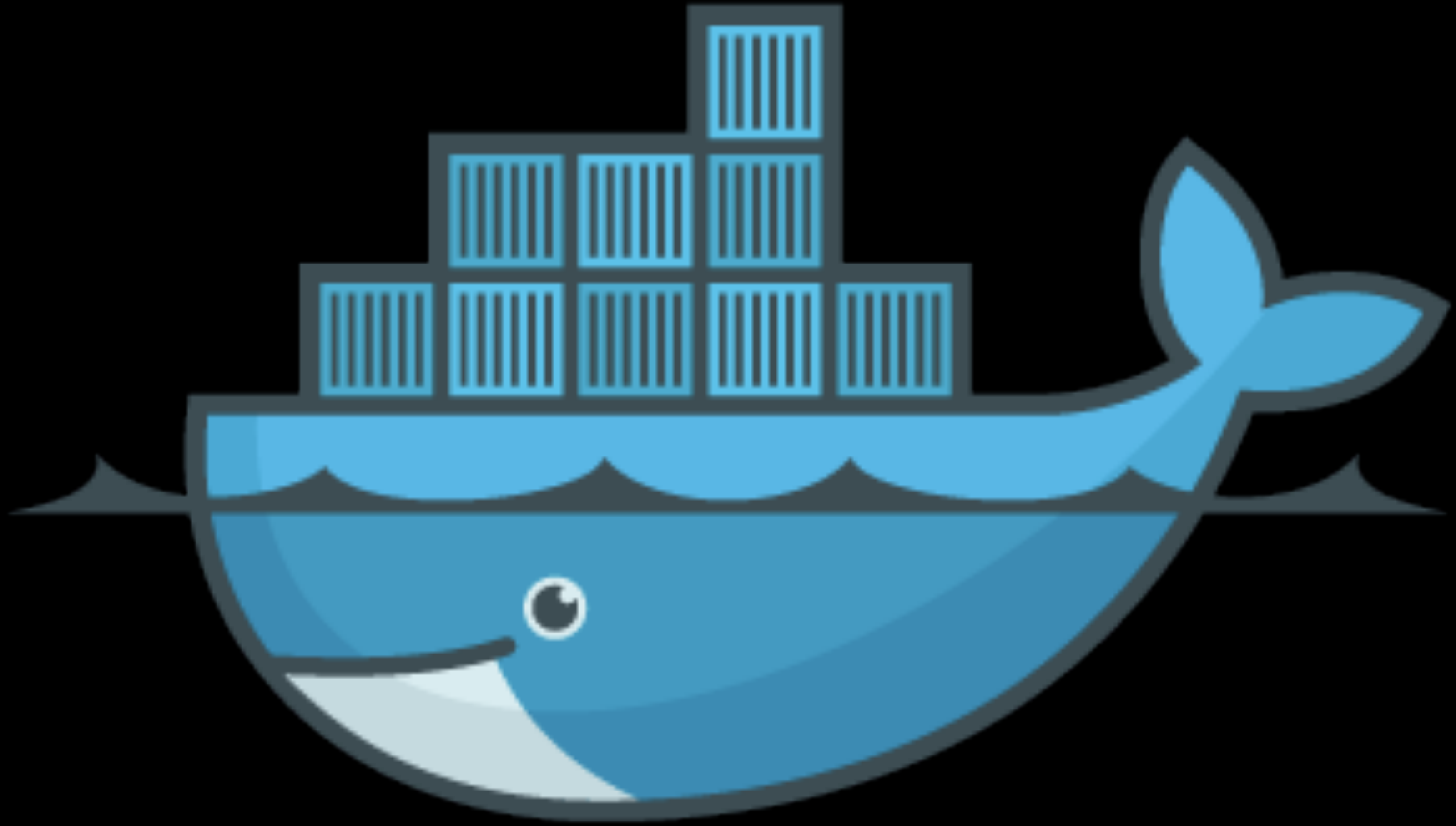
Adventures in Scaling Galaxy

https://speakerdeck.com/jxtx @jxtx / #usegalaxy



…in which I will not talk about the ~~elephant~~ whale in the room…

Galaxy’s motivating questions

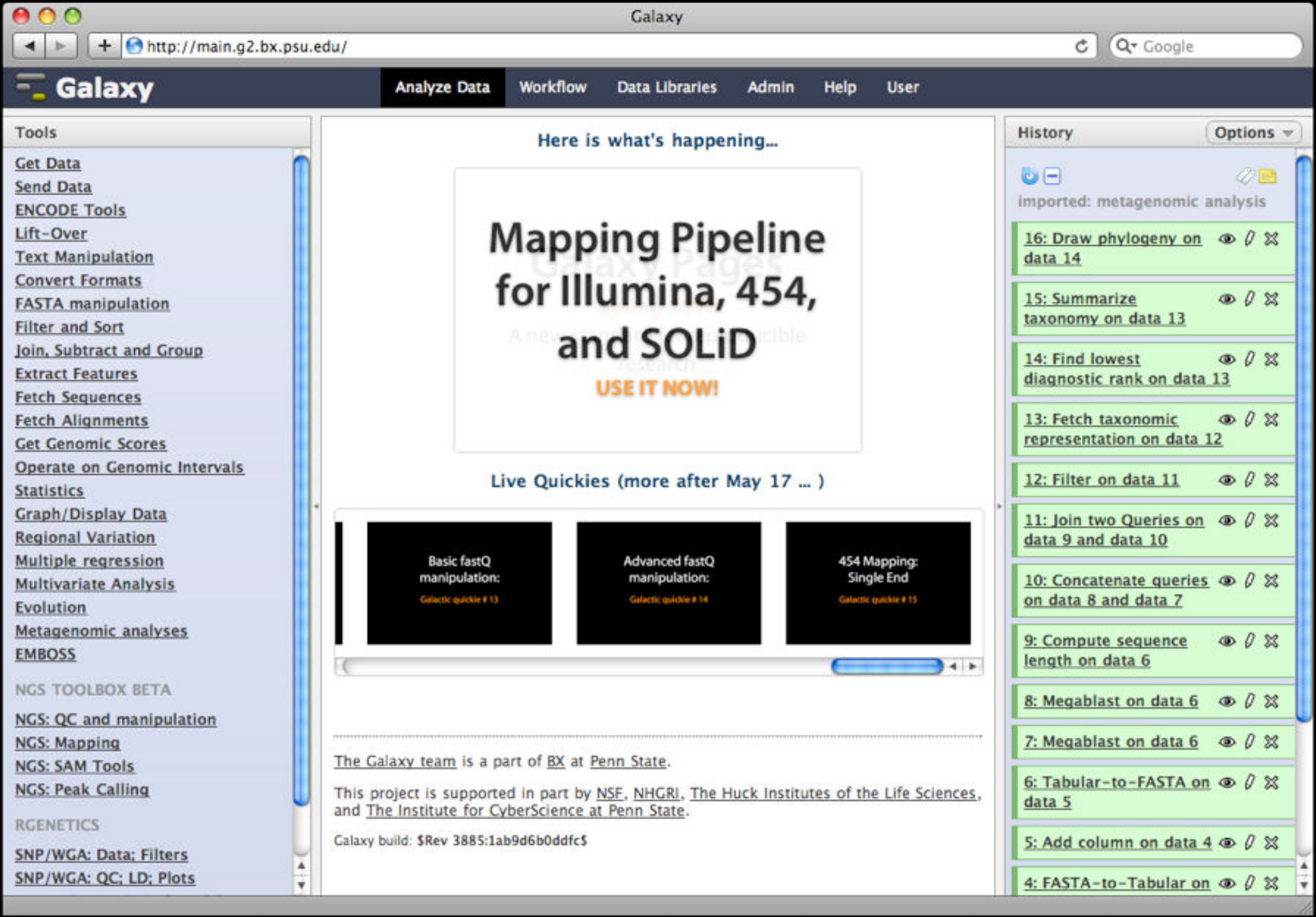
How best can data intensive methods be **accessible** to scientists?

How best to facilitate **transparent communication** of computational analyses?

How best to ensure that analyses are **reproducible***?

The state of which is* ***frighteningly* *bad, see doi:10.1038/nrg3305, doi:10.7717/peerj.148*

Galaxy: accessible analysis system



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

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: $Rev 3885:1ab9d6b0ddfc$

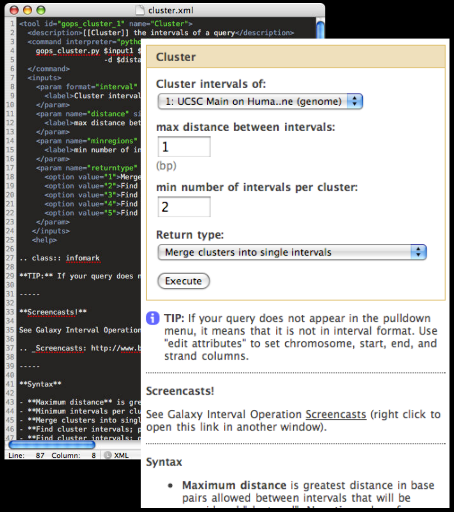
imported: metagenomic analysis

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, ...



cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="pytho
    gops_cluster.py $input1 $
                     -d $dista
  </command>
  <inputs>
    <param format="interval"
      <label>Cluster interval
    </param>
    <param name="distance" si
      <label>max distance bet
    </param>
    <param name="minregions"
      <label>min number of in
    </param>
    <param name="returntype"
      <option value="1">Merge
      <option value="2">Find
      <option value="3">Find
      <option value="4">Find
      <option value="5">Find
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does n

-----

**Screencasts!**

See Galaxy Interval Operation

.. _Screencasts: http://www.b

-----

**Syntax**

- **Maximum distance** is gre
- **Minimum intervals per clu
- **Merge clusters into singl
- **Find cluster intervals; p
- **Find cluster intervals; c
```

Line: 87 Column: 8 XML

Cluster

Cluster intervals of: 1: UCSC Main on Huma..ne (genome)

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

TIP: If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

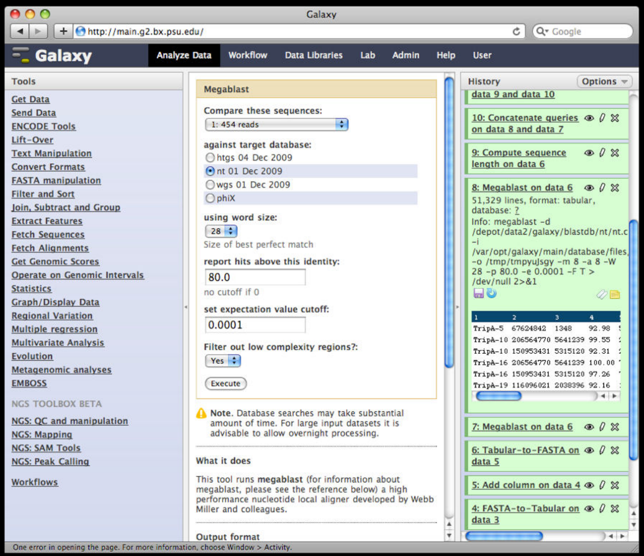
Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be

Describe analysis tool behavior abstractly



Megablast

Compare these sequences: 1: 454 reads

against target database:
○ htgs 04 Dec 2009
● nt 01 Dec 2009
○ wgs 01 Dec 2009
○ phiX

using word size: 28
Size of best perfect match

report hits above this identity: 80.0
no cutoff if 0

set expectation value cutoff: 0.0001

Filter out low complexity regions?: Yes

Note. Database searches may take substantial amount of time. For large input datasets it is advisable to allow overnight processing.

What it does

This tool runs **megablast** (for information about megablast, please see the reference below) a high performance nucleotide local aligner developed by Webb Miller and colleagues.

Output format

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

51,329 lines, format: tabular, database: ?
Info: megablast -d /depot/data2/galaxy/blastdb/nt/nt.c -i /var/opt/galaxy/main/database/files, -o /tmp/tmpyuJsgy -m 8 -a 8 -W 28 -p 80.0 -e 0.0001 -F T > /dev/null 2>&1

