

# Galaxy

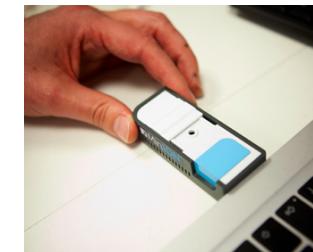
Data intensive biology *for everyone.*

[www.galaxyproject.org](http://www.galaxyproject.org)

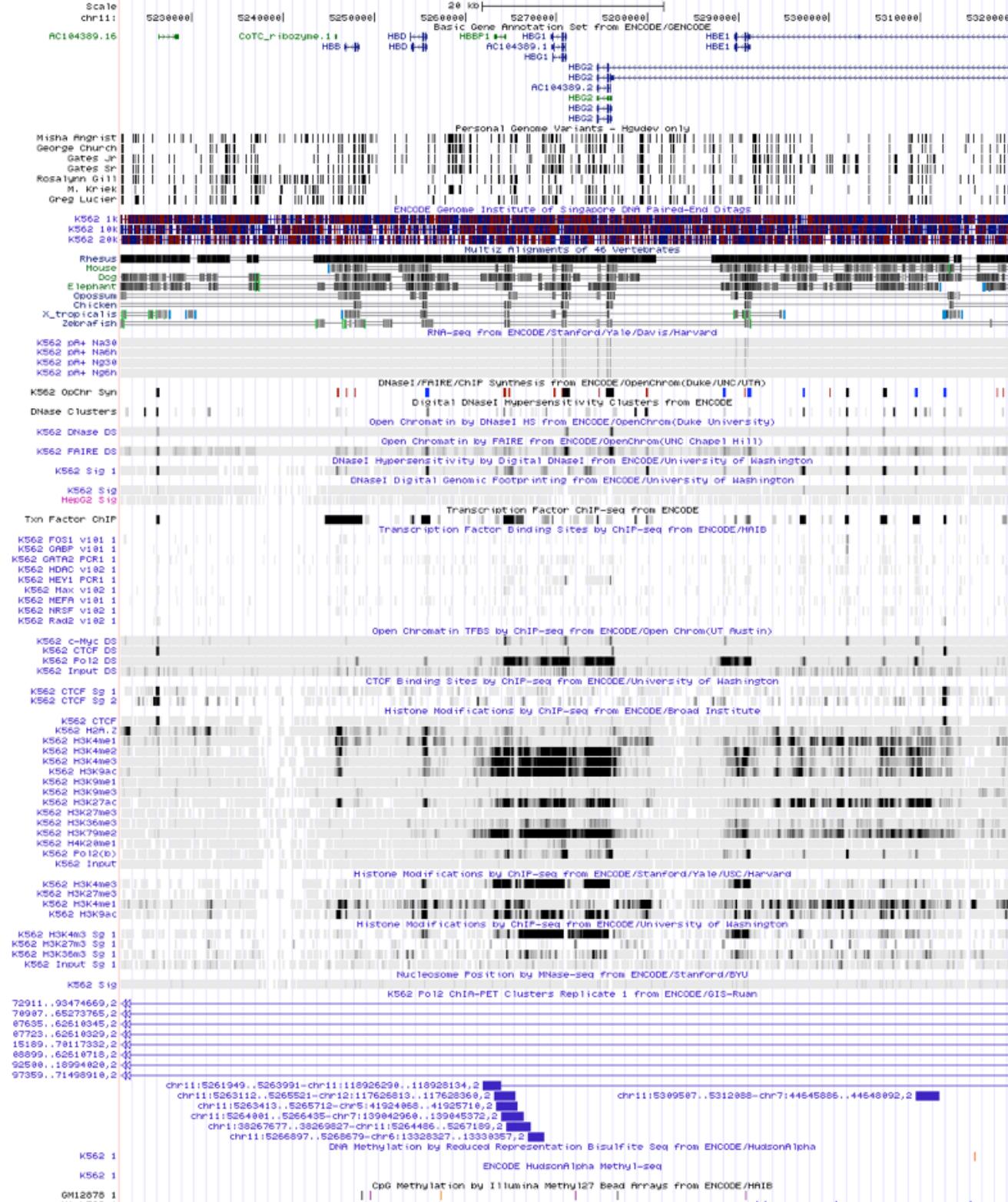


@jxtx / #usegalaxy

High-Throughput  
v  
**I ❤ SEQUENCING!**



High-throughput sequencing is  
**transformative**



# Resequencing

# De novo genome sequencing

# Direct RNA sequencing

# Open