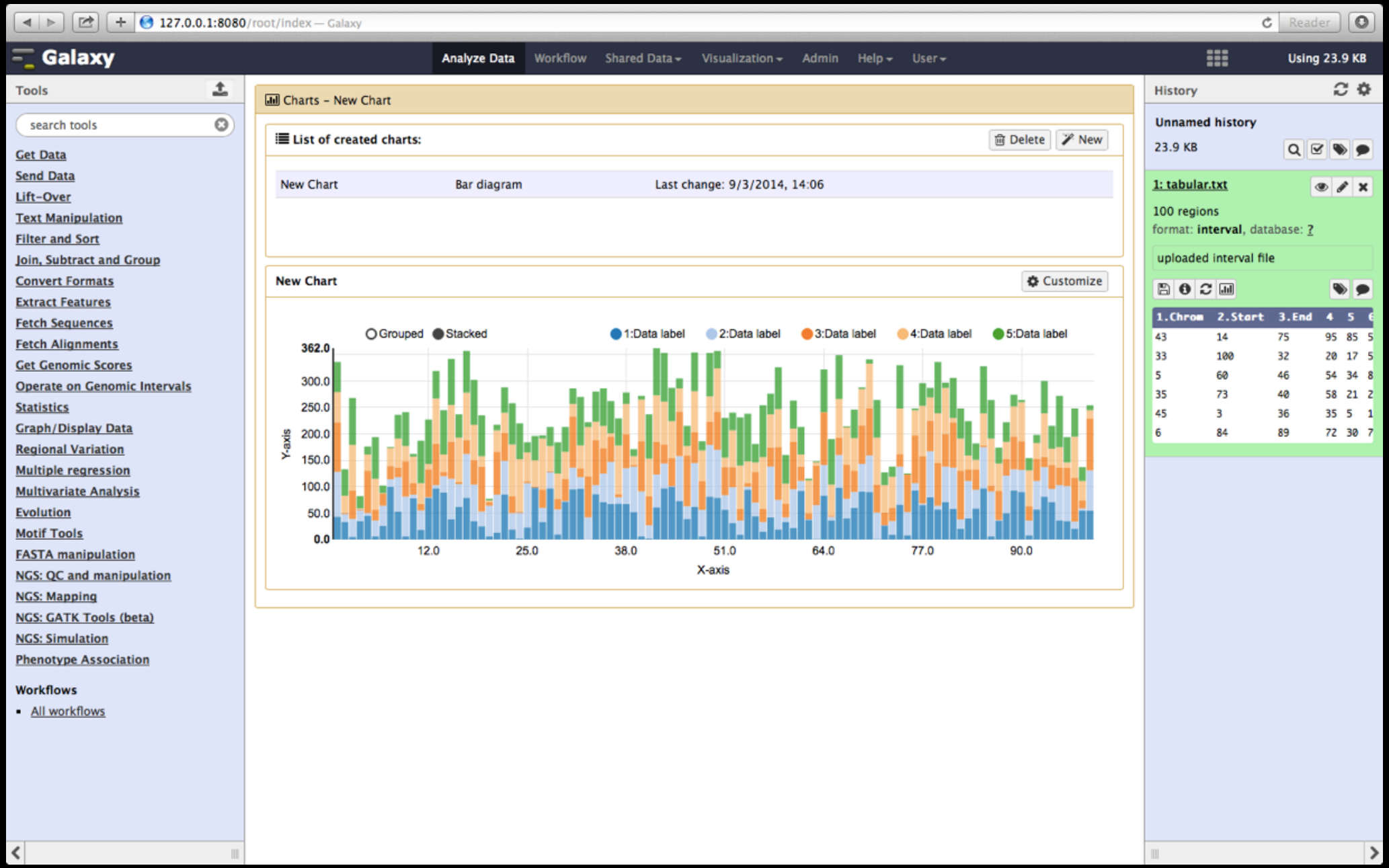


Visualization and visual analytics

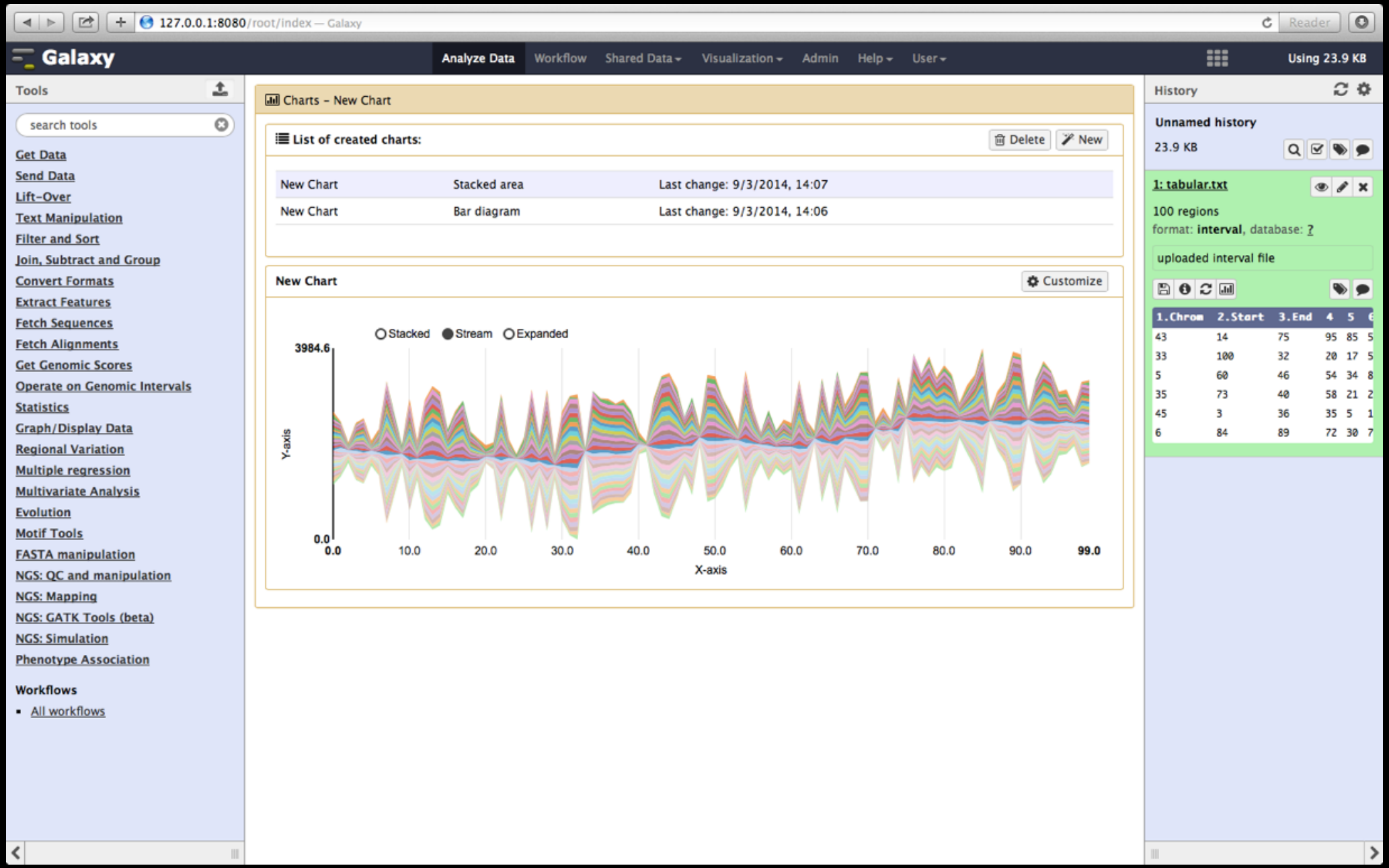
Visualization framework: Charts plugin



Visualization framework: Charts plugin



Visualization framework: Charts plugin



The free service is still the easiest way for users with no informatics infrastructure to analyze their data

How can we possibly sustain this?