```
1         2         3       4
TripA-5   67624842  1348    92.98
TripA-10  206564770 5641239 99.55
TripA-10  150953431 5315120 92.31
TripA-16  206564770 5641239 100.00
TripA-16  150953431 5315120 97.26
TripA-19  116096021 2038396 92.16
```

7: Megablast on data 6

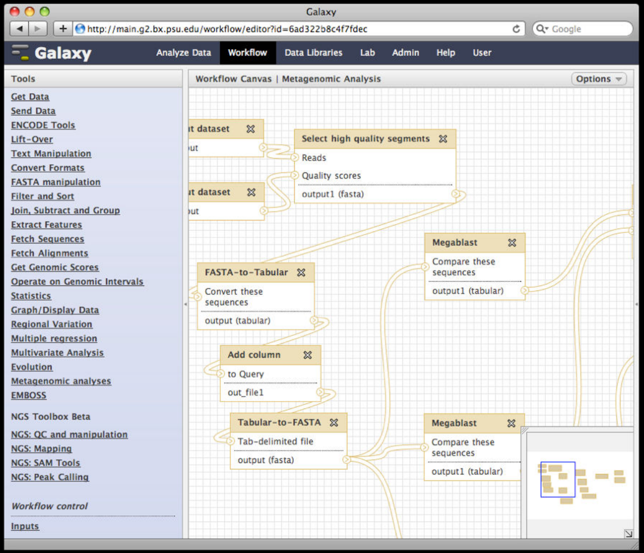
6: Tabular-to-FASTA on data 5

5: Add column on data 4

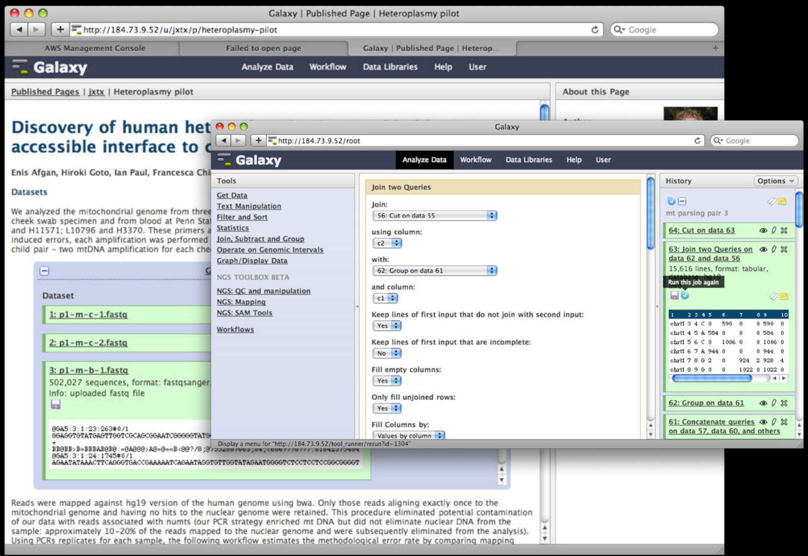
4: FASTA-to-Tabular on data 3

One error in opening the page. For more information, choose Window > Activity.

Analysis environment automatically and transparently tracks details



Workflow system for complex analysis, constructed explicitly or automatically



Discovery of human het
accessible interface to

Enis Afgan, Hiroki Goto, Ian Paul, Francesca Chi

Datasets

We analyzed the mitochondrial genome from three
cheek swab specimen and from blood at Penn Stat
and H11571; L10796 and H3370. These primers a
induced errors, each amplification was performed
child pair – two mtDNA amplification for each che

Dataset

1: p1-m-c-1.fastq

2: p1-m-c-2.fastq

3: p1-m-b-1.fastq
502,027 sequences, format: fastqsanger,
Info: uploaded fastq file