Chromatin assays (DNase, FAIRE)

# Transcription factors (ChIP-seq)

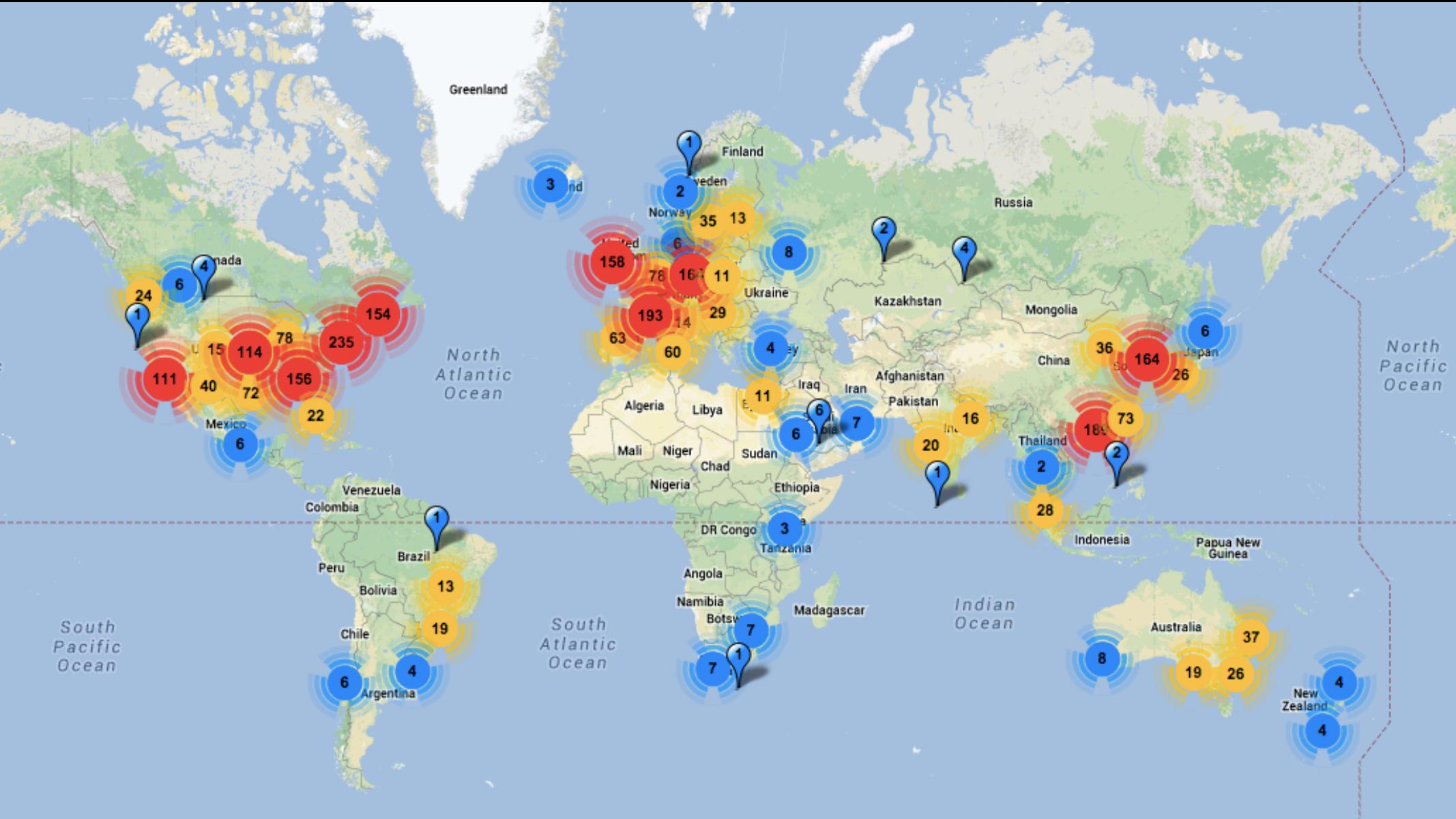
## Histones variants (ChIP-seq, MNase-seq)

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

# Methylation (Bisulfite-seq)

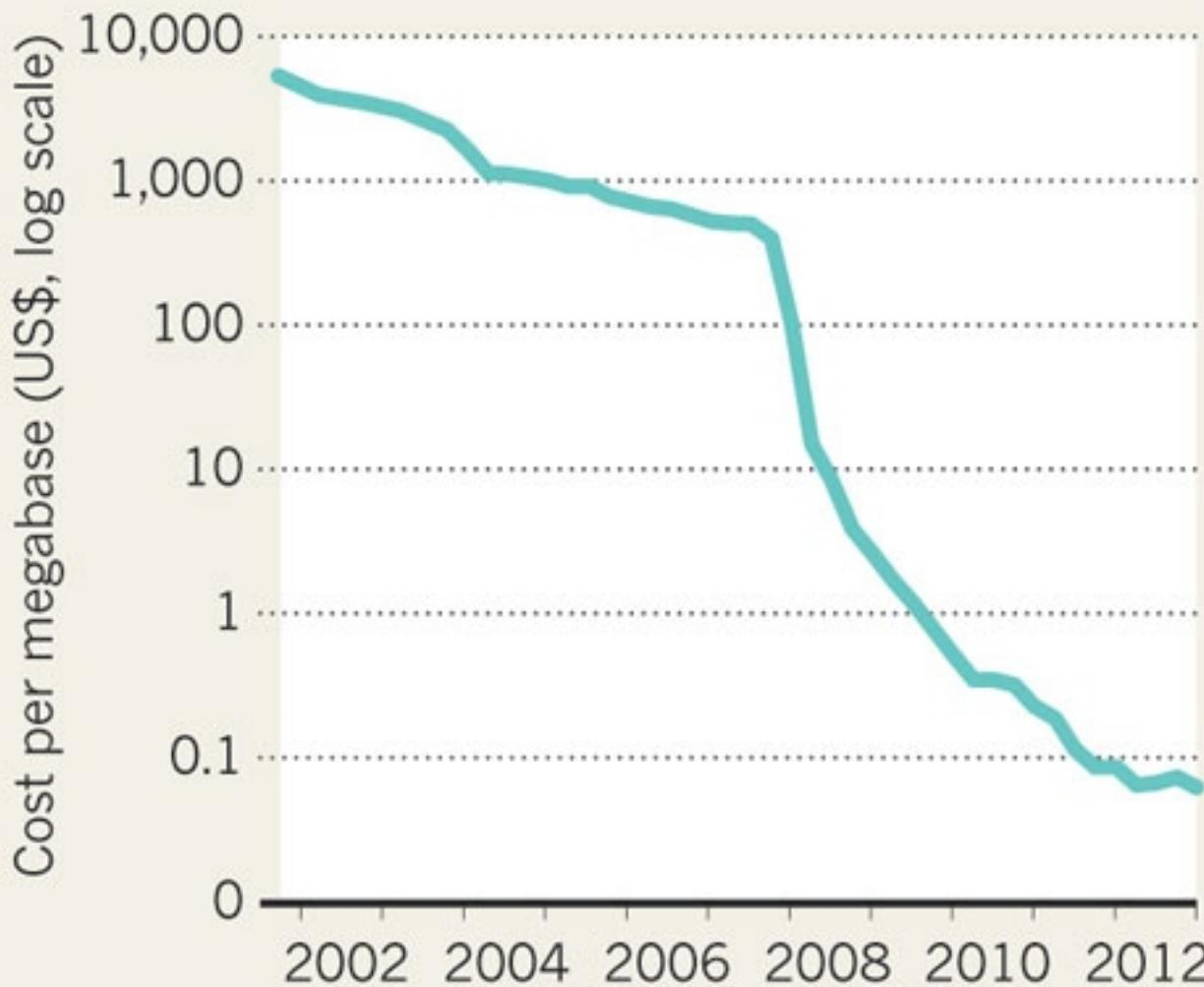
High-throughput sequencing is  
**democratizing**

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?