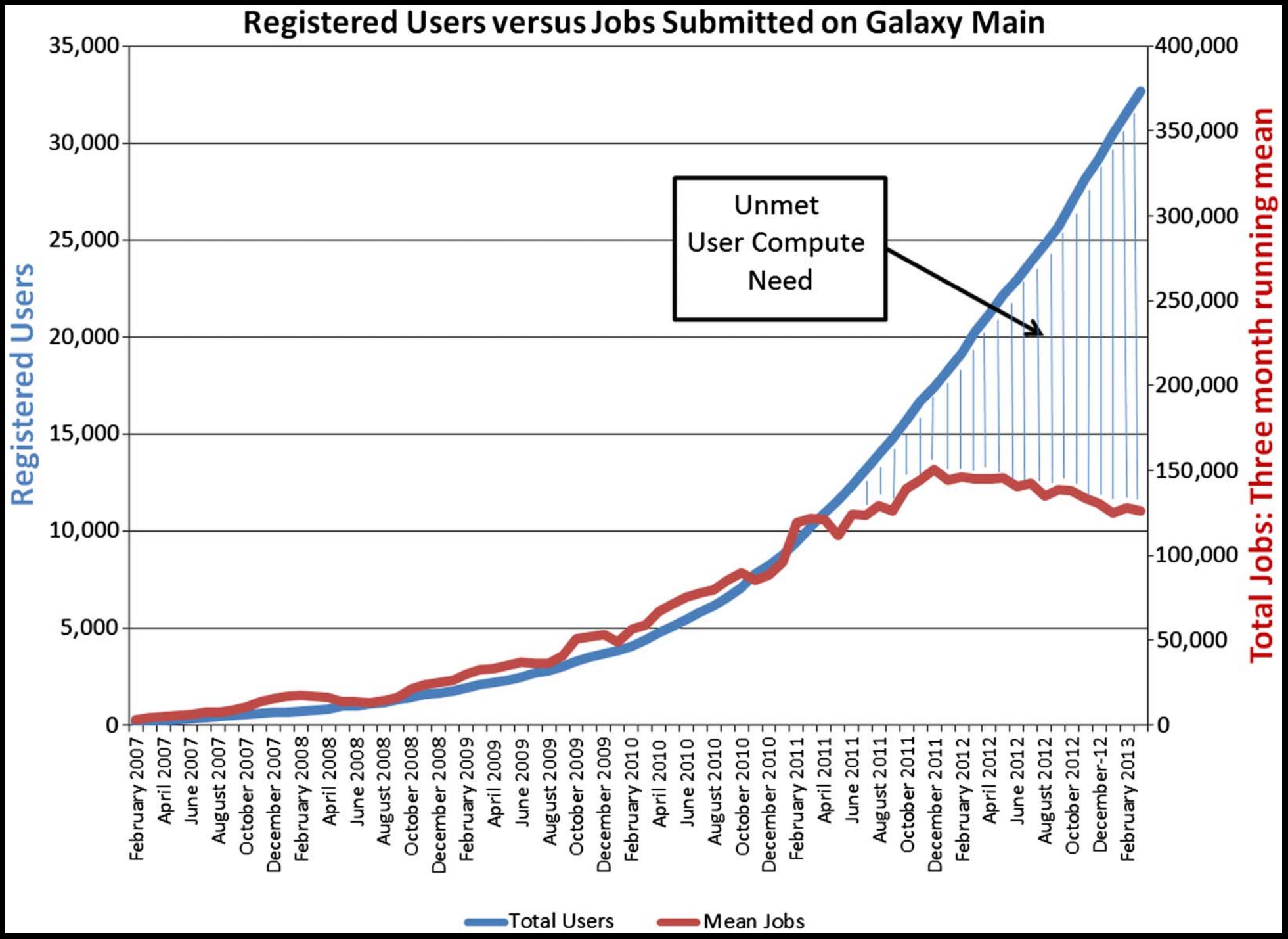
Best place to build this robust entry point is clearly a national supercomputing center

The Texas Advanced Computing Center (TACC) has already built substantial infrastructure in the context of the iPlant project

