

Galaxy

Data intensive biology *for everyone.*

www.galaxyproject.org

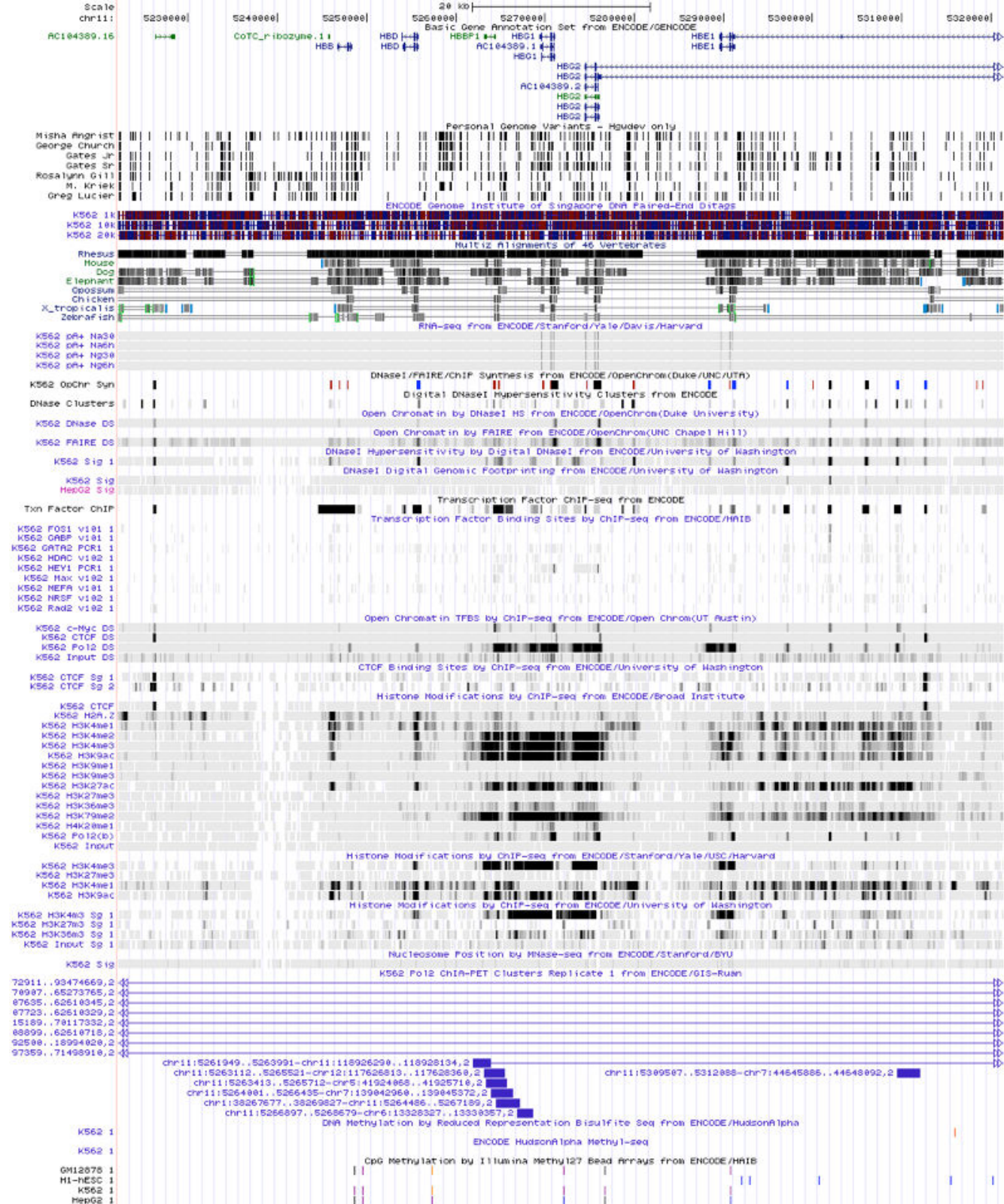


@jctx / #usegalaxy

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)