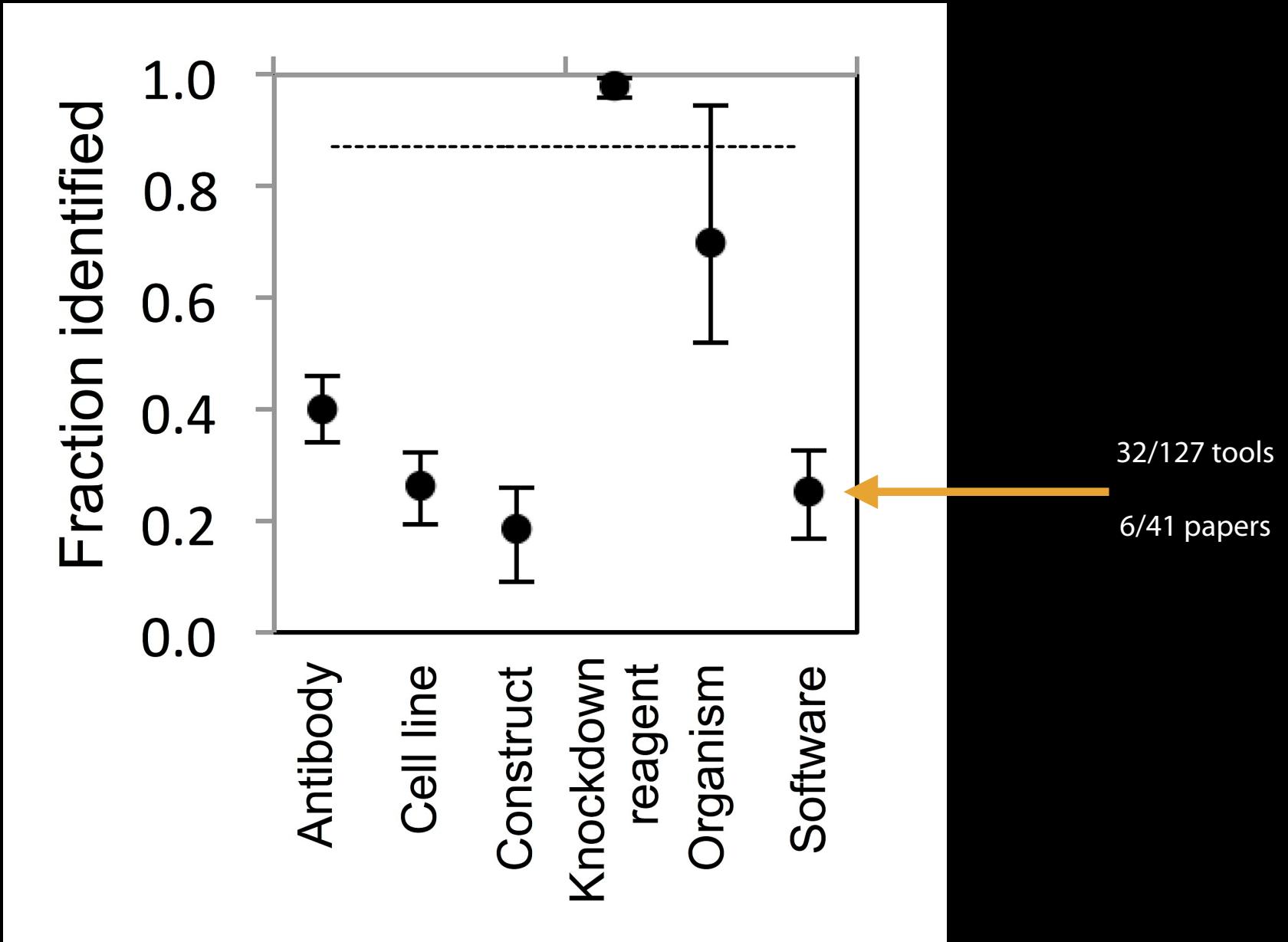
...while also ensuring that scientific results  
remain **reproducible**?

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

Galaxy

http://main.g2.bx.psu.edu/ Google

Galaxy

Analyze Data Workflow Data Libraries Admin Help User

Tools

- Get Data
- Send Data
- ENCODE Tools
- 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
- Metagenomic analyses
- EMBOSS

NGS TOOLBOX BETA

- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Peak Calling

RGENETICS

- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots

Here is what's happening...

Mapping Pipeline for Illumina, 454, and SOLiD

A new pipeline is now available for Illumina, 454, and SOLiD sequencing platforms.

USE IT NOW!

Live Quickies (more after May 17 ...)

Basic fastQ manipulation: Galactic quickie # 13

Advanced fastQ manipulation: Galactic quickie # 14

454 Mapping: Single End Galactic quickie # 15

The Galaxy team is a part of BX at Penn State.

This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

Galaxy build: \$Rev 3885:1ab9d6b0ddfc\$

History Options

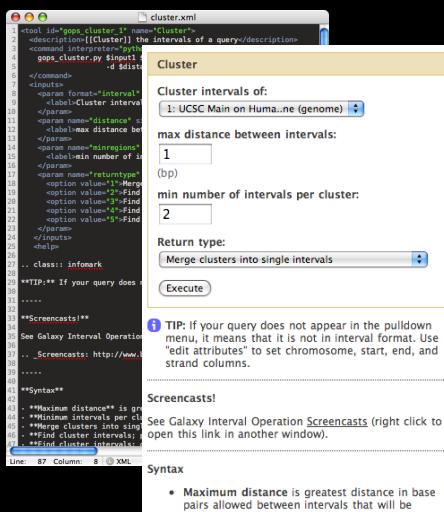
imported: metagenomic analysis

- 16: Draw phylogeny on data 14
- 15: Summarize taxonomy on data 13
- 14: Find lowest diagnostic rank on data 13
- 13: Fetch taxonomic representation on data 12
- 12: Filter on data 11
- 11: Join two Queries on data 9 and data 10
- 10: Concatenate queries on data 8 and data 7
- 9: Compute sequence length on data 6
- 8: Megablast on data 6
- 7: Megablast on data 6
- 6: Tabular-to-FASTA on data 5
- 5: Add column on data 4
- 4: FASTA-to-Tabular on