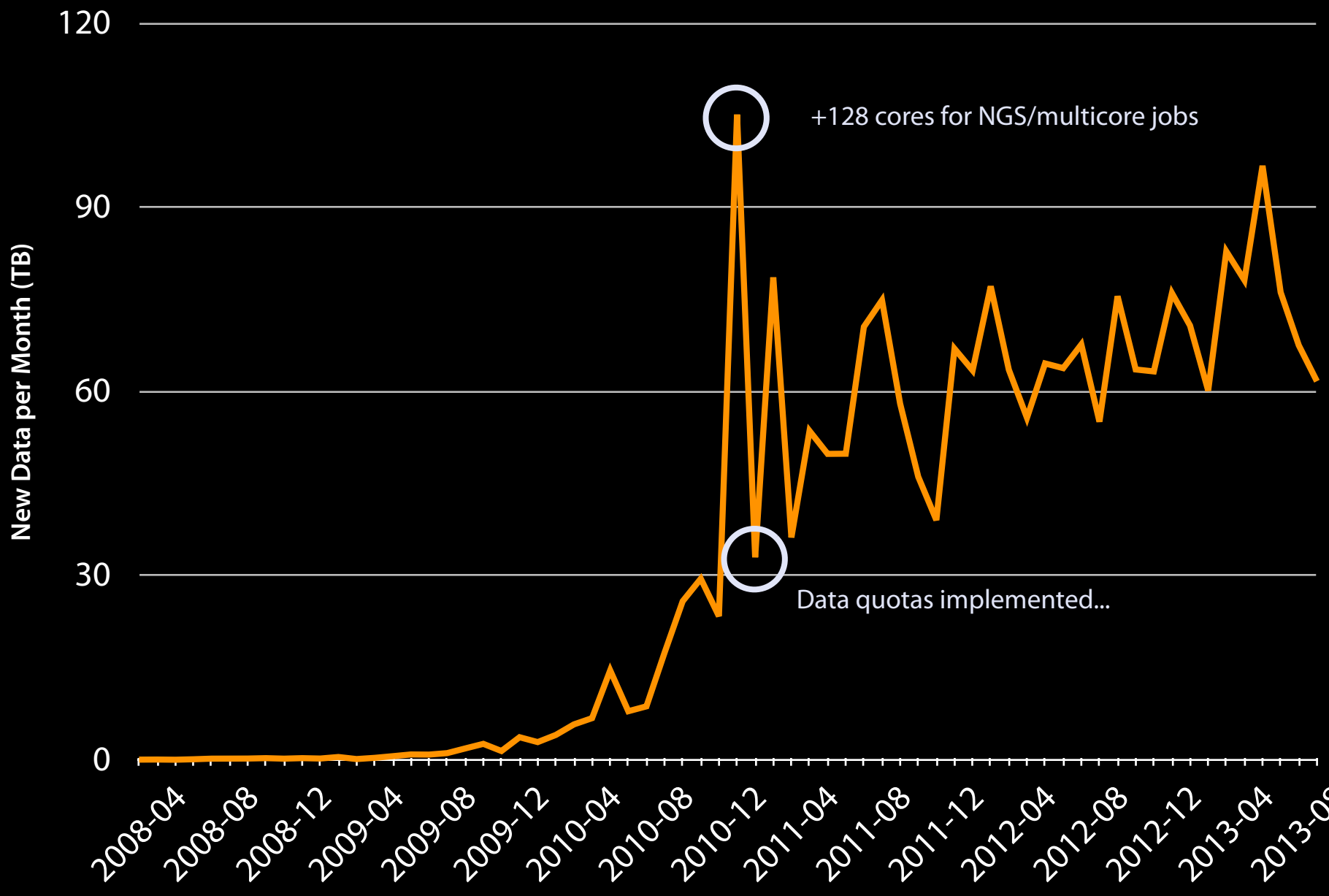
(Including multi petabyte online storage, cloud infrastructure, collocated with some of the worlds largest HPC machines)

However, the iPlant and TACC cyber-infrastructure was underused; thus we established a collaboration