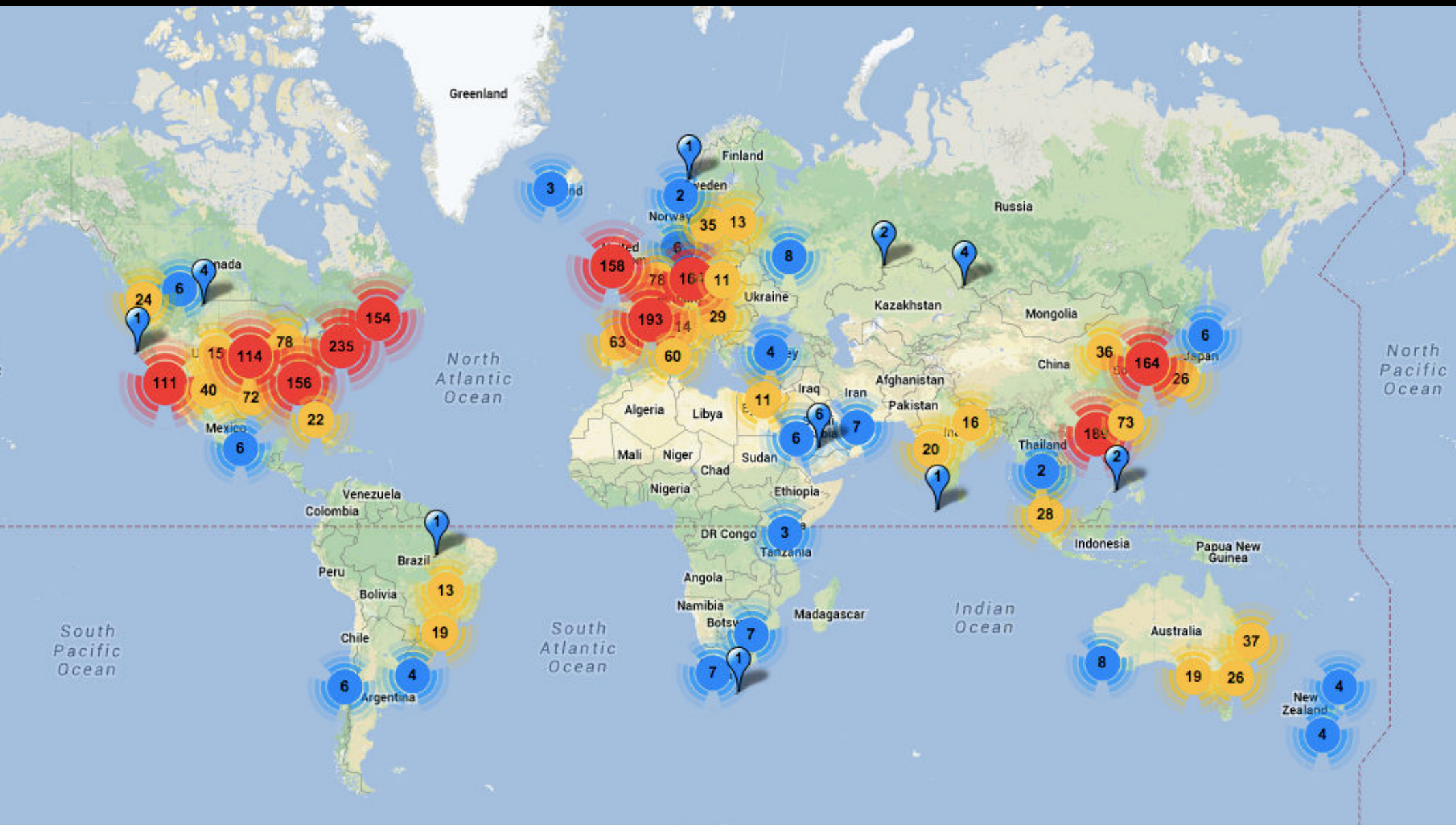
Histones variants (ChIP-seq, MNase-seq)

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

Methylation (Bisulfite-seq)

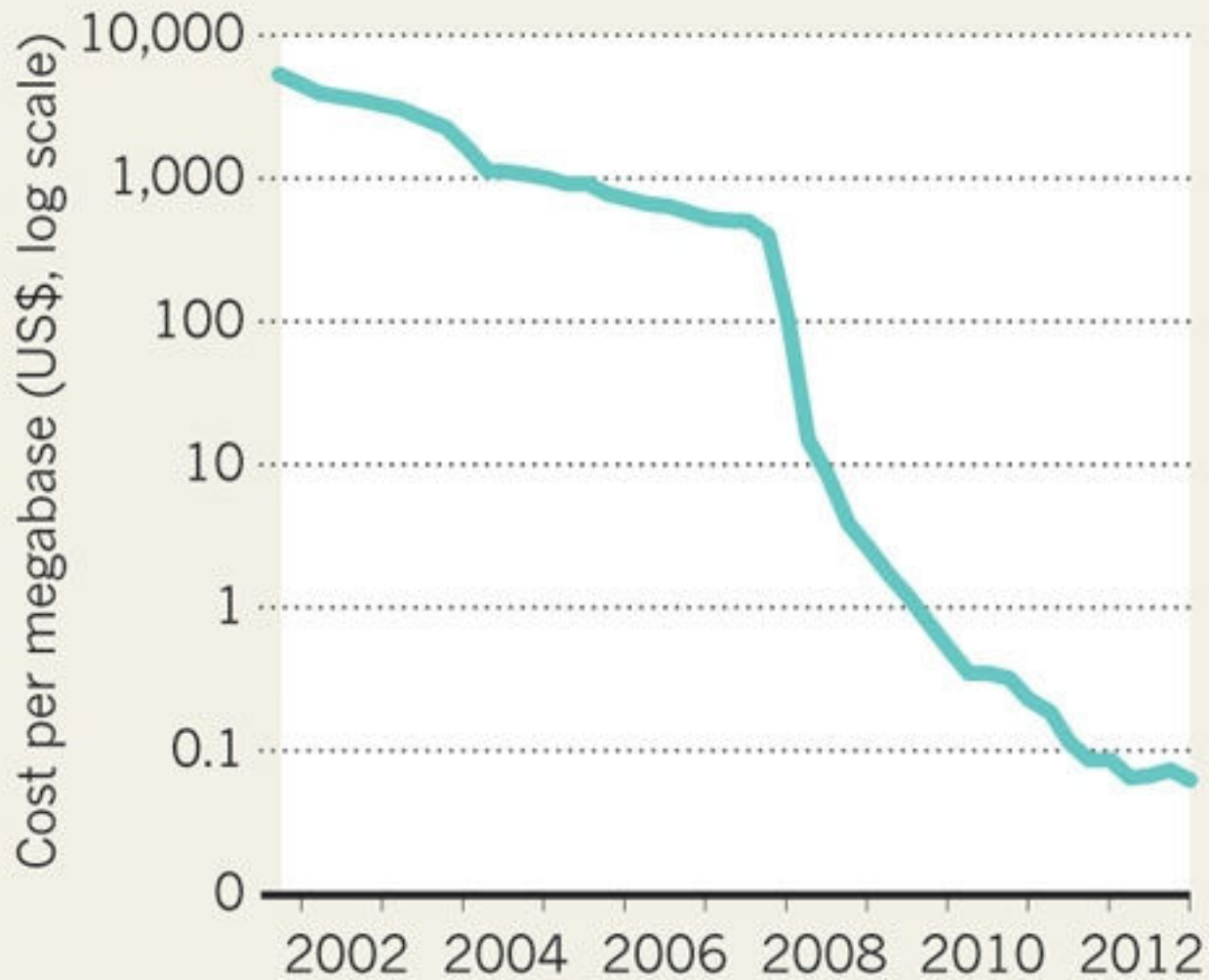
High-throughput sequencing is
democratizing

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**?