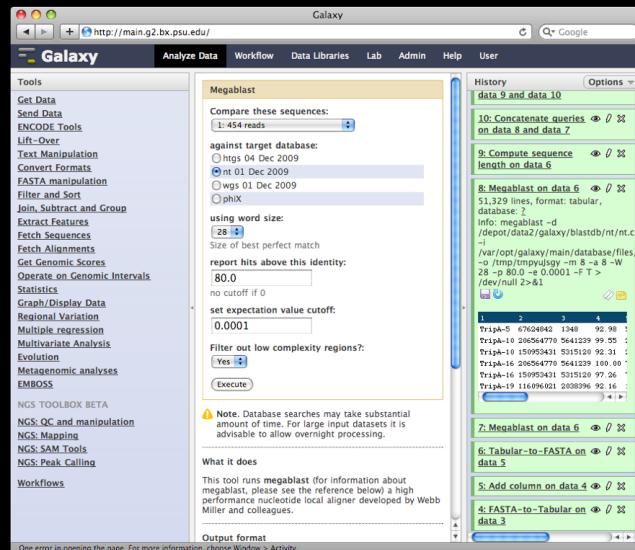
**A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

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

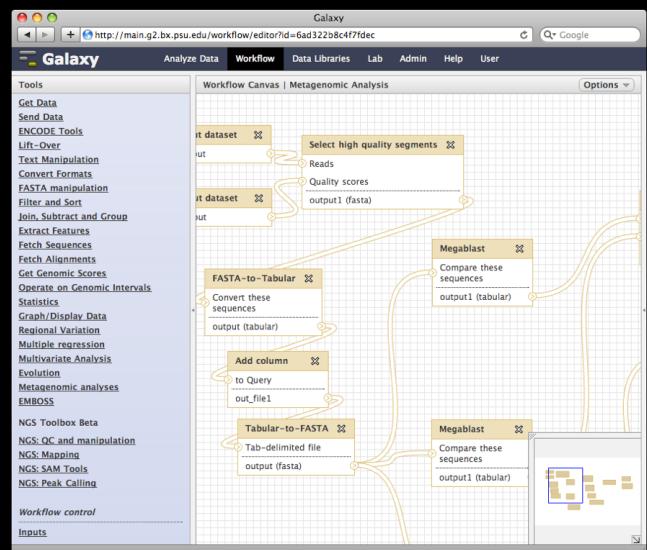
**An open extensible platform** for sharing tools, datatypes, workflows, ...



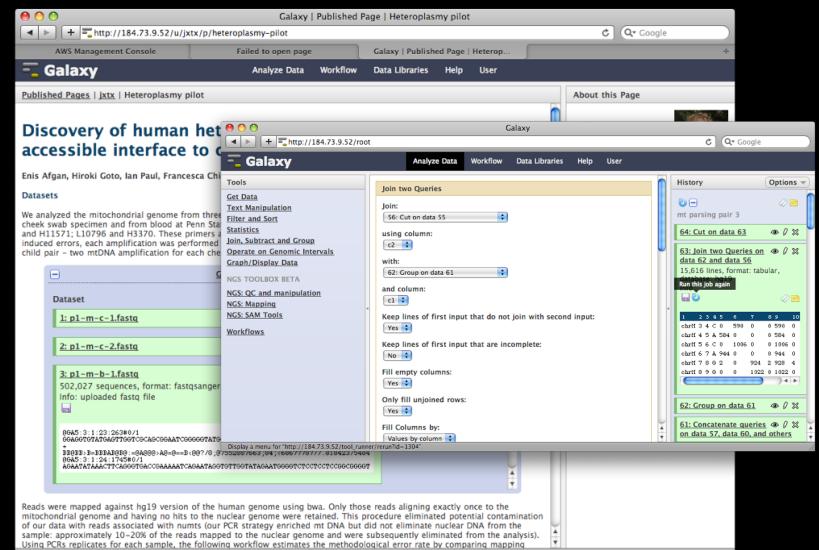
# Describe analysis tool behavior abstractly



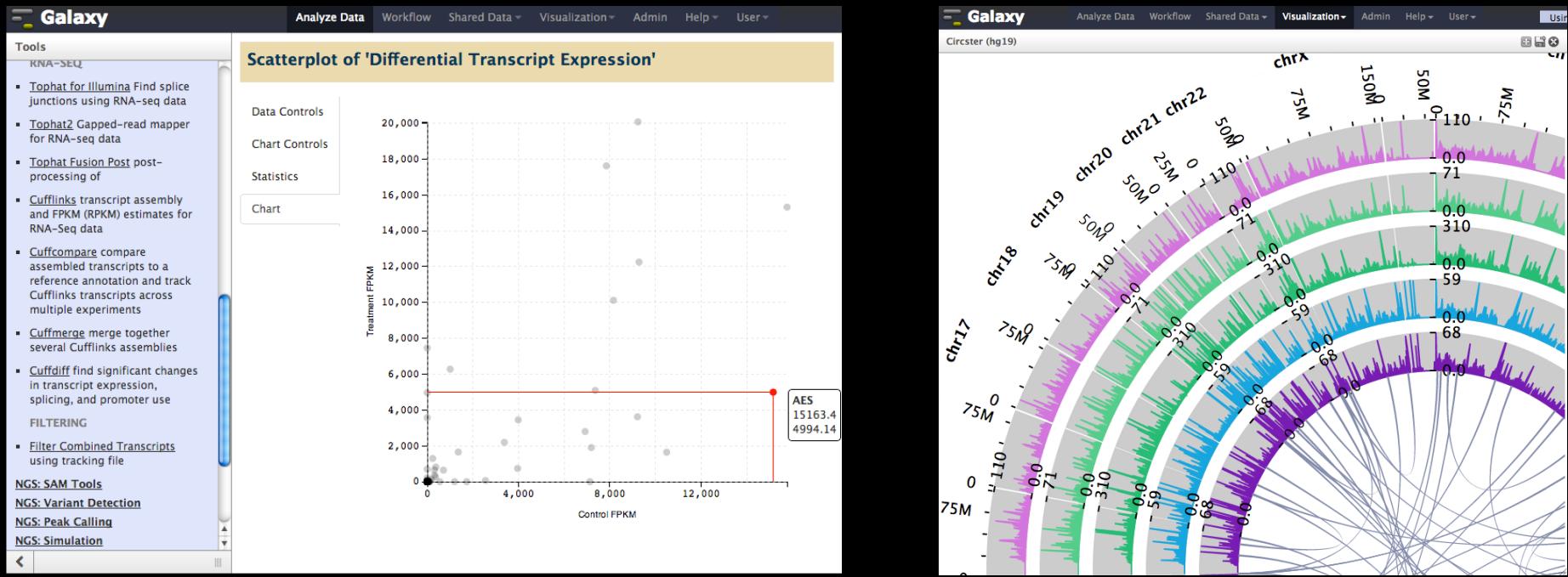
Analysis environment automatically  
and transparently tracks details



# 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

127.0.0.1:8080/root/index — Galaxy

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Using 23.9 KB

Tools

search tools

Get Data Send Data Lift-Over Text Manipulation Filter and Sort Join, Subtract and Group Convert Formats 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 FASTA manipulation NGS: QC and manipulation NGS: Mapping NGS: GATK Tools (beta) NGS: Simulation Phenotype Association Workflows All workflows

Charts – New Chart

Start Configuration Add Data Draw

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?