```
@GA5:3:1:23:263#0/1
GGAGGTGTATGAGTTGGTCGCAGCGGAATCGGGGGTATG
+
BB@BB>B=BBBAB@B@:=@A@@@>A@=@==B<@@?/8;@
@GA5:3:1:24:1745#0/1
AGAATATAAACTTCAGGGTGACCGAAAAAATCAGAATAGGTGTTGGTATAGAATGGGGTCTCCTCCTCCGGCGGGGT
```

Join two Queries

Join: 56: Cut on data 55

using column: c2

with: 62: Group on data 61

and column: c1

Keep lines of first input that do not join with second input: Yes

Keep lines of first input that are incomplete: No

Fill empty columns: Yes

Only fill unjoined rows: Yes

Fill Columns by: Values by column

Display a menu for "http://184.73.9.52/tool_runner/rerun?id=1304"

mt parsing pair 3

64: Cut on data 63

63: Join two Queries on data 62 and data 56
15,616 lines, format: tabular,

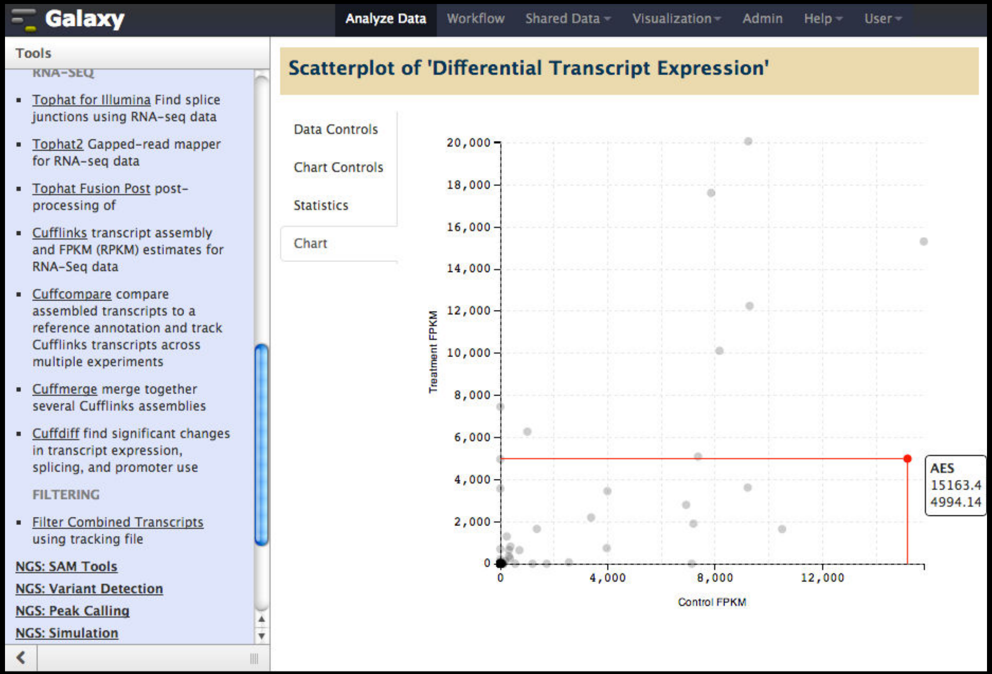
```
1     2 3 4 5  6     7     8 9     10
chrM  3 4 C 0  590   0     0 590   0
chrM  4 5 A 584 0    0     0 584   0
chrM  5 6 C 0  1006  0     0 1006  0
chrM  6 7 A 944 0    0     0 944   0
chrM  7 8 G 2  0     924   2 928   4
chrM  8 9 G 0  0     1022  0 1022  0
```

62: Group on data 61

61: Concatenate queries on data 57, data 60, and others

Reads were mapped against hg19 version of the human genome using bwa. Only those reads aligning exactly once to the mitochondrial genome and having no hits to the nuclear genome were retained. This procedure eliminated potential contamination of our data with reads associated with numts (our PCR strategy enriched mt DNA but did not eliminate nuclear DNA from the sample: approximately 10-20% of the reads mapped to the nuclear genome and were subsequently eliminated from the analysis). Using PCRs replicates for each sample, the following workflow estimates the methodological error rate by comparing mapping

Pervasive sharing, and publication of documents with integrated analysis



Tools

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat2 Gapped-read mapper for RNA-seq data
- Tophat Fusion Post post-processing of
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

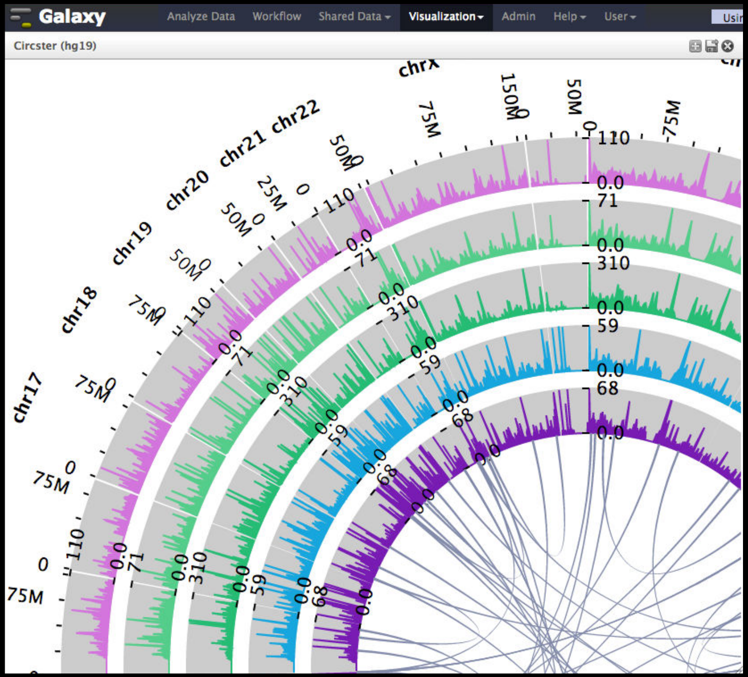
NGS: SAM Tools

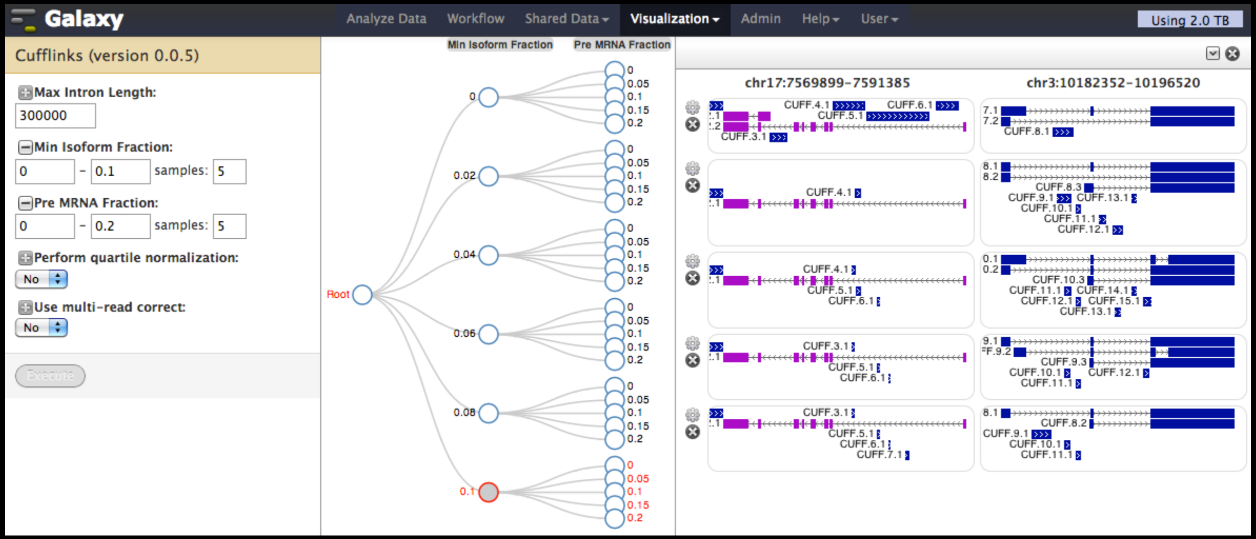
NGS: Variant Detection

NGS: Peak Calling

NGS: Simulation