Since October 2013 Galaxy Main has run from TACC

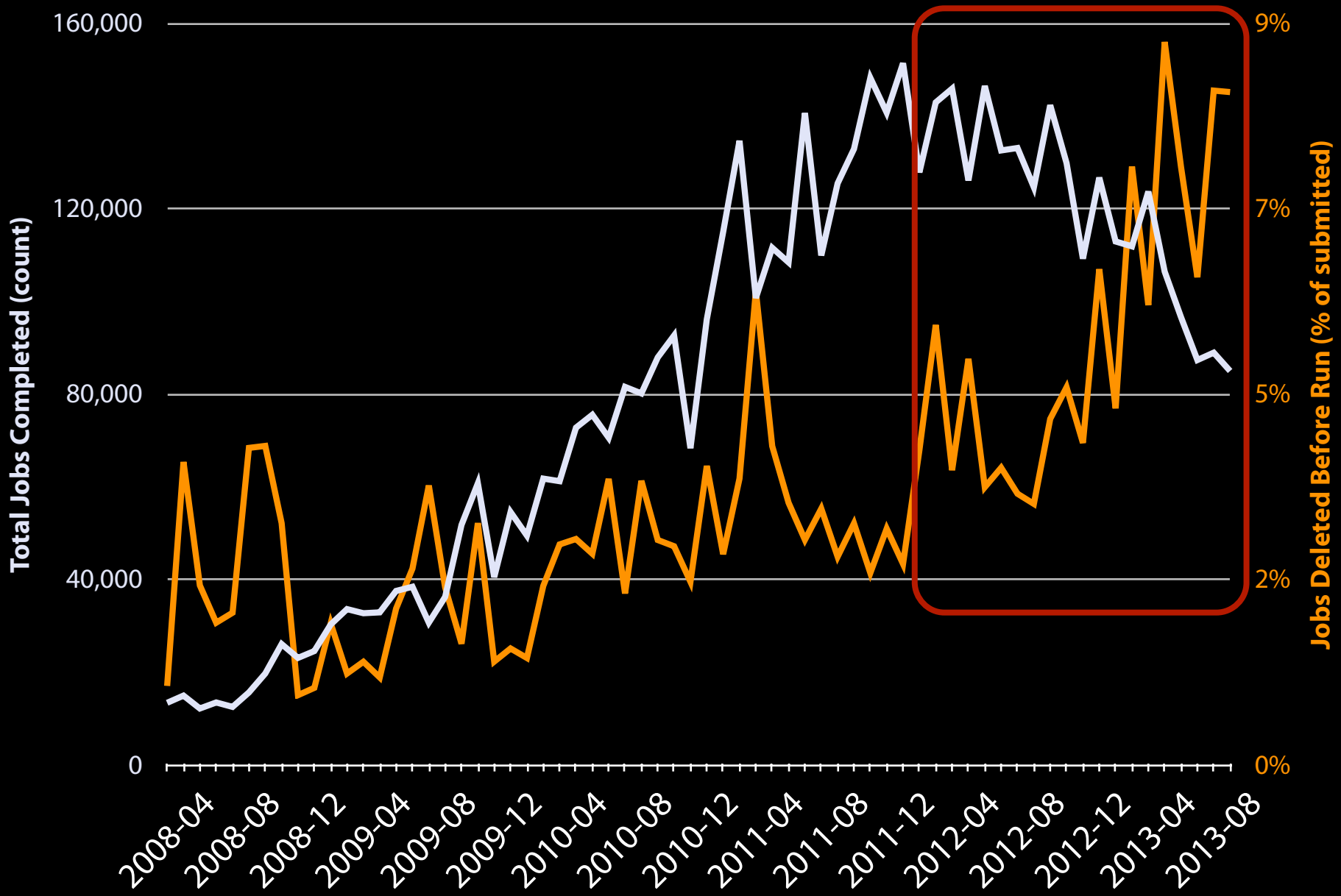

Registered Users versus Jobs Submitted on Galaxy Main
Registered Users
Total Jobs: Three month running mean
Unmet
User Compute
Need
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
February 2007
April 2007
June 2007
August 2007
October 2007
December 2007
February 2008
April 2008
June 2008
August 2008
October 2008
December 2008
February 2009
April 2009
June 2009
August 2009
October 2009
December 2009
February 2010
April 2010
June 2010
August 2010
October 2010
December 2010
February 2011
April 2011
June 2011
August 2011
October 2011
December 2011
February 2012
April 2012
June 2012
August 2012
October 2012
December-12
February 2013
Total Users
Mean Jobs

usegalaxy.org data growth



Nate Coraor

usegalaxy.org frustration growth



Nate Coraor

How can this possibly scale?