History

Unnamed history 23.9 KB

1: tabular.txt 100 regions format: interval, database: ? uploaded interval file

1. Chrom	2. Start	3. End	4	5	6
43	14	75	95	85	5
33	100	32	20	17	5
5	60	46	54	34	8
35	73	40	58	21	2
45	3	36	35	5	1
6	84	89	72	30	7

# Visualization framework: Charts plugin

127.0.0.1:8080/root/index — Galaxy

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Using 23.9 KB

Tools

search tools

Get Data Send Data Lift-Over Text Manipulation Filter and Sort Join, Subtract and Group Convert Formats 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 FASTA manipulation NGS: QC and manipulation NGS: Mapping NGS: GATK Tools (beta) NGS: Simulation Phenotype Association Workflows All workflows

Charts – New Chart

List of created charts:

New Chart Bar diagram Last change: 9/3/2014, 14:06 Delete New

New Chart

Customize

Grouped Stacked

1:Data label 2:Data label 3:Data label 4:Data label 5:Data label

Y-axis X-axis

History

Unnamed history 23.9 KB

1: tabular.txt 100 regions format: interval, database: ? uploaded interval file

1. Chrom	2. Start	3. End	4	5	6
43	14	75	95	85	5
33	100	32	20	17	5
5	60	46	54	34	8
35	73	40	58	21	2
45	3	36	35	5	1
6	84	89	72	30	7

# Visualization framework: Charts plugin

127.0.0.1:8080/root/index — Galaxy

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Using 23.9 KB

Tools

search tools

Get Data Send Data Lift-Over Text Manipulation Filter and Sort Join, Subtract and Group Convert Formats 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 FASTA manipulation NGS: QC and manipulation NGS: Mapping NGS: GATK Tools (beta) NGS: Simulation Phenotype Association Workflows All workflows

Charts – New Chart

List of created charts:

New Chart	Stacked area	Last change: 9/3/2014, 14:07
New Chart	Bar diagram	Last change: 9/3/2014, 14:06

Delete New

New Chart

Customize

Stacked Stream Expanded

Y-axis X-axis

1.Chrom	2.Start	3.End	4	5	6
43	14	75	95	85	5
33	100	32	20	17	5
5	60	46	54	34	8
35	73	40	58	21	2
45	3	36	35	5	1
6	84	89	72	30	7

History Unnamed history 23.9 KB

1: tabular.txt 100 regions format: interval, database: ? uploaded interval file

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

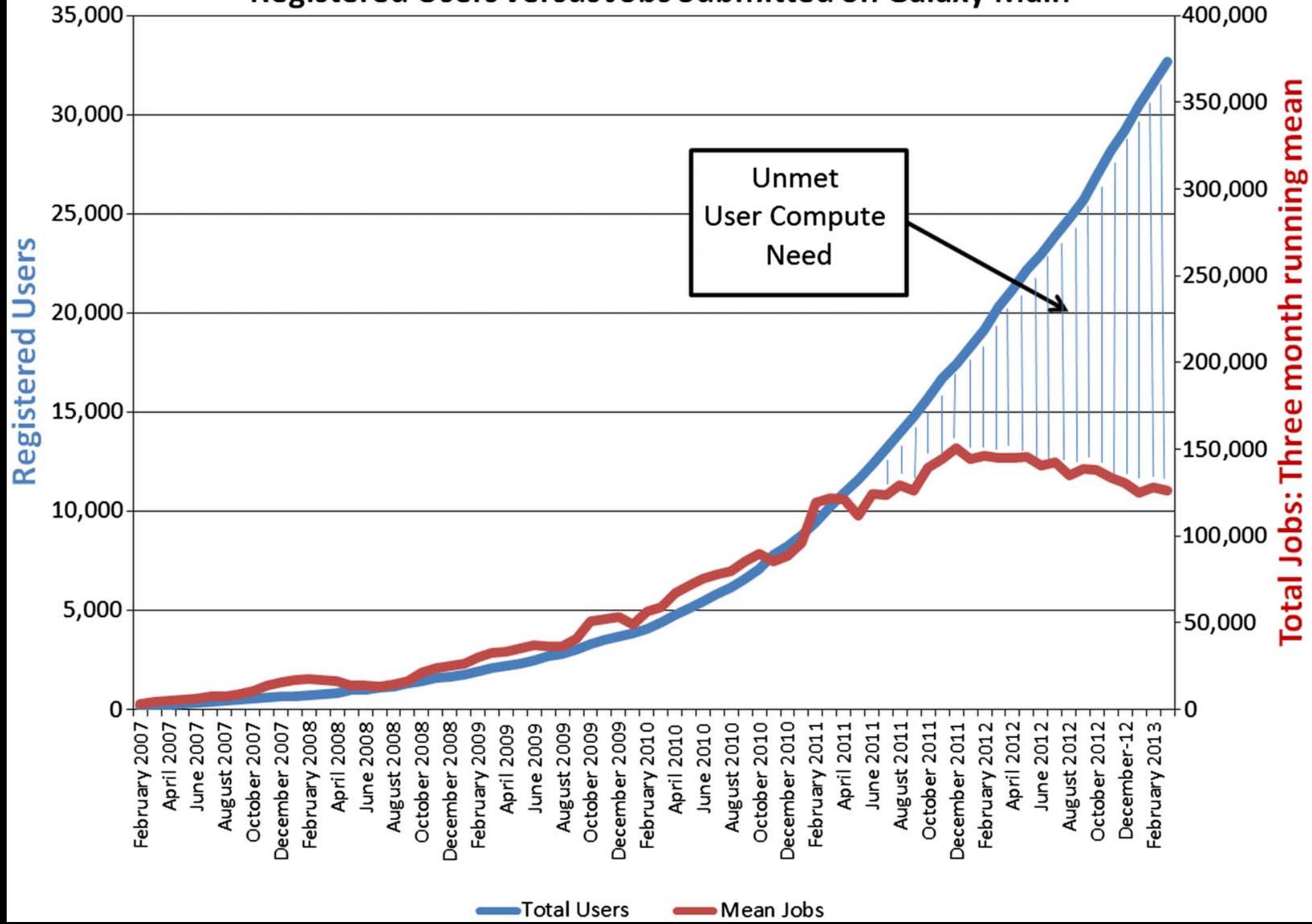
The Texas Advanced Computing Center (TACC) has already built substantial infrastructure in the context of the iPlant project