Galaxy: accessible analysis system

The screenshot displays the Galaxy web interface in a browser window. The address bar shows <http://main.g2.bx.psu.edu/>. The top navigation bar includes links for **Analyze Data**, **Workflow**, **Data Libraries**, **Admin**, **Help**, and **User**.

Tools Panel (Left): A list of available tools categorized into sections: **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**, and **SNP/WGA: QC; LD; Plots**.

Main Content Area: A central box titled "Here is what's happening..." features a large text overlay: "Mapping Pipeline for Illumina, 454, and SOLiD" with a subtext "A new accessible research tool" and a prominent "USE IT NOW!" button. Below this, a section titled "Live Quickies (more after May 17 ...)" displays three cards: "Basic fastQ manipulation: Galactic quickie # 13", "Advanced fastQ manipulation: Galactic quickie # 14", and "454 Mapping: Single End: Galactic quickie # 15".

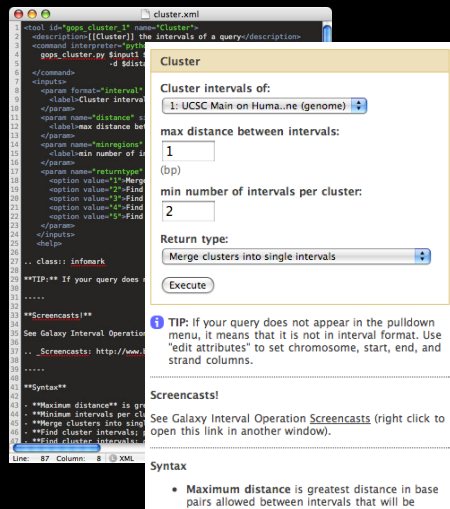
History Panel (Right): A list of workflow steps, each with a name, an eye icon, and a delete icon. The steps are: 16: Draw phylogeny on data 14, 15: Summarize taxonomy on data 13, 14: Find lowest diagnostic rank on data 13, 13: Fetch taxonomic representation on data 12, 12: Filter on data 11, 11: Join 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, and 4: FASTA-to-Tabular on data 4.

Footer: Text at the bottom states: "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\$".

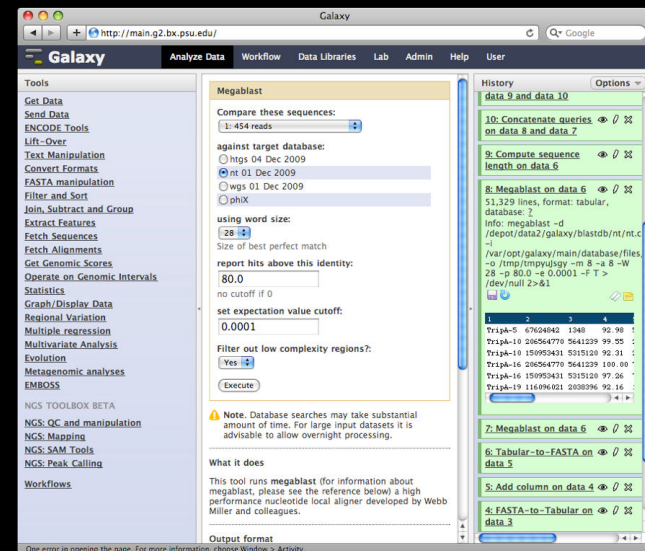
A free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

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

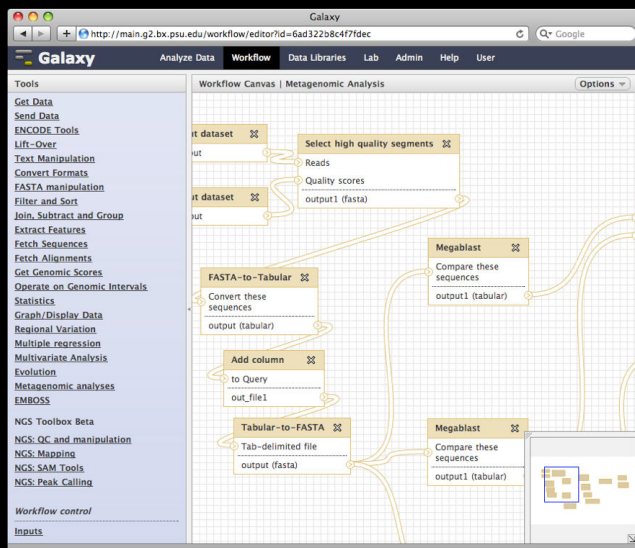
An open extensible platform for sharing tools, datatypes, workflows, ...



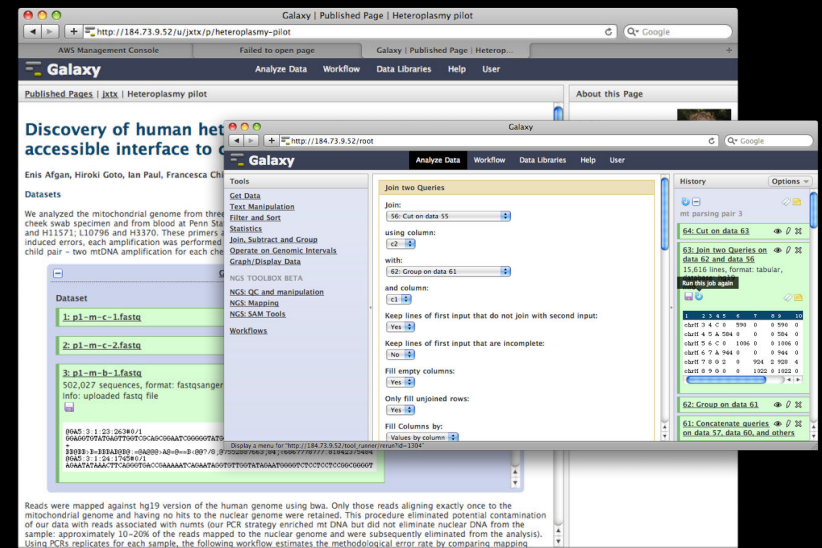
Describe analysis tool
behavior abstractly