1. Leverage exisiting public cyber-infrastructure

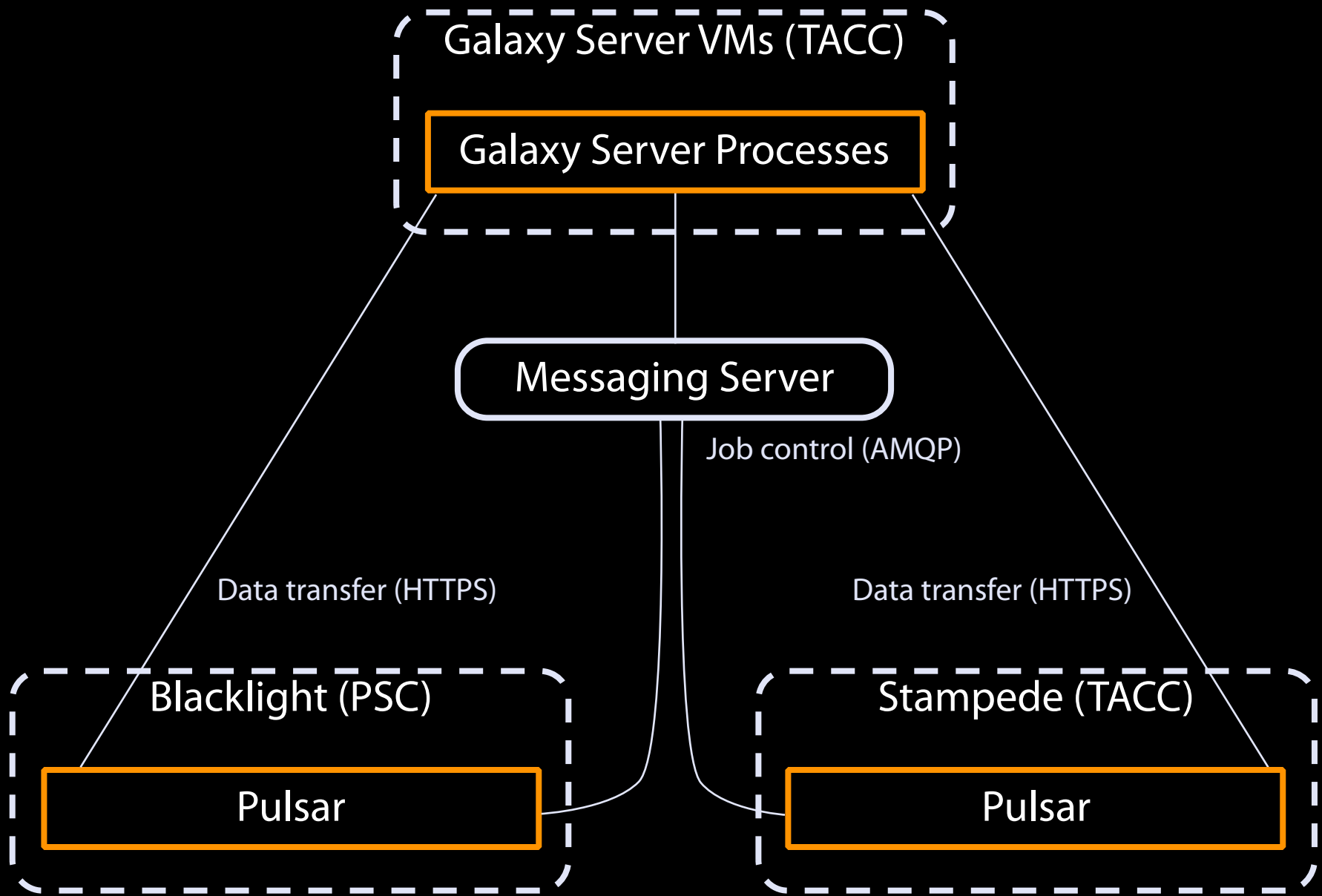
2. Decentralize, provide many deployment models (cloud and local)

Pulsar

Galaxy job runner that can
run almost anywhere

No shared filesystem, stages all necessary
Galaxy components

John Chilton


Galaxy Server VMs (TACC)
Galaxy Server Processes
Messaging Server
Job control (AMQP)
Data transfer (HTTPS)
Data transfer (HTTPS)
Blacklight (PSC)
Pulsar
Stampede (TACC)
Pulsar

Nate Coraor

Some thoughts on the future...

- Scale of analyses is increasing not just in data size, but complexity of workflows and numbers of samples: throughput and reliability for workflows is increasingly important, as well as intuitive user interfaces for processing many samples

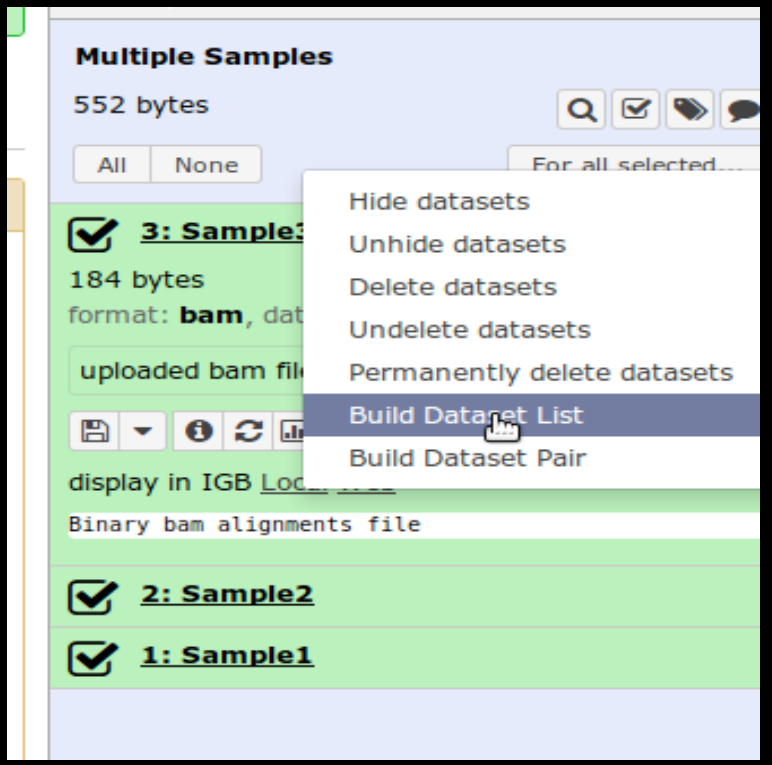
- New computational models and special purpose hardware will almost certainly be more commonly used, how do we best adapt generic tools and workflows to specific execution environments?