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

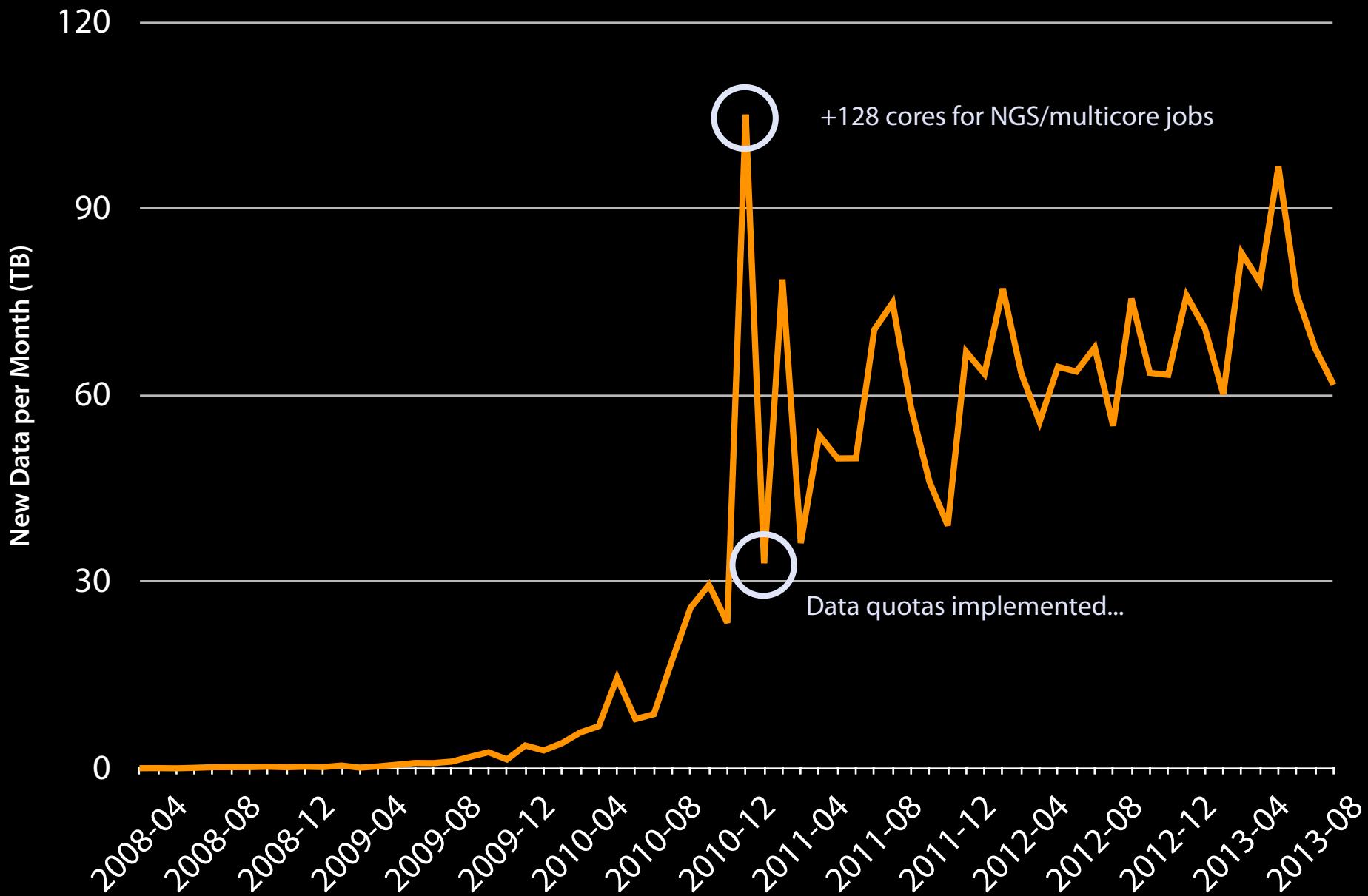
However, the iPlant and TACC cyber-infrastructure was underused; thus we established a collaboration