Scatterplot of 'Differential Transcript Expression'





Cufflinks (version 0.0.5)

Max Intron Length: 300000

Min Isoform Fraction: 0 - 0.1 samples: 5

Pre MRNA Fraction: 0 - 0.2 samples: 5

Perform quartile normalization: No

Use multi-read correct: No

Execute

Visualization and visual analytics

Wait… what… a *free* web-service for high throughput sequence data analysis?!

Galaxy as a Service

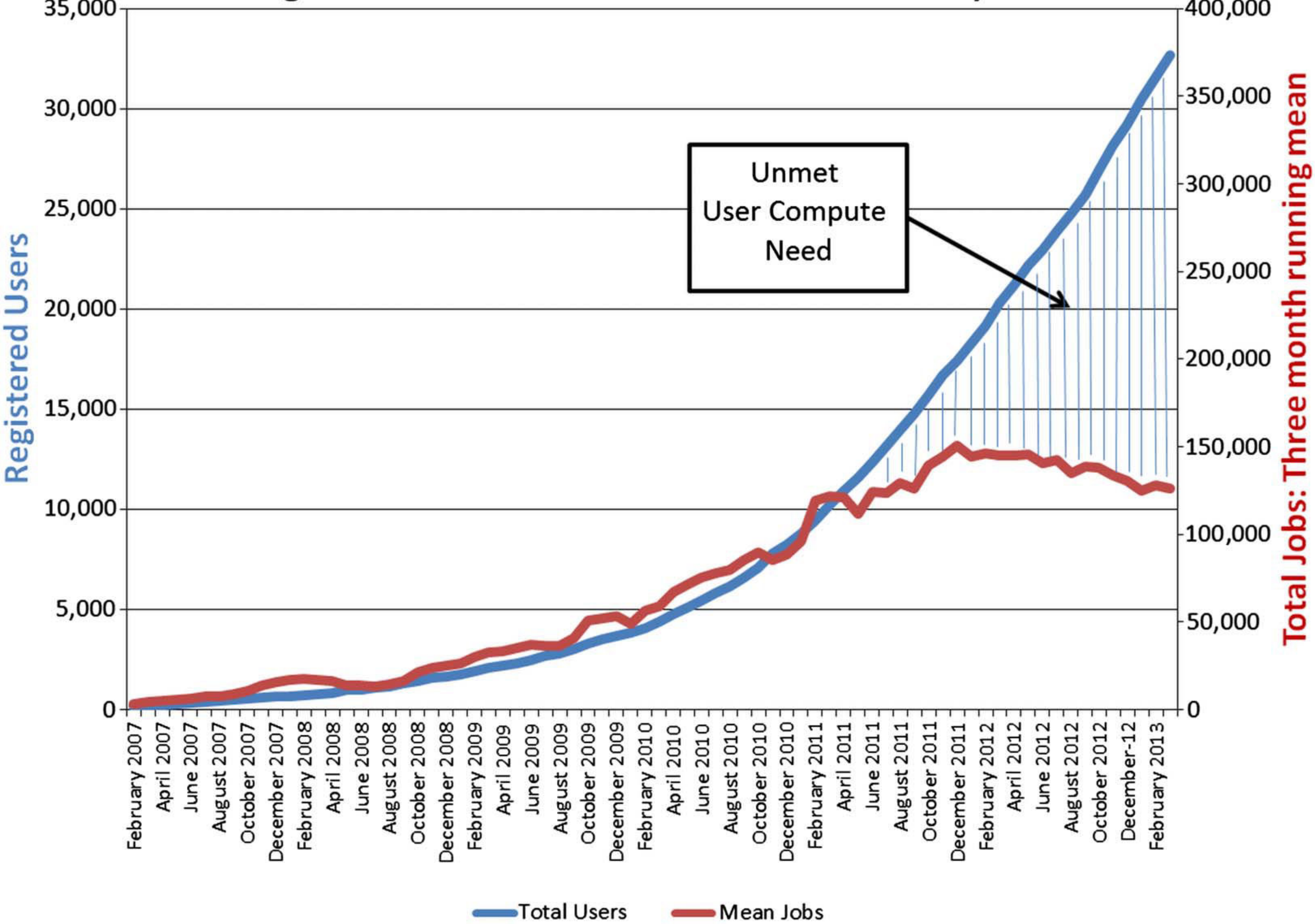
Public web site that anyone can use for free

~1,200 new users, ~20 TB of user data uploaded, and ~180,000 analysis jobs *per month*

Since 2010, disk quotas (250Gb per user) and compute limits (4 concurrent analyses)

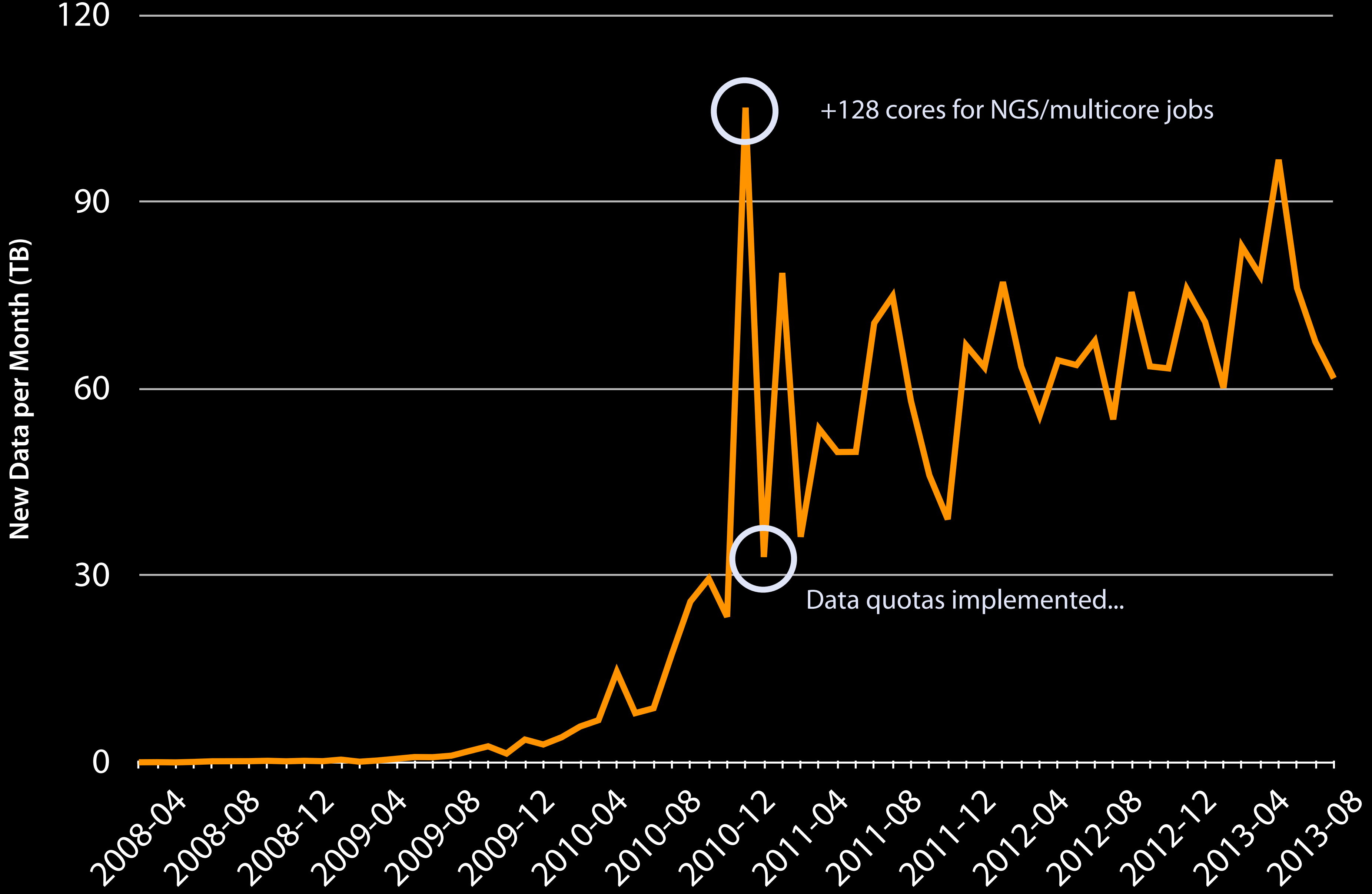
http://usegalaxy.org

Registered Users versus Jobs Submitted on Galaxy Main



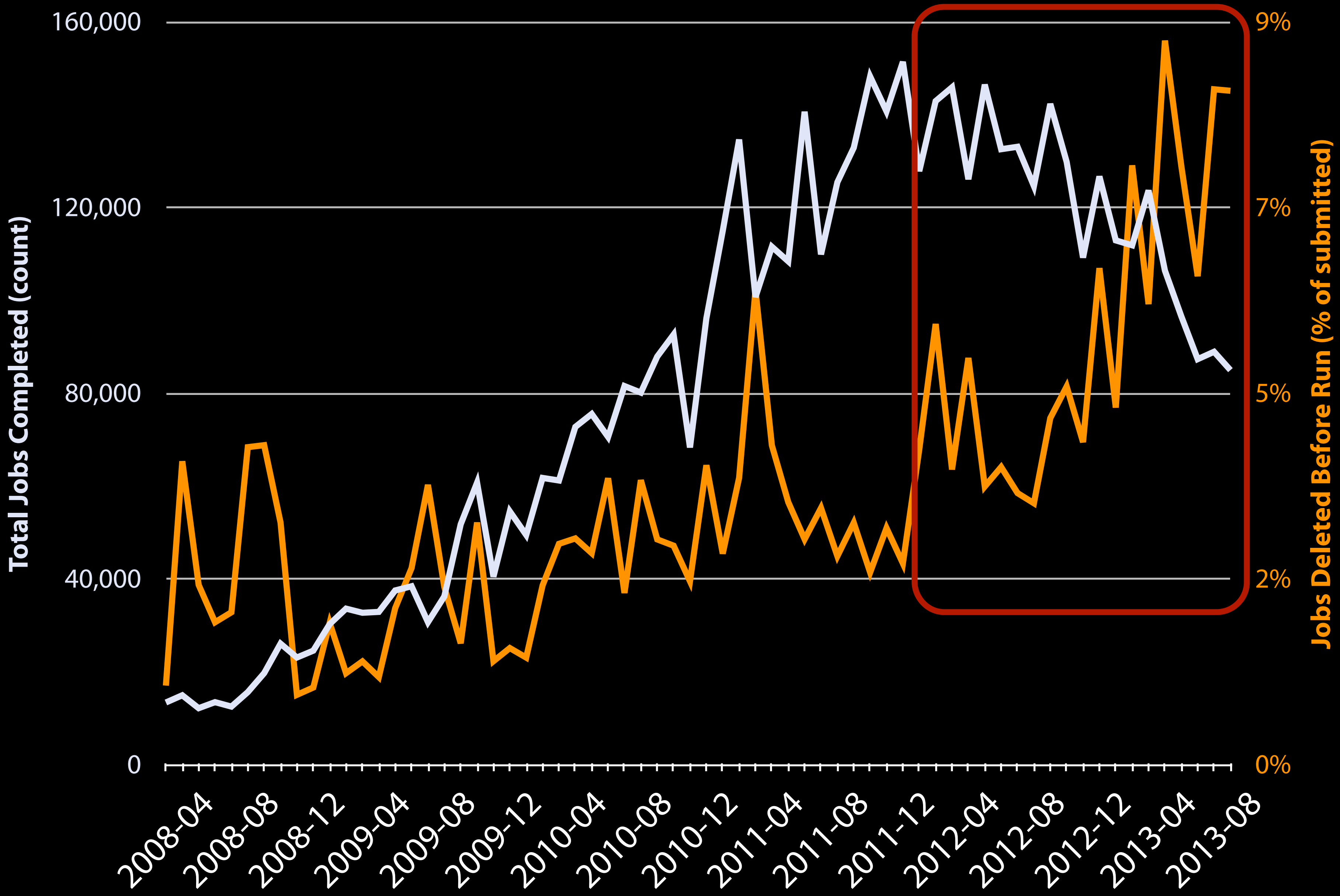
Unmet User Compute Need

usegalaxy.org data growth



Nate Coraor

usegalaxy.org frustration growth



Nate Coraor

Scaling plan one: Decentralize!

Local Galaxy Deployment

Galaxy is designed for local installation and customization... just download and run

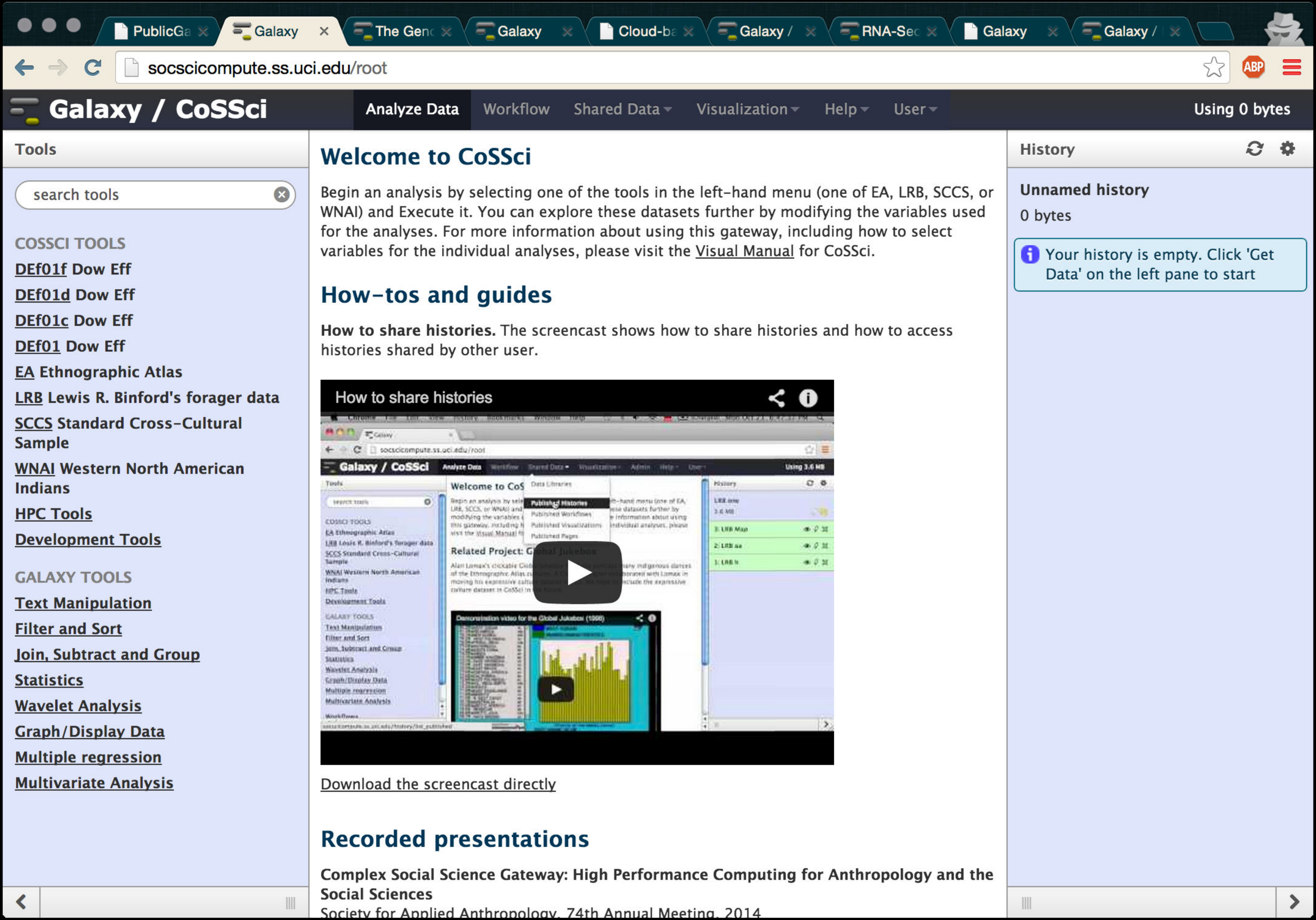
Pluggable interfaces to compute resources, easily connect to one or more existing clusters

Ideally, allow users to take advantage of whatever computational resources they already have access to.

More than 60 known public Galaxy servers

Ballaxy for structure based computational biology,
Cistrome for regulatory sequence analysis,
Genomic Hyperbrowser: statistical integration of genomic data,
GigaGalaxy: integrating workflows published in GigaScience,
Pathogen Portal:comparative analysis of host response to pathogens,
...

Dozens of large scale private Galaxy instances



COSSCI TOOLS

DEf01f Dow Eff

DEf01d Dow Eff

DEf01c Dow Eff

DEf01 Dow Eff

EA Ethnographic Atlas

LRB Lewis R. Binford's forager data

SCCS Standard Cross-Cultural Sample

WNAI Western North American Indians

HPC Tools

Development Tools

GALAXY TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Statistics

Wavelet Analysis

Graph/Display Data

Multiple regression

Multivariate Analysis

Welcome to CoSSci