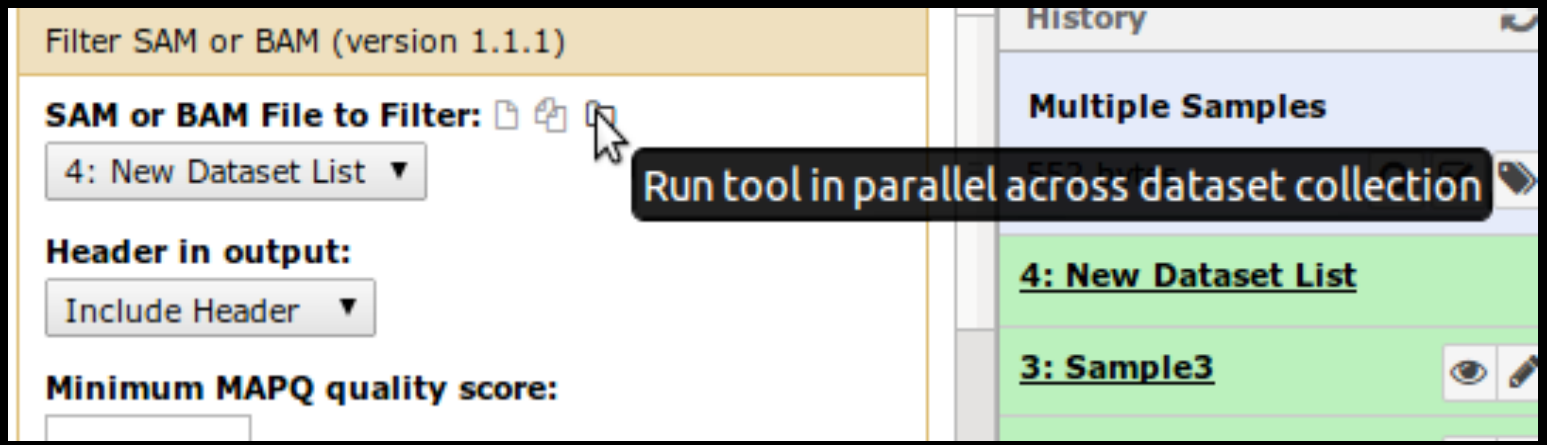
- Infrastructure is only useful when combined with the right incentives, what are the right ways to incentivize reproducible publications and curation of best practice workflows?

- As these workflows emerge, the existence of an accessible open framework facilitates rapid translation from research to clinical application, what are the right summaries and visualizations for this environment?

Dataset collections

- Group datasets into collections.
- Sample identifier tracked through complex workflows.
- New tool parameter types for consuming collections.
- Extensible plugin framework for defining types of collections.