Since October 2013 Galaxy Main has run from TACC

## Registered Users versus Jobs Submitted on Galaxy Main

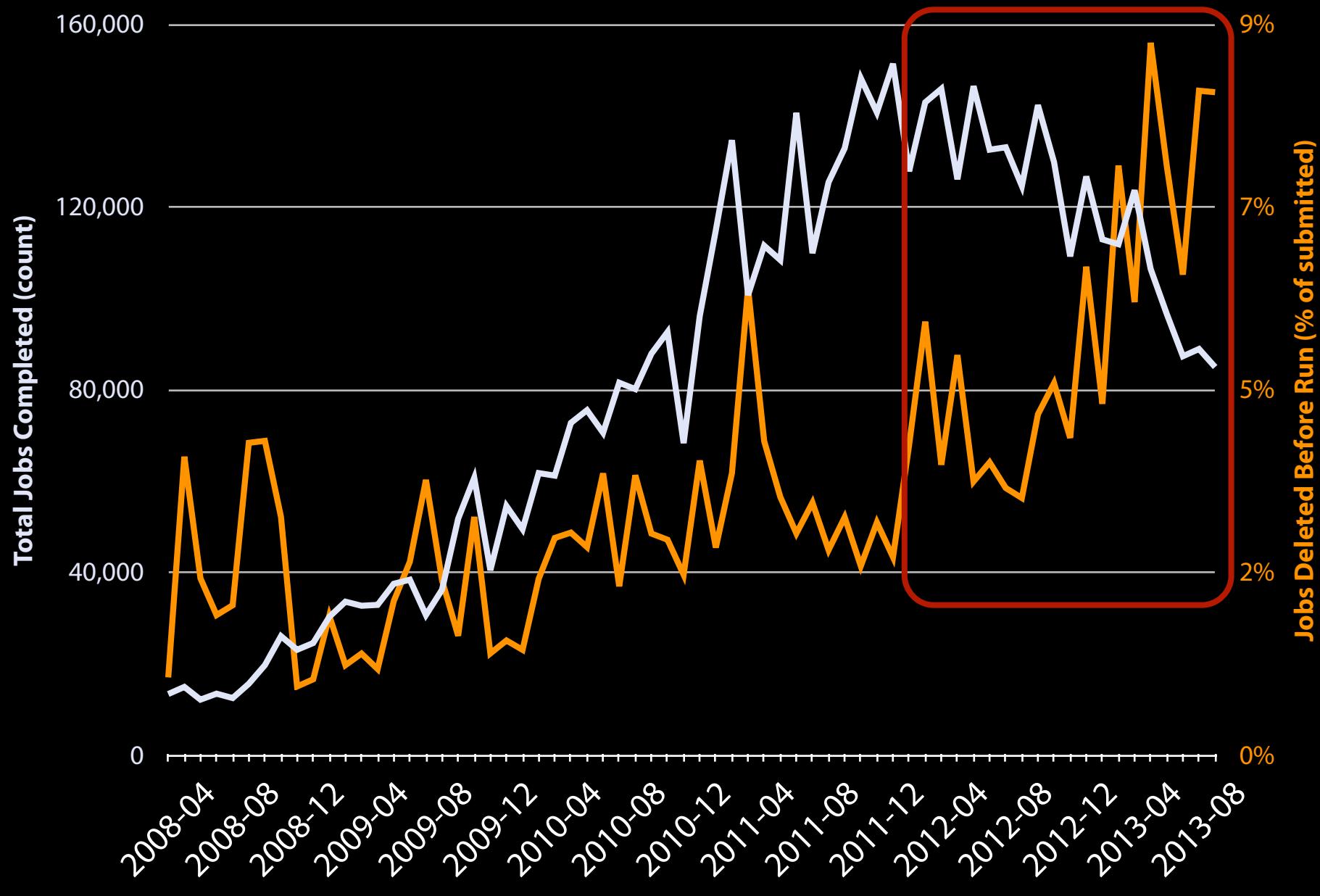


# usegalaxy.org data growth



Nate Coraor

# usegalaxy.org frustration growth



Nate Coraor

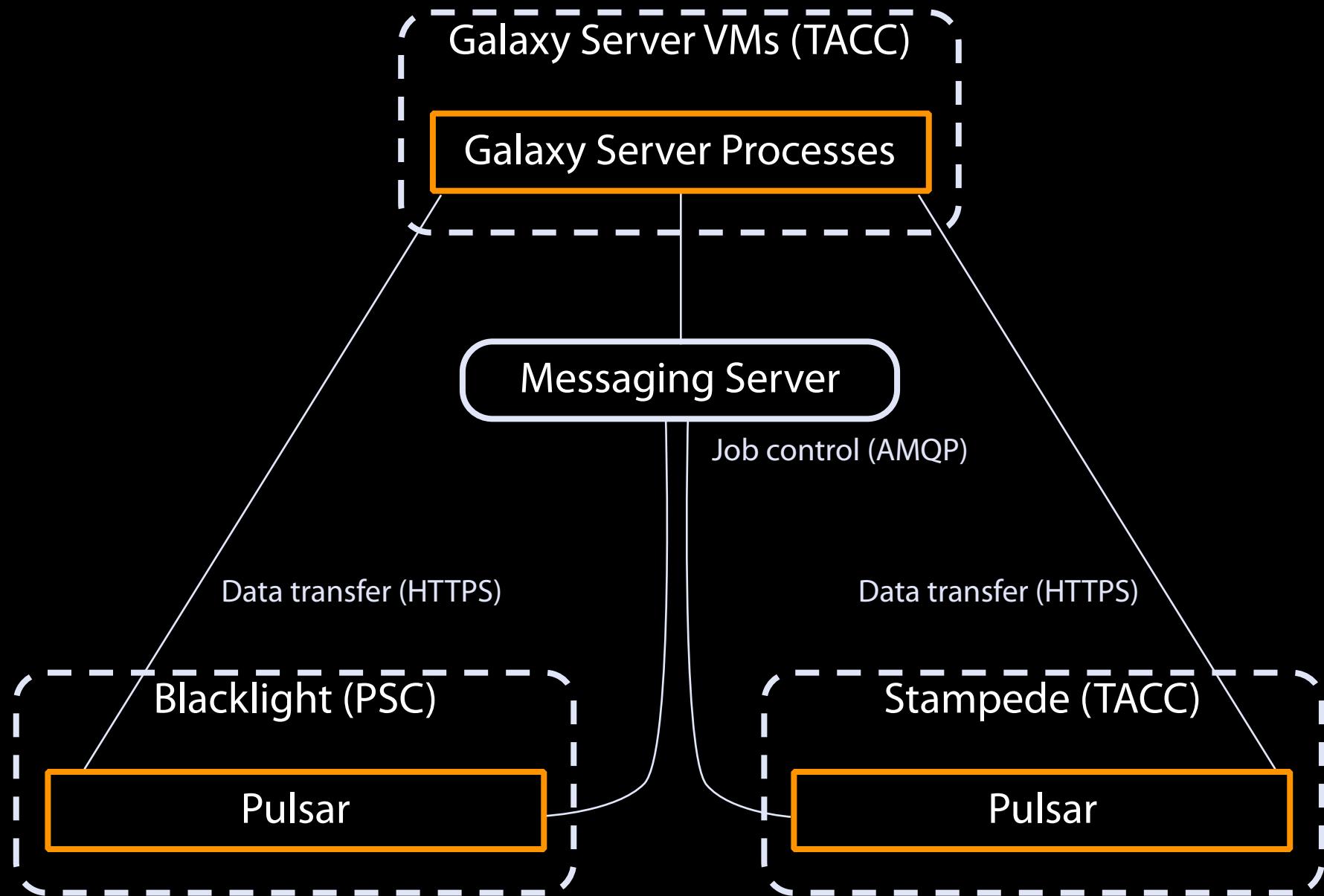
## How can this possibly scale?