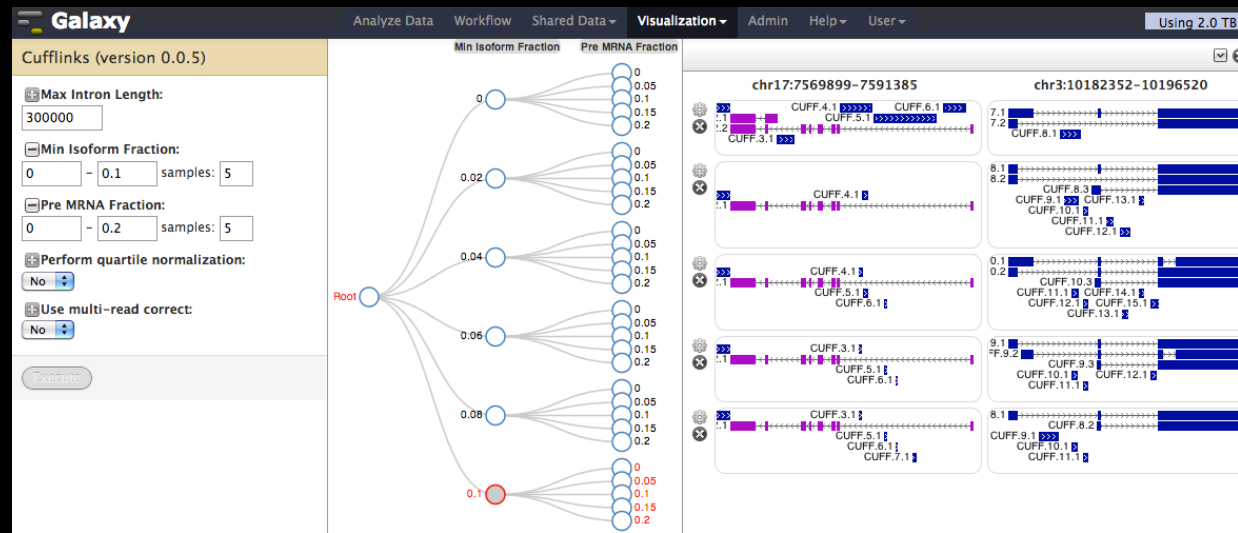
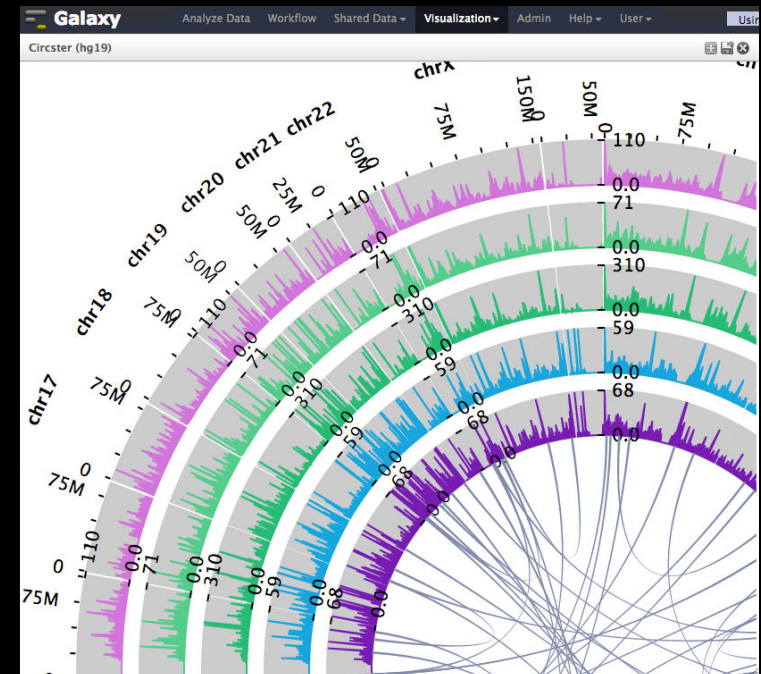
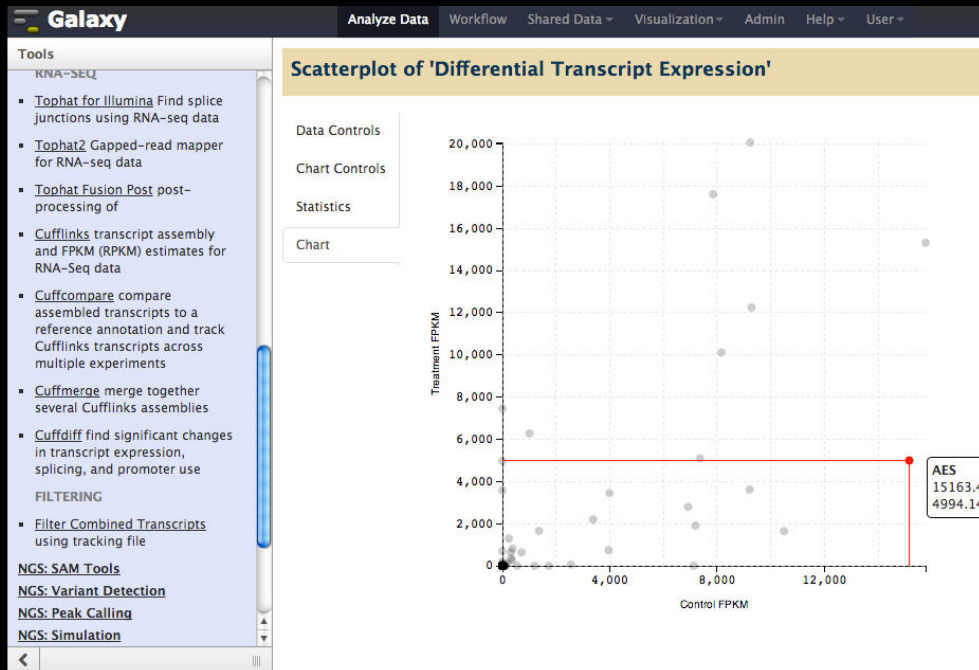
Analysis environment automatically
and transparently tracks details