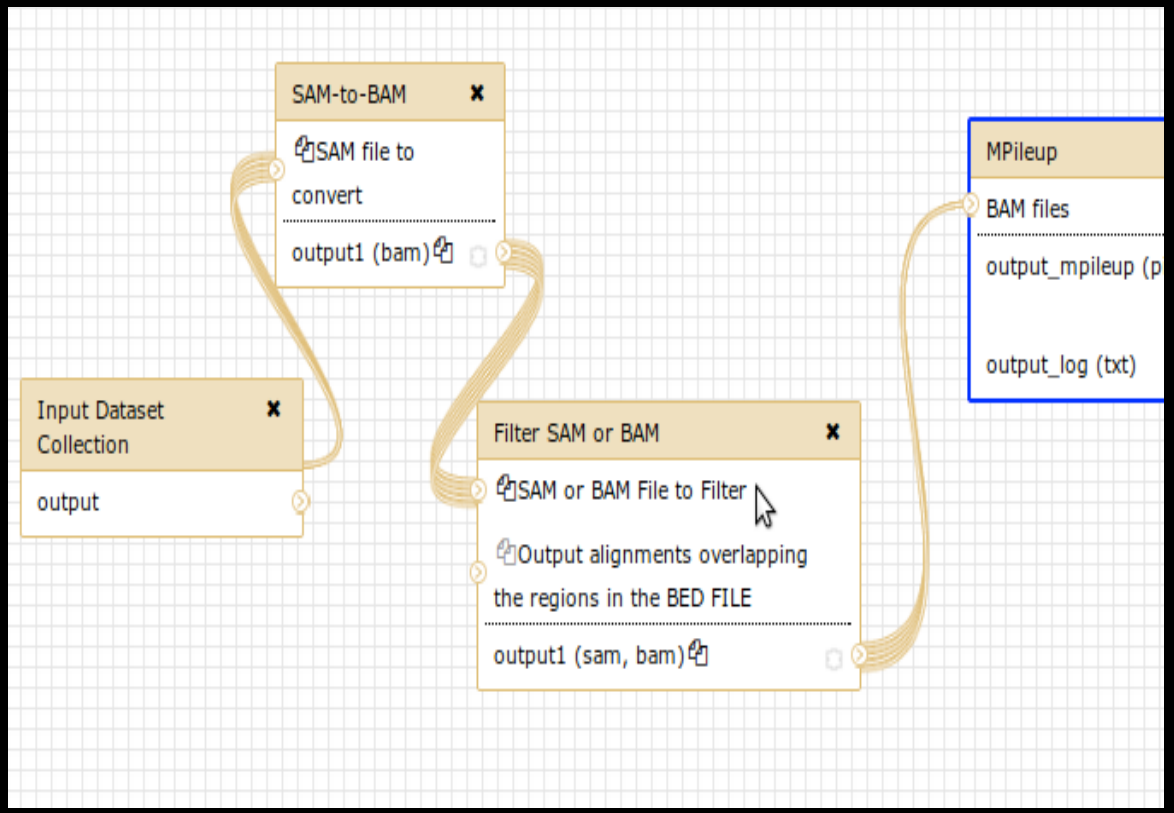




Map and Reduce workflows

not just useful for new tools!

- **Map** - run multiple instances of nearly all existing tools across collections in parallel.
- **Reduce** - steps leveraging existing tool constructs.
- Extract such steps from histories into workflows - or build these workflows from scratch.
- Very common analysis pattern across many kinds of '*omics*'.

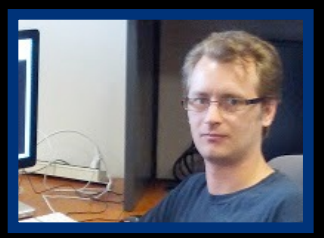


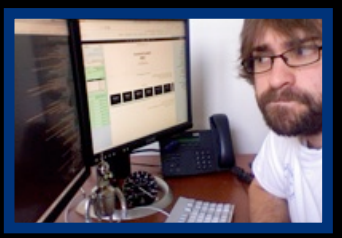
Engineering


Enis Afgan


Dannon Baker


Dan Blankenberg

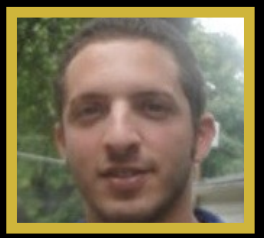

Dave Bouvier

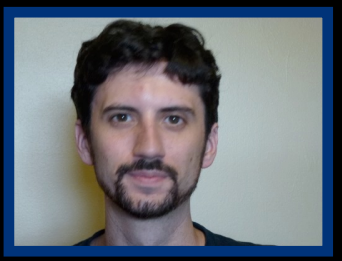

Nate Coraor


Martin Čech


John Chilton


Carl Eberhard


Sam Guerler


Nick Stoler

Support and outreach


Dave Clements


Jennifer Jackson

Leadership


James Taylor


Anton Nekrutenko


Jeremy Goecks

Supported by the **NHGRI** (HG005542, HG004909, HG005133, HG006620), **NSF** (DBI-0850103), Penn State University, Johns Hopkins University, and the Pennsylvania Department of Public Health