Begin an analysis by selecting one of the tools in the left-hand menu (one of EA, LRB, SCCS, or WNAI) and Execute it. You can explore these datasets further by modifying the variables used for the analyses. For more information about using this gateway, including how to select variables for the individual analyses, please visit the Visual Manual for CoSSci.

How-tos and guides

How to share histories. The screencast shows how to share histories and how to access histories shared by other user.

How to share histories

Download the screencast directly

Recorded presentations

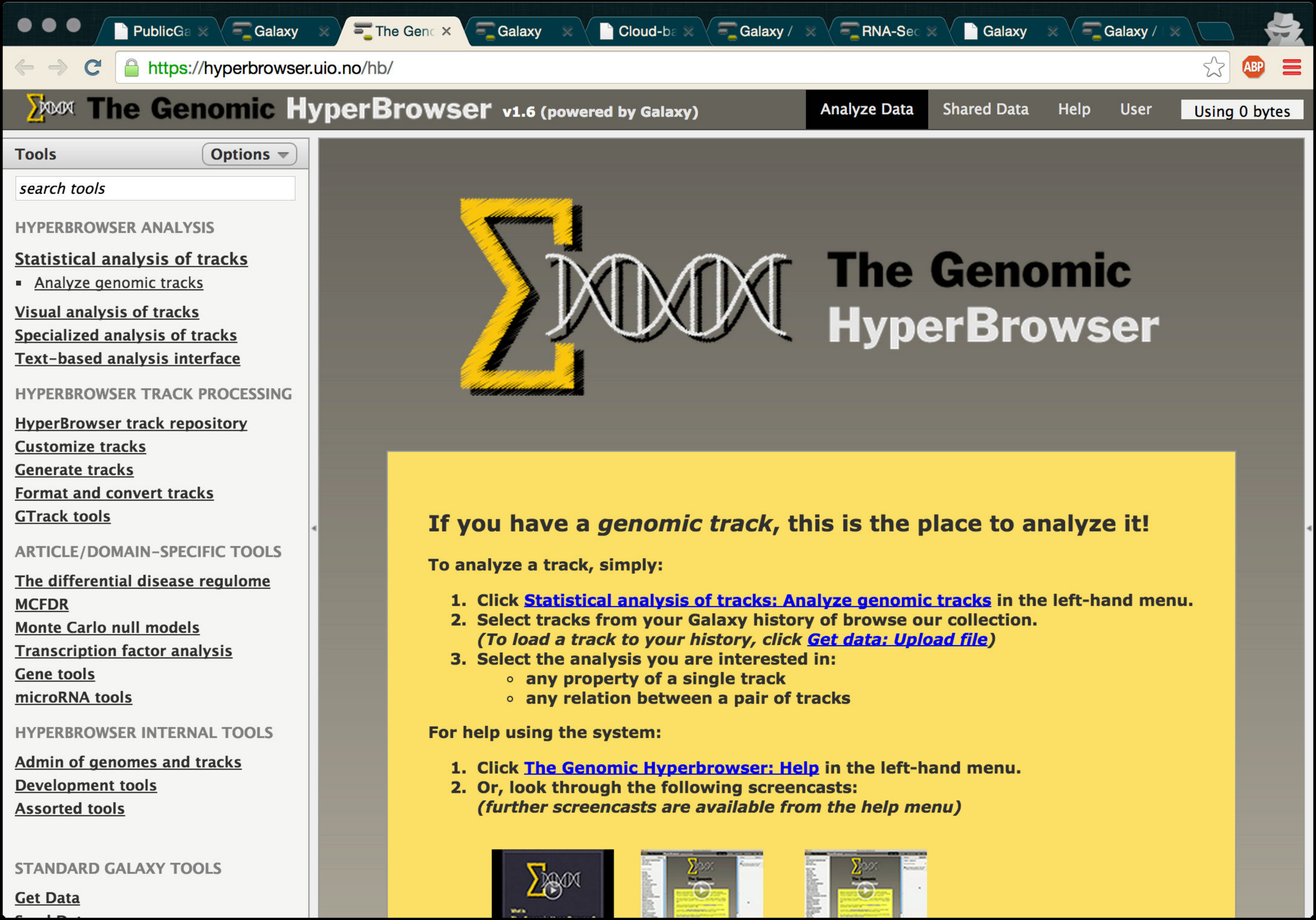
Complex Social Science Gateway: High Performance Computing for Anthropology and the Social Sciences

Society for Applied Anthropology, 74th Annual Meeting, 2014

Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start



HYPERBROWSER ANALYSIS

- Statistical analysis of tracks
  - Analyze genomic tracks
- Visual analysis of tracks
- Specialized analysis of tracks
- Text-based analysis interface

HYPERBROWSER TRACK PROCESSING

- HyperBrowser track repository
- Customize tracks
- Generate tracks
- Format and convert tracks
- GTrack tools

ARTICLE/DOMAIN-SPECIFIC TOOLS

- The differential disease regulome
- MCFDR
- Monte Carlo null models
- Transcription factor analysis
- Gene tools
- microRNA tools

HYPERBROWSER INTERNAL TOOLS

- Admin of genomes and tracks
- Development tools
- Assorted tools

STANDARD GALAXY TOOLS

- Get Data

The Genomic HyperBrowser

If you have a *genomic track*, this is the place to analyze it!

To analyze a track, simply:

1. Click Statistical analysis of tracks: Analyze genomic tracks in the left-hand menu.
2. Select tracks from your Galaxy history of browse our collection.
   (To load a track to your history, click Get data: Upload file)
3. Select the analysis you are interested in:
   - any property of a single track
   - any relation between a pair of tracks

For help using the system:

1. Click The Genomic Hyperbrowser: Help in the left-hand menu.
2. Or, look through the following screencasts:
   (further screencasts are available from the help menu)

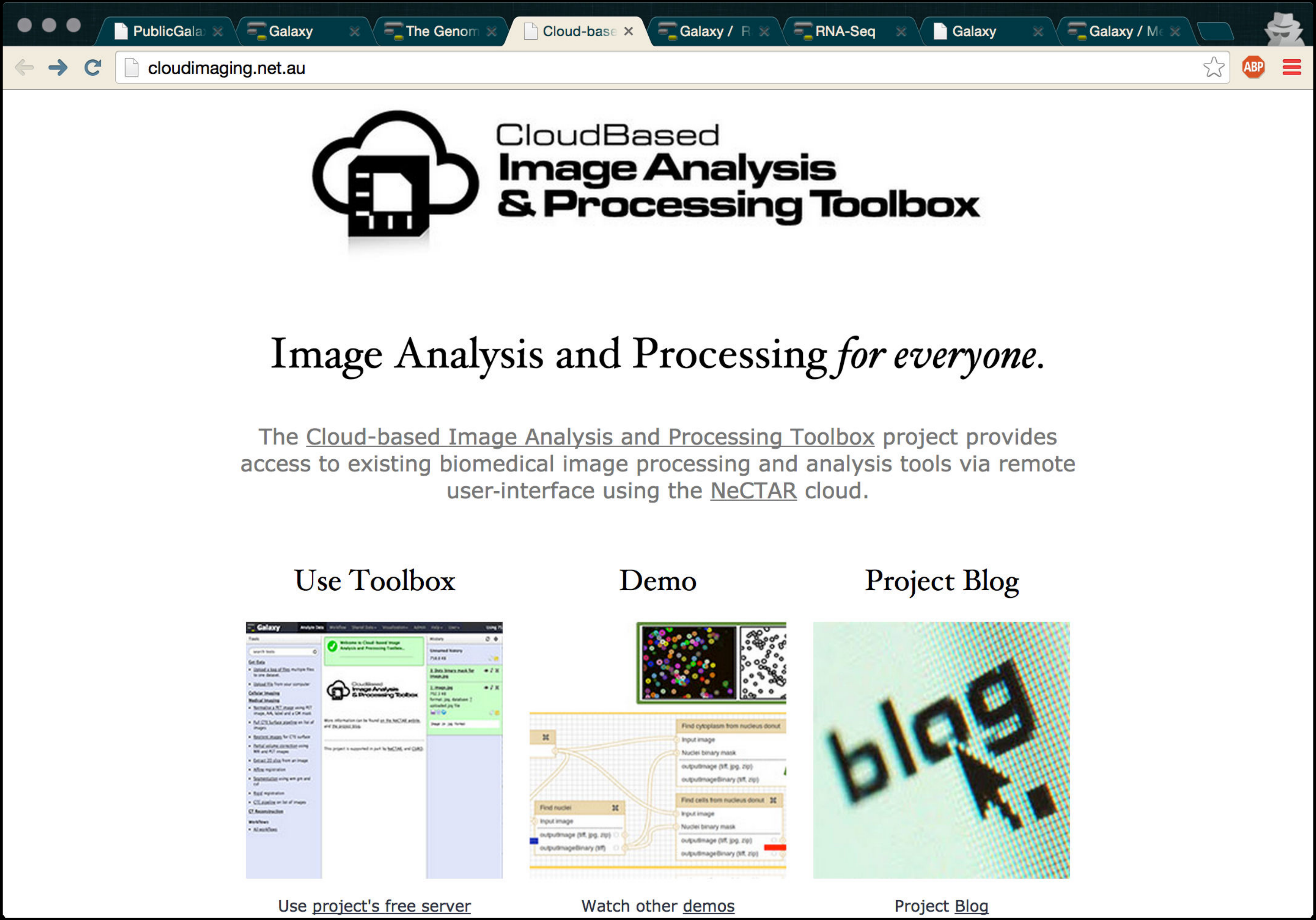