Workflow system for complex analysis,
constructed explicitly or automatically



Pervasive sharing, and publication
of documents with integrated analysis

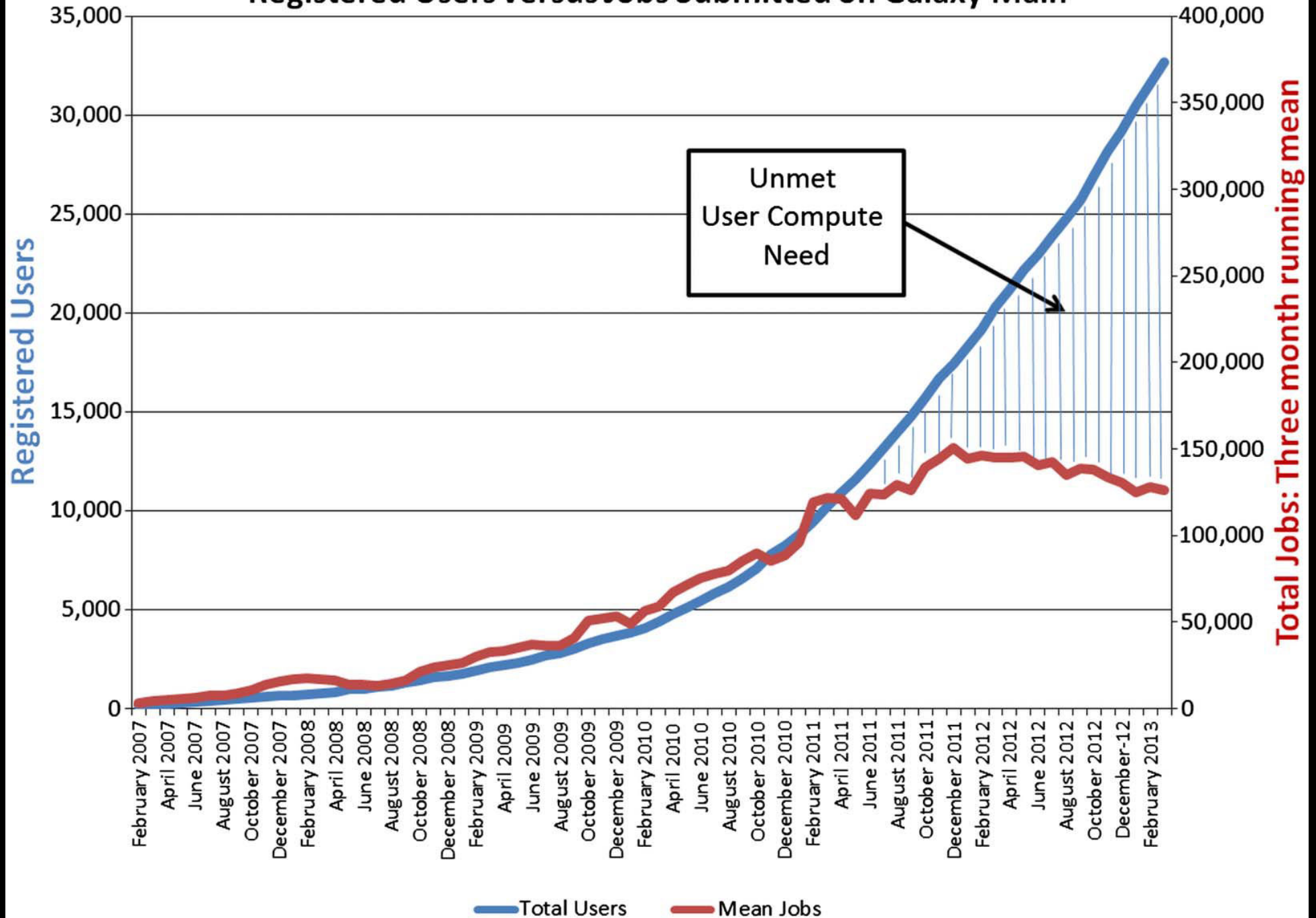


Visualization and visual analytics

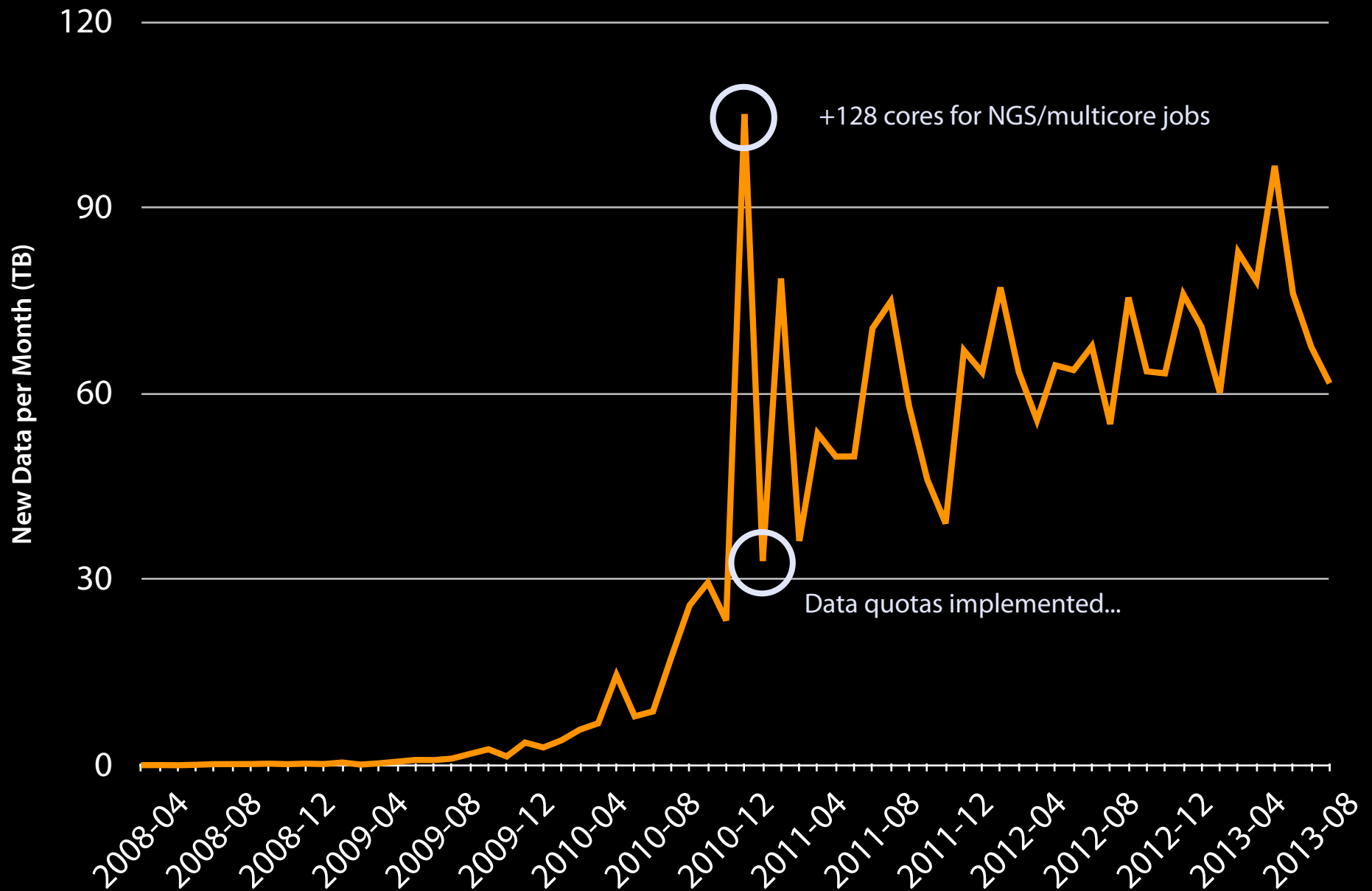