1. Leverage existing 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

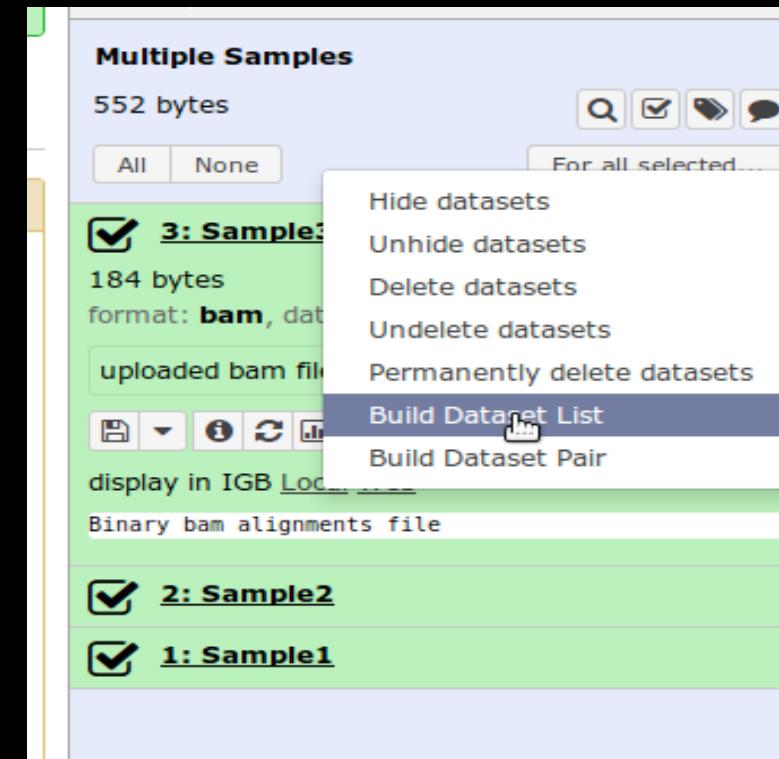


## Some thoughts on the future...

- 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.



The interface shows a tool configuration window for "Filter SAM or BAM (version 1.1.1)". The "SAM or BAM File to Filter:" dropdown is set to "4: New Dataset List". A tooltip "Run tool in parallel across dataset collection" points to this dropdown. The "History" panel shows the following dataset list:

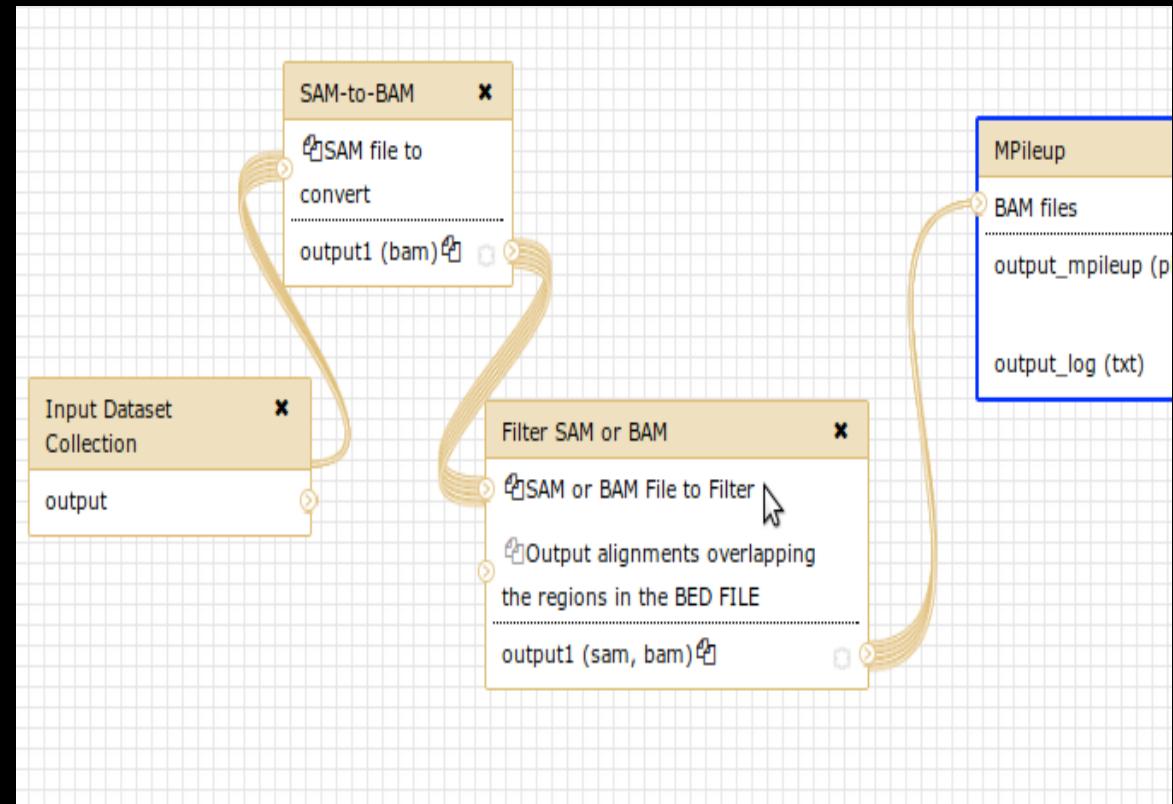
- 4: New Dataset List
- 3: Sample3

The "Header in output:" dropdown is set to "Include Header". The "Minimum MAPQ quality score:" input field is present.

# 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

---



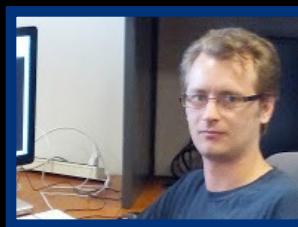
Enis Afgan



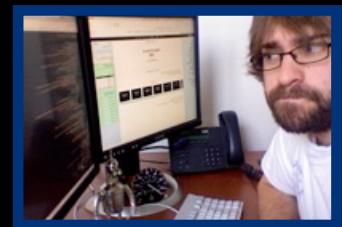
Dannon Baker



Dan Blankenberg



Dave Bouvier



Nate Coraor



Martin Čech



John Chilton



Carl Eberhard



Sam Guerler



Nick Stoler

## Support and outreach

---



Dave Clements



Jennifer Jackson



James Taylor



Anton Nekrutenko



Jeremy Goecks

## Leadership

---

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