Image Analysis and Processing *for everyone*.

The Cloud-based Image Analysis and Processing Toolbox project provides access to existing biomedical image processing and analysis tools via remote user-interface using the NeCTAR cloud.

Use Toolbox

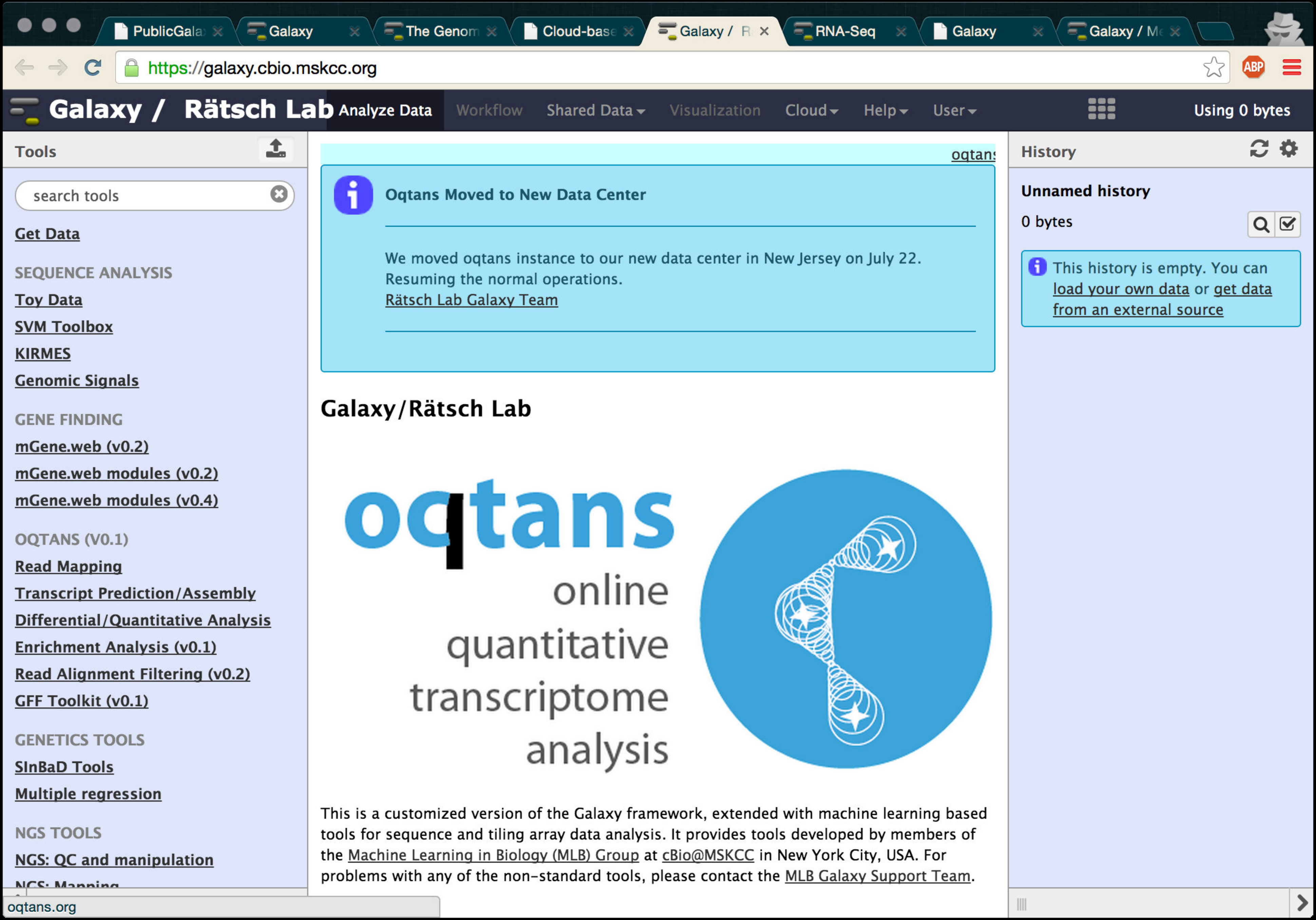
Use project's free server

Demo

Watch other demos

Project Blog

Project Blog



Get Data

SEQUENCE ANALYSIS

Toy Data
SVM Toolbox
KIRMES
Genomic Signals

GENE FINDING

mGene.web (v0.2)
mGene.web modules (v0.2)
mGene.web modules (v0.4)

OQTANS (V0.1)

Read Mapping
Transcript Prediction/Assembly
Differential/Quantitative Analysis
Enrichment Analysis (v0.1)
Read Alignment Filtering (v0.2)
GFF Toolkit (v0.1)

GENETICS TOOLS

SInBaD Tools
Multiple regression

NGS TOOLS

NGS: QC and manipulation
NGS: Mapping

Oqtans Moved to New Data Center

We moved oqtans instance to our new data center in New Jersey on July 22. Resuming the normal operations.
Rätsch Lab Galaxy Team

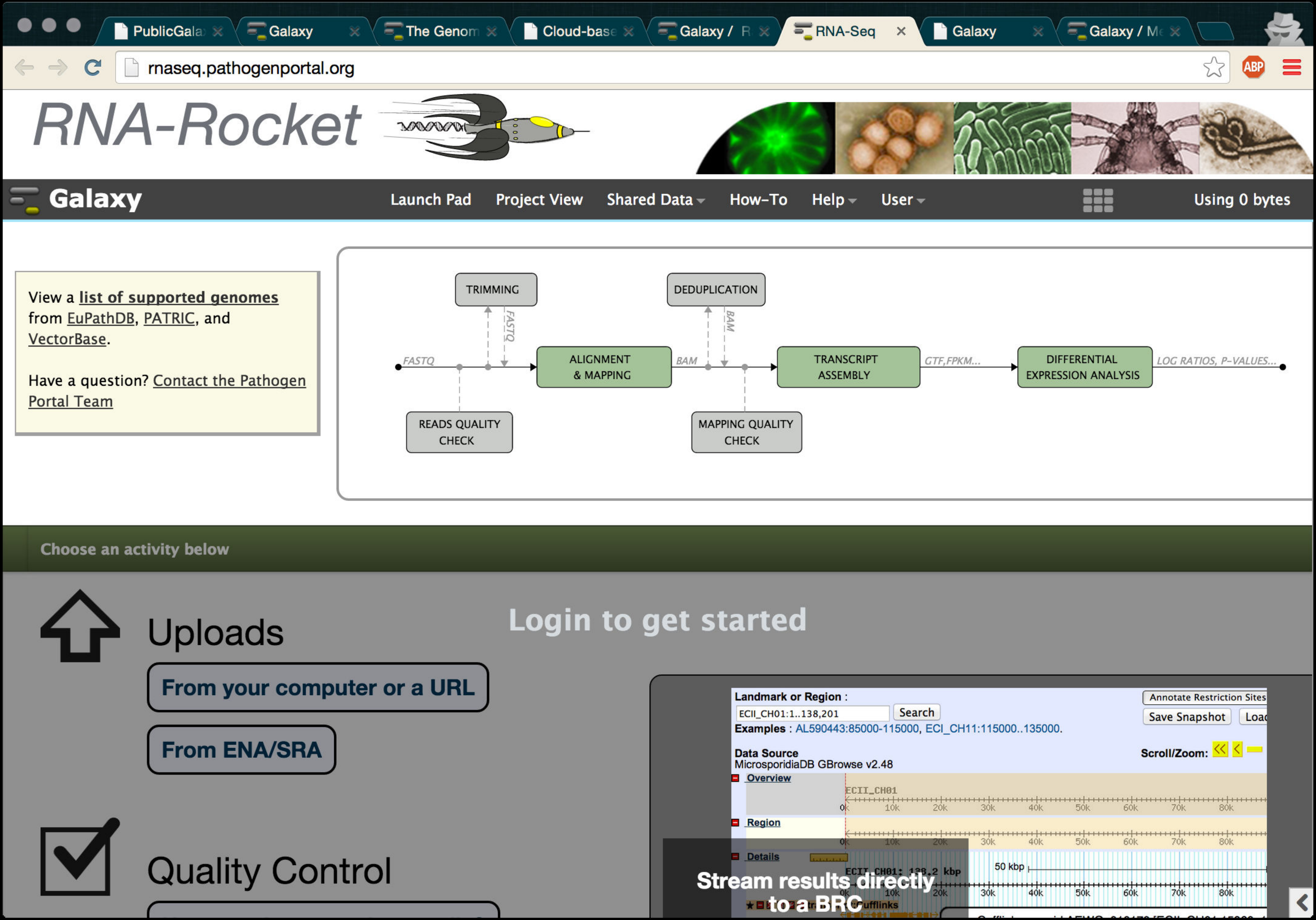
Galaxy/Rätsch Lab

This is a customized version of the Galaxy framework, extended with machine learning based tools for sequence and tiling array data analysis. It provides tools developed by members of the Machine Learning in Biology (MLB) Group at cBio@MSKCC in New York City, USA. For problems with any of the non-standard tools, please contact the MLB Galaxy Support Team.

Unnamed history

0 bytes

This history is empty. You can load your own data or get data from an external source



View a list of supported genomes from EuPathDB, PATRIC, and VectorBase.

Have a question? Contact the Pathogen Portal Team

Examples : AL590443:85000-115000, ECI_CH11:115000..135000.

Data Source
MicrosporidiaDB GBrowse v2.48

Welcome to **Galaxy** on the Cloud(s)

CloudMan: a general purpose deployment manager for ANY cloud

CloudMan

amazon
web services™

nectar

Google
Cloud Platform

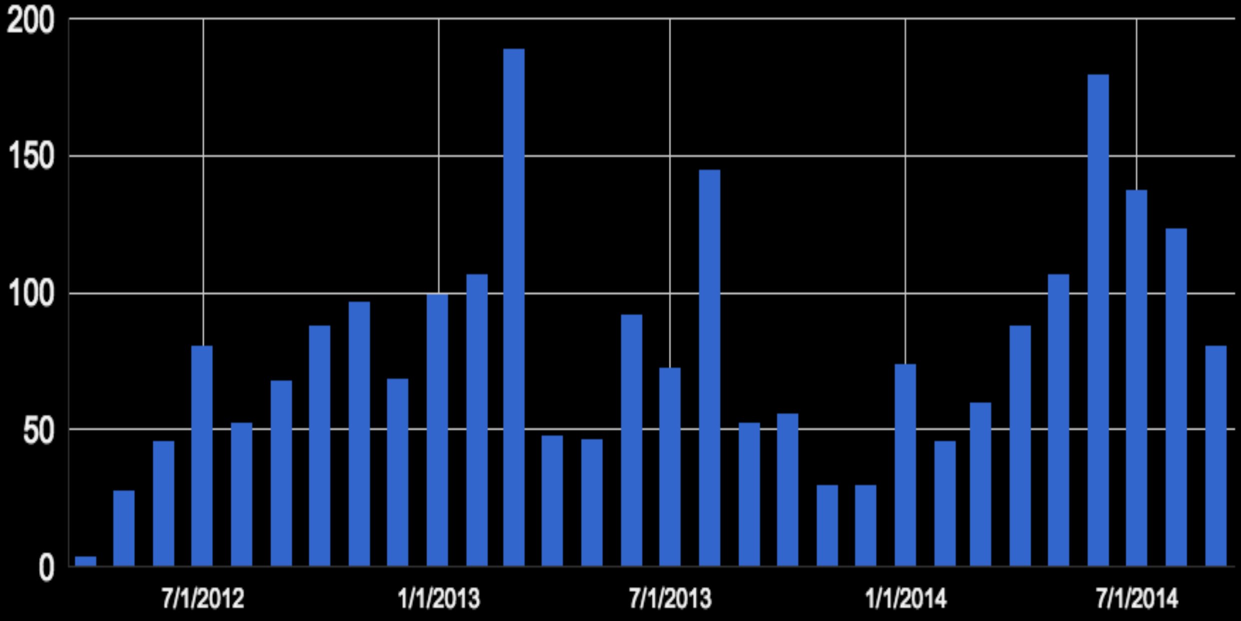
iPlant
Collaborative™

Atmosphere

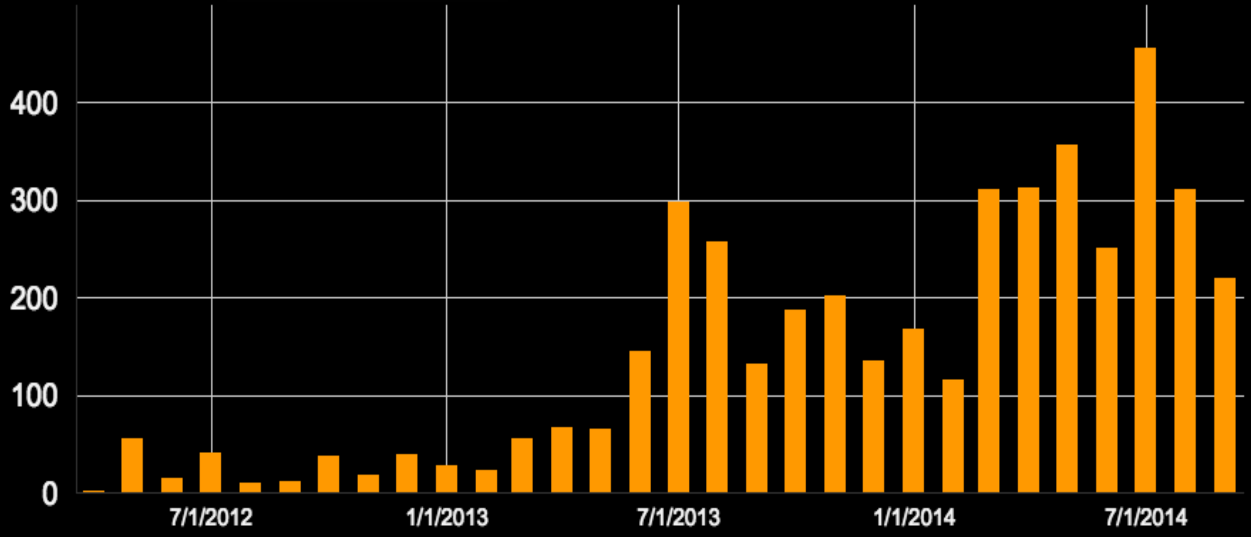
Enis Afgan, Dannon Baker

CloudMan

amazon web services™



nectar



Enis Afgan, Dannon Baker

That was a great plan!