The free service is still the easiest way for users with no informatics infrastructure to analyze their data

How can we possibly sustain this?

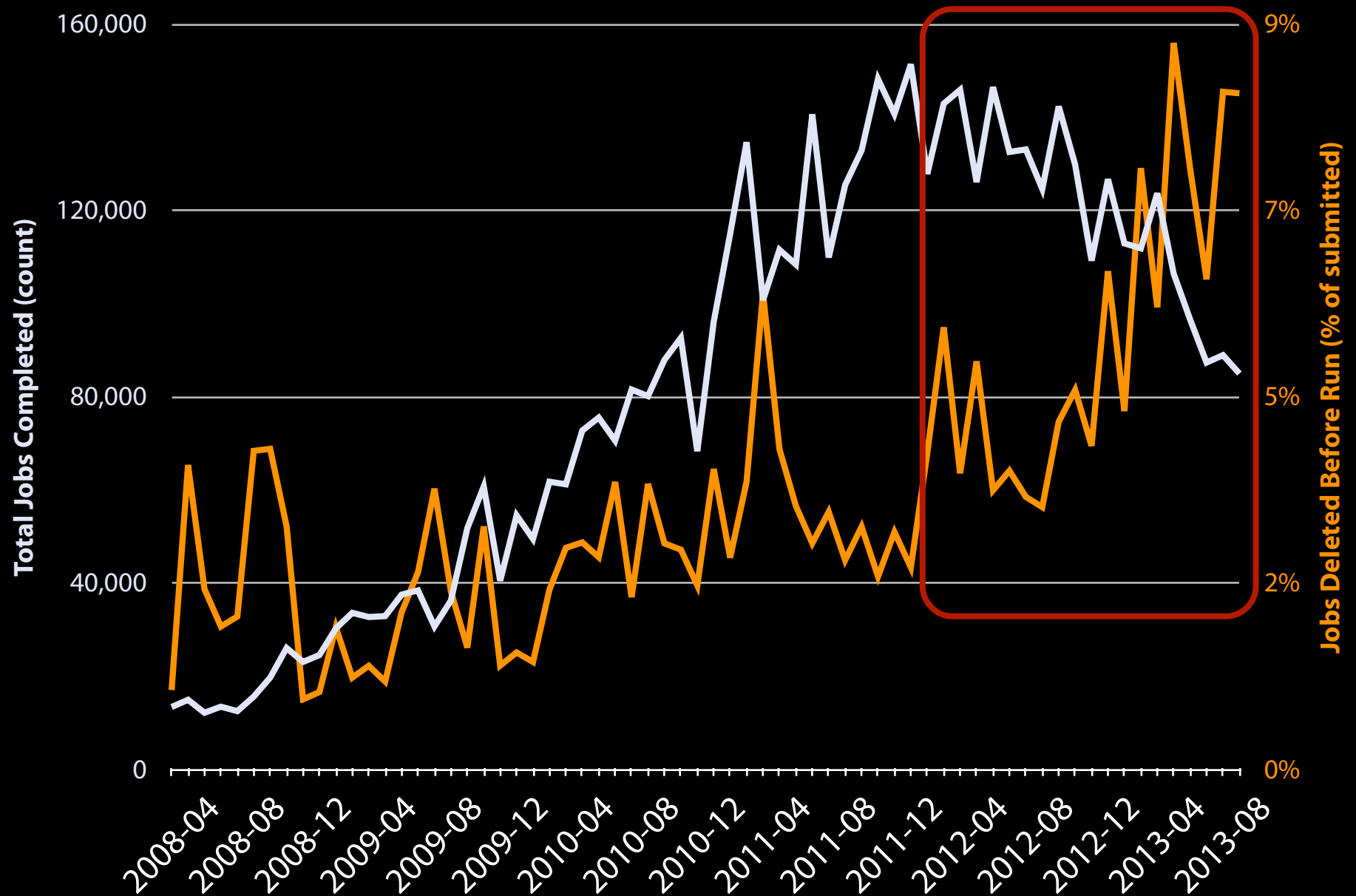
Registered Users versus Jobs Submitted on Galaxy Main



usegalaxy.org data growth



usegalaxy.org frustration growth



How can this possibly scale?

1. Leverage existing public cyber-infrastructure
2. Decentralize, provide many deployment models (cloud and local — not talking about this today)

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

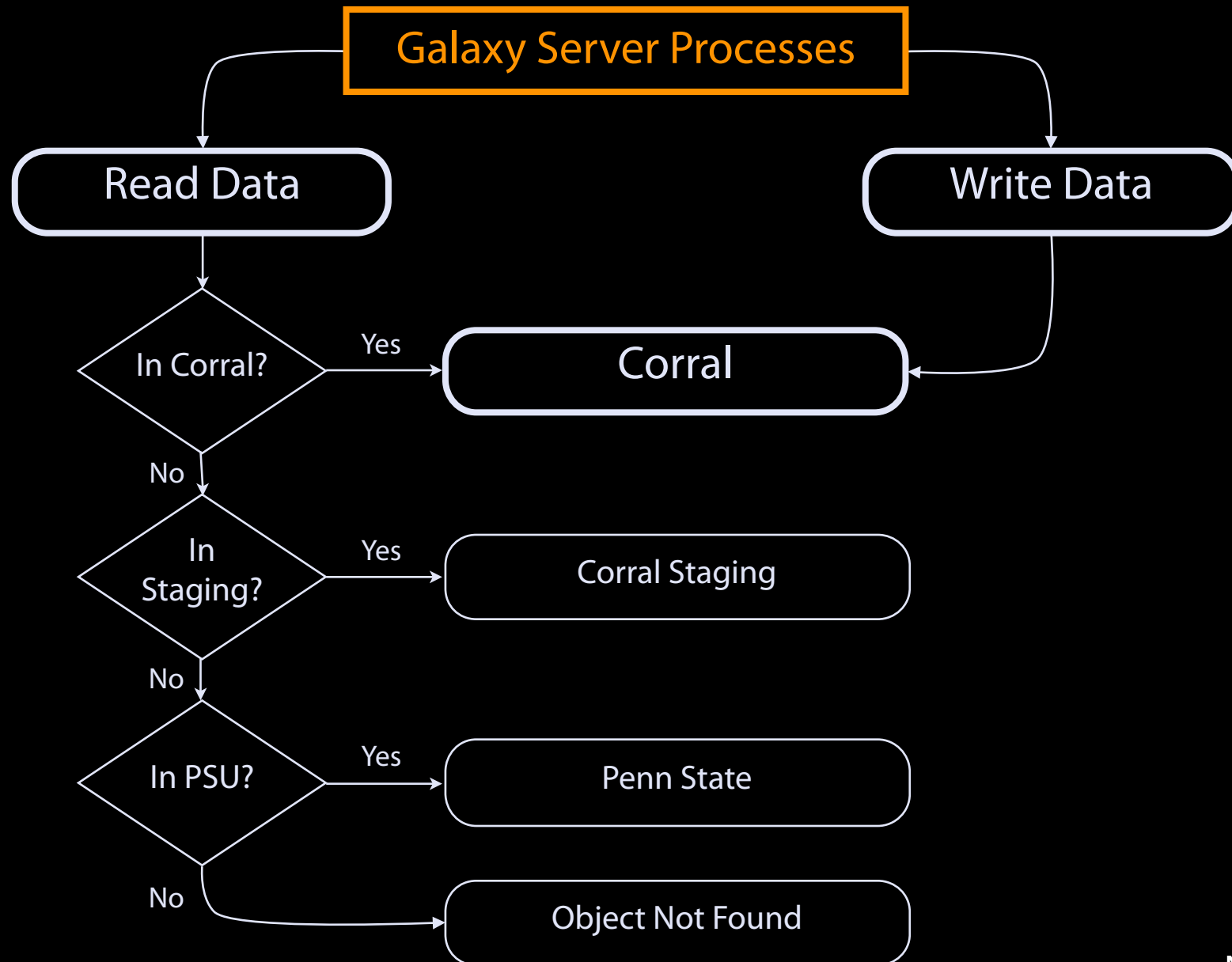
The Texas Advanced Computing Center (TACC) has already built substantial infrastructure in the context of the iPlant project