…but users still want one easy to use gateway

Scaling plan two: beg, borrow, steal!

TACC

TEXAS ADVANCED COMPUTING CENTER

PITTSBURGH
SUPERCOMPUTING
CENTER

XSEDE

Extreme Science and Engineering
Discovery Environment

iPlant
Collaborative™

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

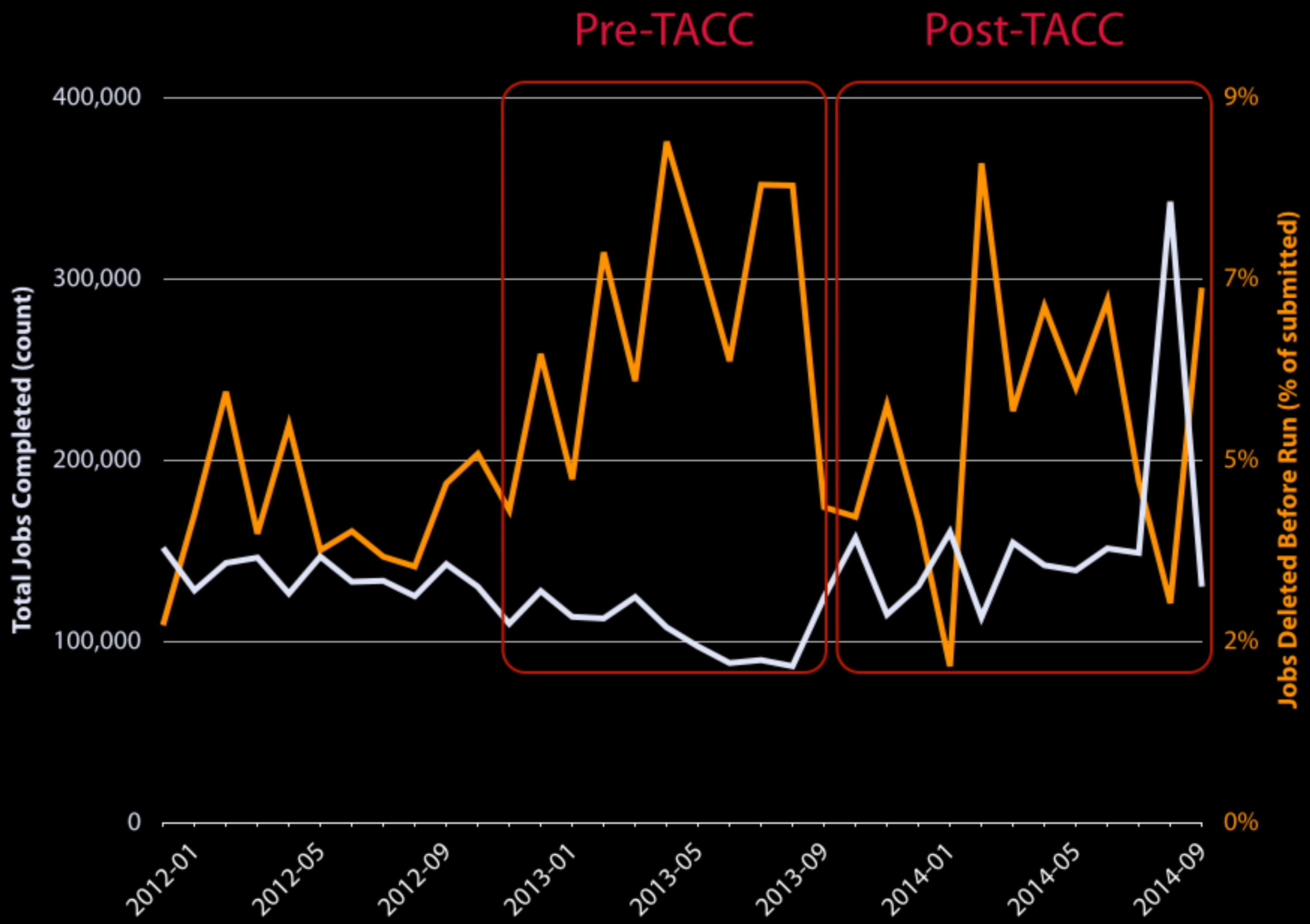
The Texas Advanced Computing Center (TACC) has already built substantial infrastructure in the context of the iPlant project

(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

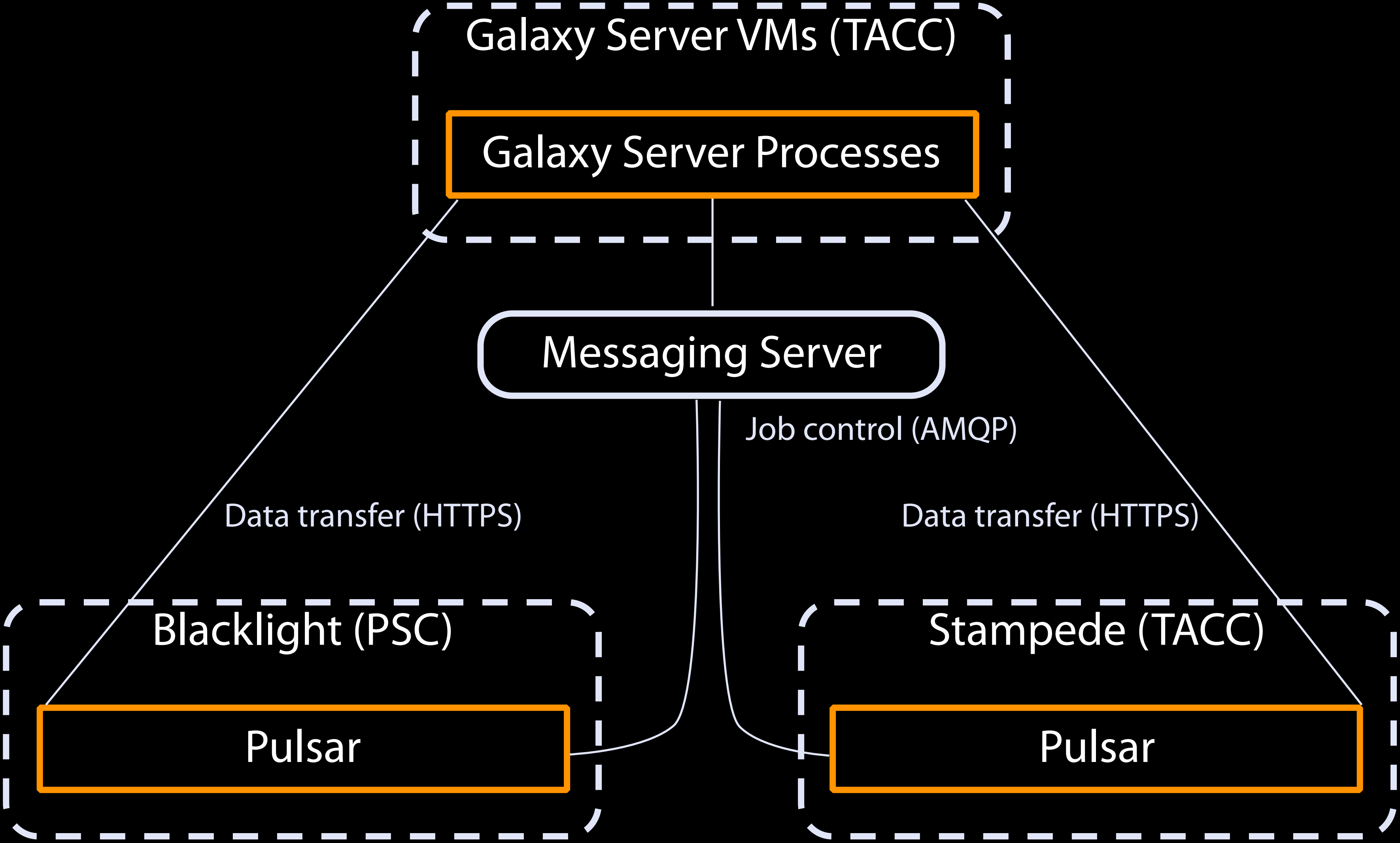
Since October 2013 Galaxy Main has run from TACC

usegalaxy.org frustration growth



Still not enough!

Pulsar: Galaxy job runner that can run almost anywhere. No shared filesystem, stages all necessary Galaxy components


Galaxy Server VMs (TACC)

Galaxy Server Processes

Messaging Server

Job control (AMQP)

Data transfer (HTTPS)

Data transfer (HTTPS)

Blacklight (PSC)

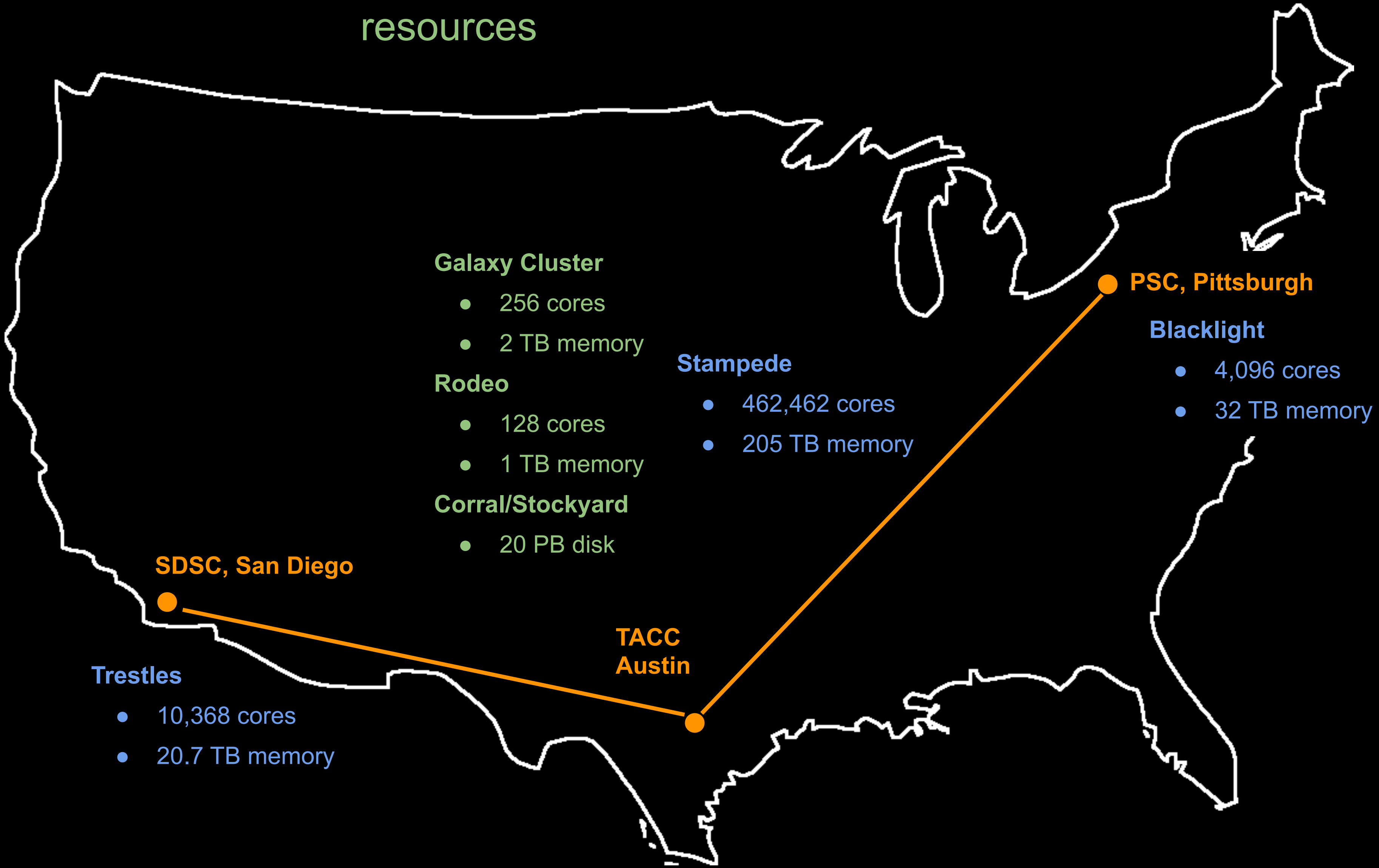
Pulsar

Stampede (TACC)

Pulsar

John Chilton

- Dedicated resources
- Shared resources



SDSC, San Diego

Trestles

- 10,368 cores
- 20.7 TB memory

TACC Austin

Galaxy Cluster

- 256 cores
- 2 TB memory

Rodeo

- 128 cores
- 1 TB memory

Corral/Stockyard

- 20 PB disk

Stampede

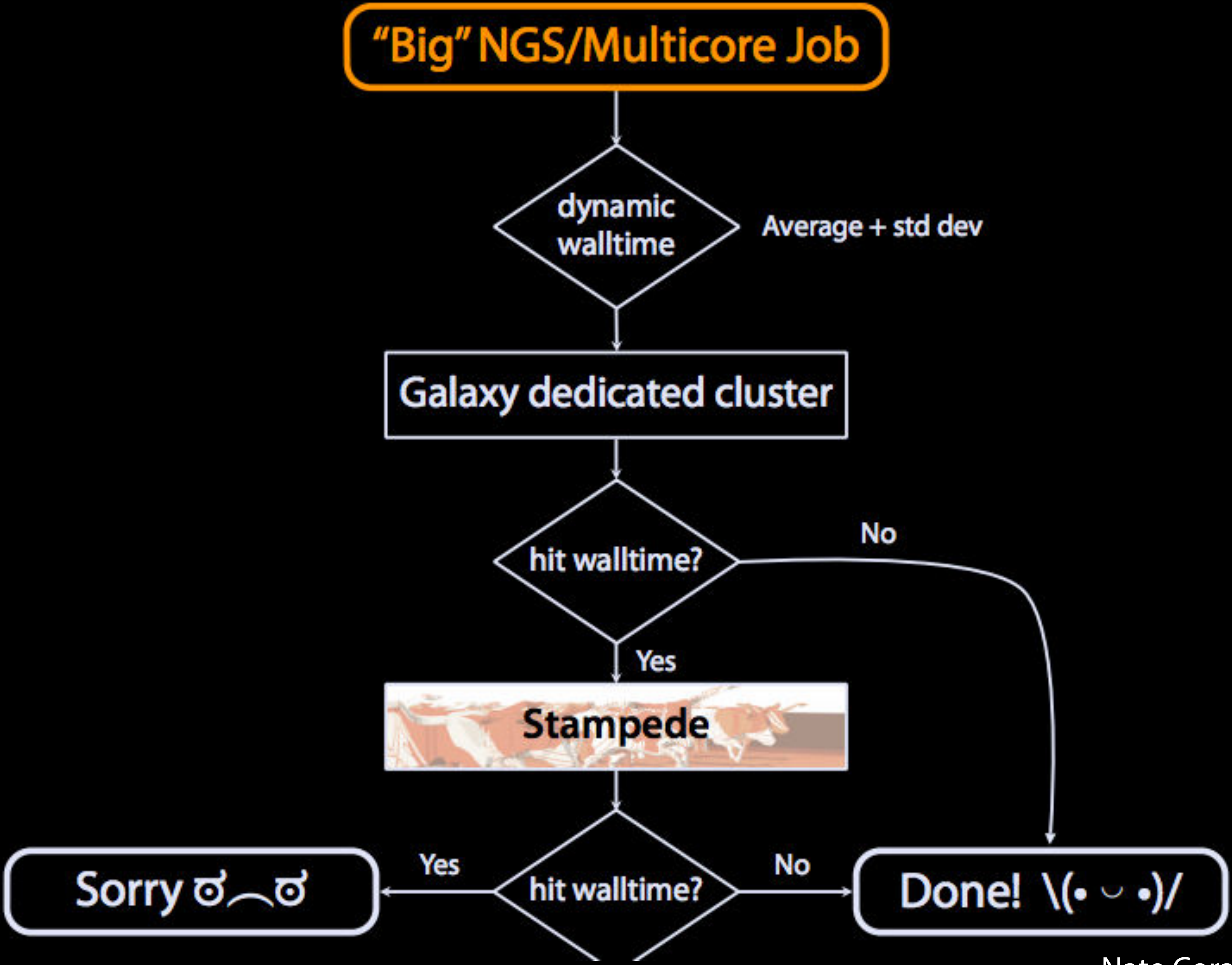
- 462,462 cores
- 205 TB memory

PSC, Pittsburgh

Blacklight

- 4,096 cores
- 32 TB memory

Nate Coraor



"Big" NGS/Multicore Job

dynamic walltime

Average + std dev

Galaxy dedicated cluster

hit walltime?

No

Yes

Stampede

Yes

hit walltime?

No

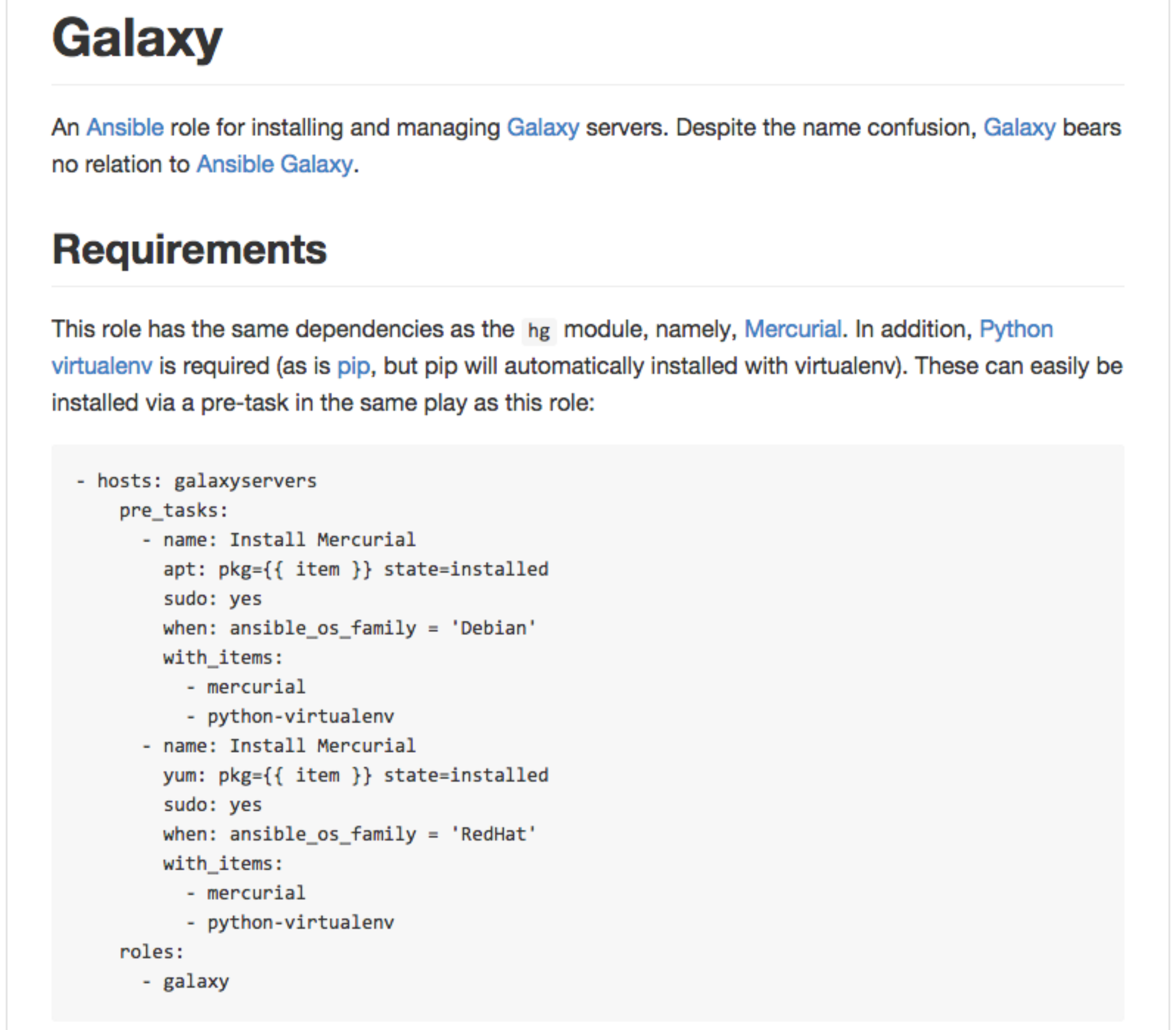
Sorry ಠ︵ಠ

Done! \(• ◡ •)/

Nate Coraor

Result: No waiting for jobs
to run on usegalaxy.org!

(for now…)



Galaxy

An Ansible role for installing and managing Galaxy servers. Despite the name confusion, Galaxy bears no relation to Ansible Galaxy.

Requirements

This role has the same dependencies as the `hg` module, namely, Mercurial. In addition, Python virtualenv is required (as is pip, but pip will automatically installed with virtualenv). These can easily be installed via a pre-task in the same play as this role:

```
- hosts: galaxyservers
    pre_tasks:
      - name: Install Mercurial
        apt: pkg={{ item }} state=installed
        sudo: yes
        when: ansible_os_family = 'Debian'
        with_items:
          - mercurial
          - python-virtualenv
      - name: Install Mercurial
        yum: pkg={{ item }} state=installed
        sudo: yes
        when: ansible_os_family = 'RedHat'
        with_items:
          - mercurial
          - python-virtualenv
    roles:
      - galaxy
```

Bringing it all together: automate all the things!
Unified **ansible** playbook for Galaxy main, cloud, and local deployments

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Čech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

James Taylor

http://wiki.galaxyproject.org/GalaxyTeam

Computational Biology, Genomics, and Bioinformatics at Johns Hopkins University

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Biomedical Engineering

Joel Bader
Mike Beer
Rachel Karchin
Steven Salzberg

Computer Science

Alexis Battle
Ben Langmead
Suchi Saria

Applied Math

Don Geman

JOHNS HOPKINS
SCHOOL *of* MEDICINE

Oncology

Elana Fertig
Luigi Marchionni
Robert Scharpf
Sarah Wheelan

Medicine

Lilian Florea
Mihaela Pertea
Jiang Qian

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Biostatistics

Kasper Hansen
Hongkai Ji
Jeff Leek
Ingo Ruczinski
Cristian Tomasetti

JOHNS HOPKINS
KRIEGER SCHOOL
of ARTS & SCIENCES

Biology

James Taylor

http://ccb.jhu.edu

JOHNS HOPKINS
KRIEGER SCHOOL
of ARTS & SCIENCES

JOHNS HOPKINS
SCHOOL *of* MEDICINE

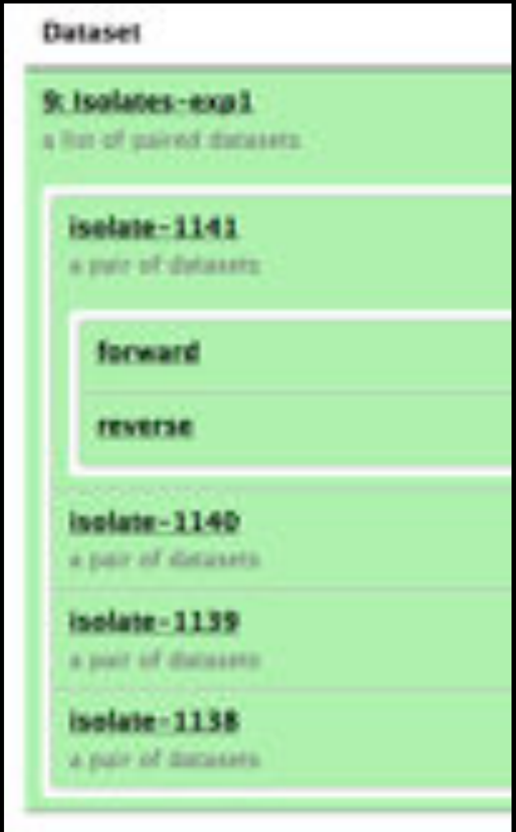
Tenure-Track Faculty Position in Data Intensive Biology

The Department of **Biology** seeks to hire a tenure-track Assistant Professor who applies data intensive approaches to investigate biological problems in creative and innovative ways… Candidates who apply **computational, quantitative, or data intensive methods in any area of Biology** will be considered…

Bloomberg Distinguished Professorship in Evolutionary Genomics.

The Johns Hopkins University is searching for an outstanding senior scientist in the area of **Evolutionary Genomics** for an endowed chair as a Bloomberg Distinguished Professor. This position will be held jointly between the Department of **Biology** (Krieger School of Arts and Sciences) and the **Institute for Genetic Medicine** (JHU School of Medicine).

More Info: http://www.bio.jhu.edu/Events/Jobs/Default.aspx
Or contact me: james@taylorlab.org



Dataset

9: Isolates-exp1
a list of paired datasets

isolate-1141
a pair of datasets

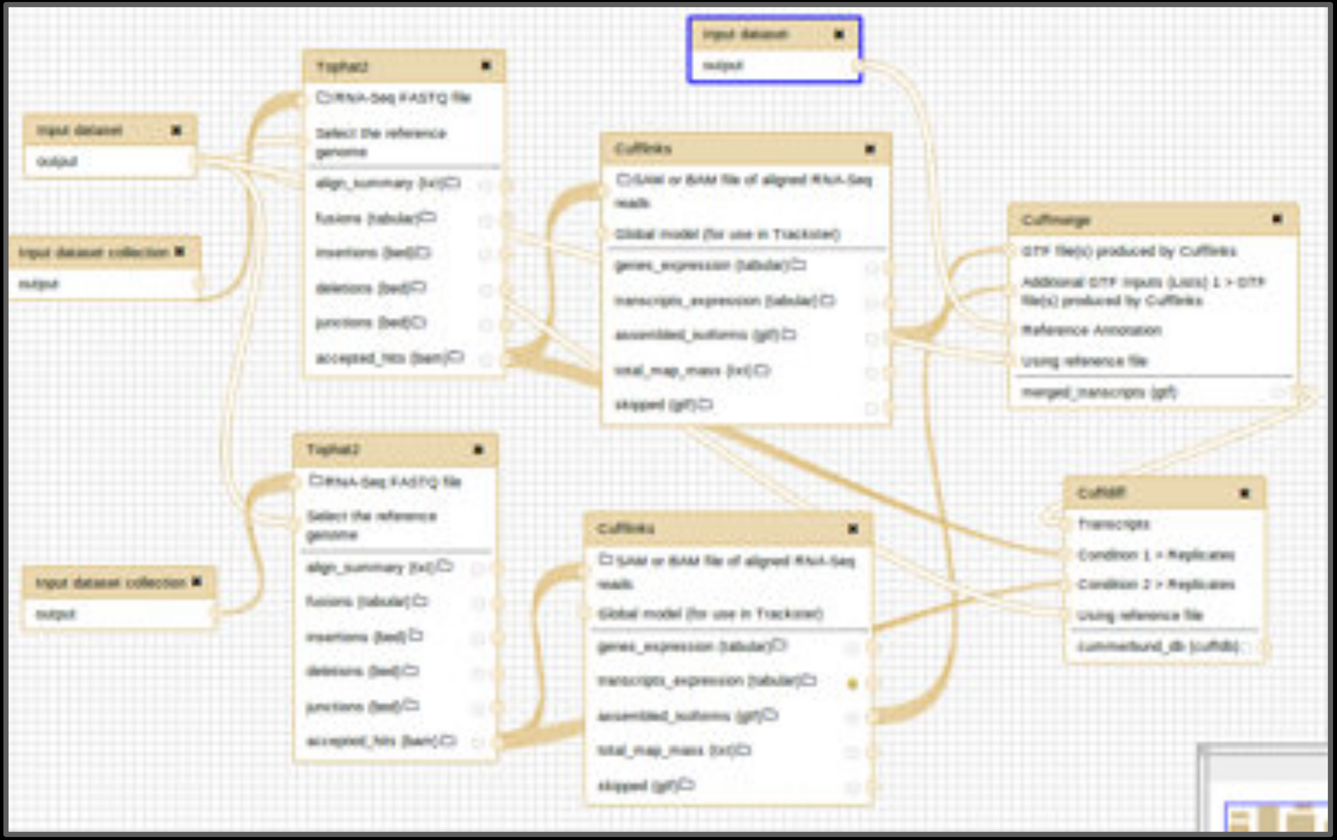
forward

reverse

isolate-1140
a pair of datasets

isolate-1139
a pair of datasets

isolate-1138
a pair of datasets



John Chilton