(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

Transparent Migrations using Galaxy's Hierarchical Object Store



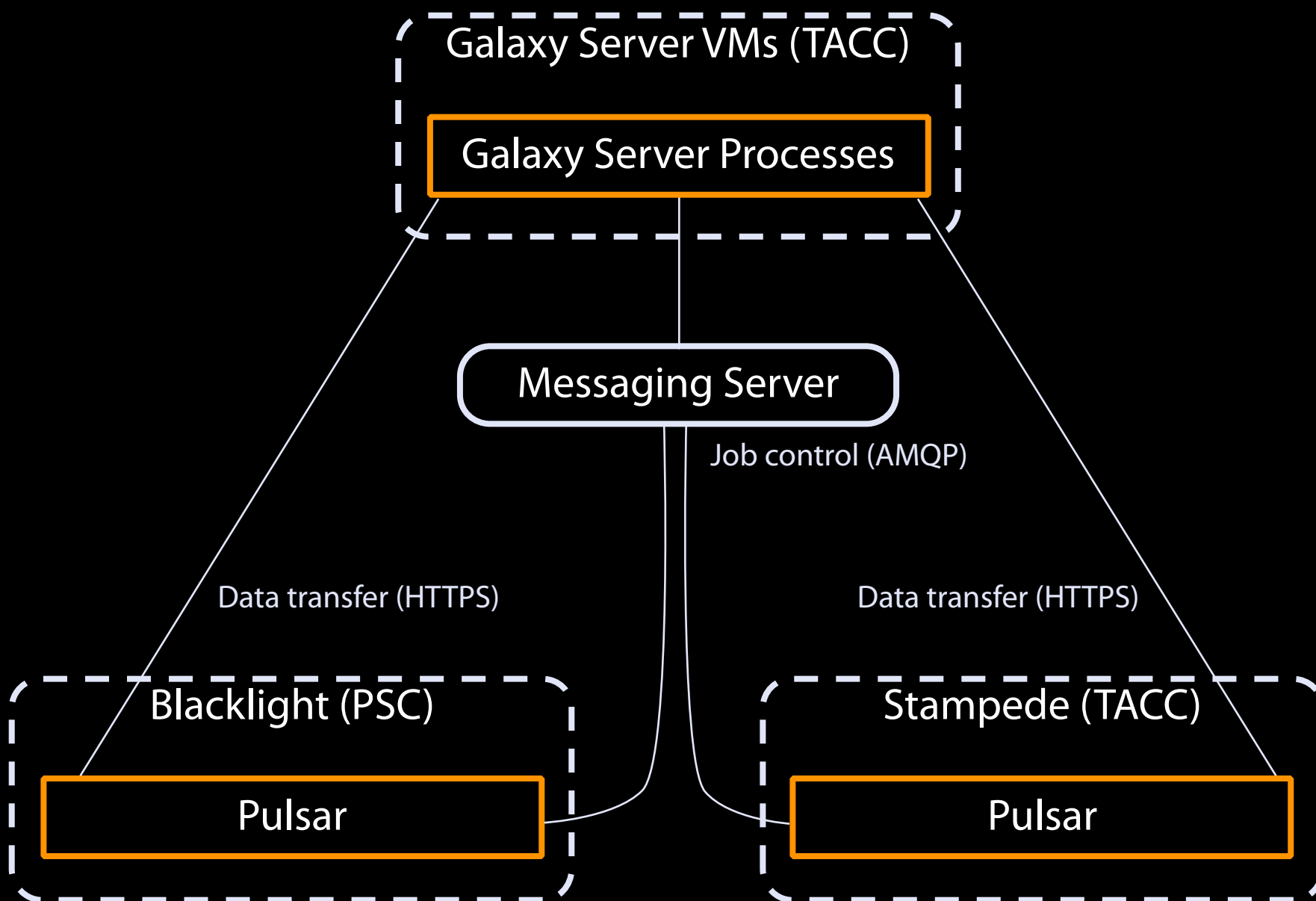
Expanding to more XSEDE resources

Galaxy can already run jobs on almost any batch system, but most XSEDE resources do not provide direct access for job submission...

Pulsar

Galaxy job runner that can
run almost anywhere

No shared filesystem, stages all necessary
Galaxy components



Moving long running jobs out to XSEDE

- Problem:
 - Jobs wait in the queue for a long time
 - Jobs may fail immediately upon run due to bad parameters
 - Most jobs run quickly! Can we relocate the long ones?
- Goals:
 - Shorten wait from submission to start
 - Allow testing params without waiting
- Solutions:
 - Set a short walltime, resubmit jobs to bigger resources (new code)
 - User selection of resources (Stampede - longer wait to start, but more concurrent jobs allowed)
 - Create “development” queues w/ short walltime

State of Affairs

- Today
 - Galaxy Test jobs to Stampede and Blacklight
 - Galaxy Main jobs to Stampede
- Up next
 - Galaxy Main jobs to Blacklight
 - Optimize Trinity tools for Blacklight
 - Linking XSEDE allocations to Galaxy accounts

Credits

- Texas Advanced Computing Center
 - Dan Stanzione
 - Matt Vaughn
 - Chris Jordan
 - Mike Packard
 - Nathaniel Mendoza
- iPlant Collaborative
 - Stephen Goff
- Pittsburgh Supercomputing Center
 - Philip Blood
 - Kathy Benninger
 - Robert Budden
 - Jared Yanovich
 - Josephine Palencia
 - J. Ray Scott
 - Joe Lappa

... and the Galaxy Team and community

Galaxy is supported in part by NSF, NHGRI, Pennsylvania Department of Public Health, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University

Engineering



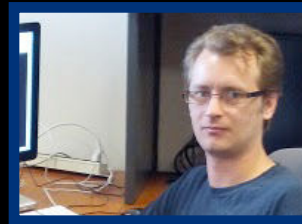
Enis Afgan



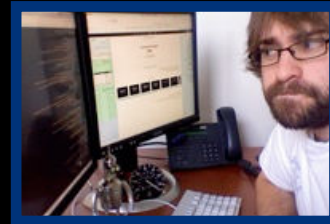
Dannon Baker



Dan Blankenberg



Dave Bouvier



Nate Coraor



Martin Čech



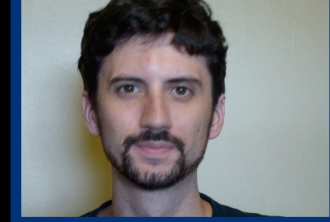
John Chilton



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