Galaxy

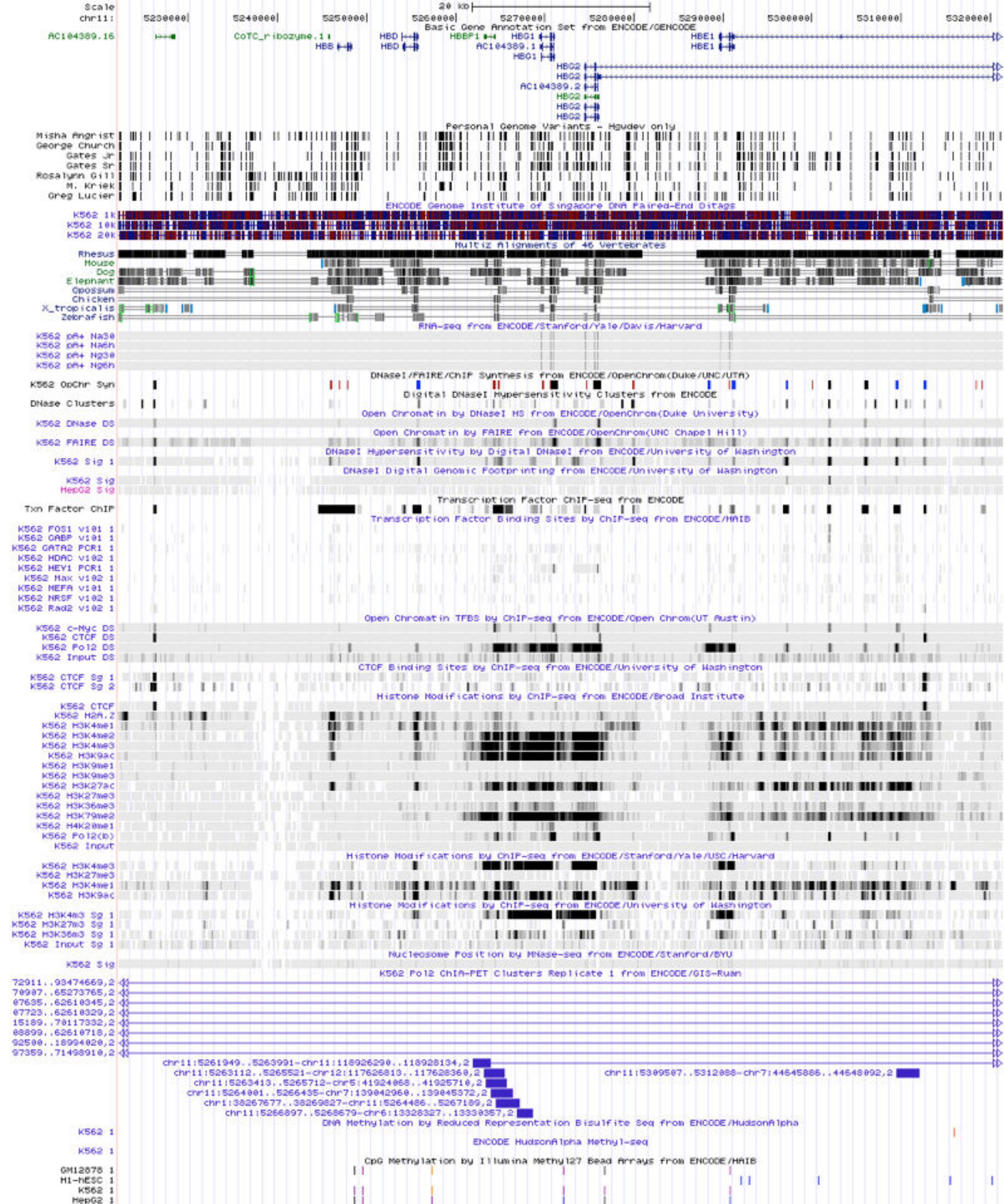
Data intensive biology *for everyone.*

www.galaxyproject.org

@jxtx / #usegalaxy

I ❤ High-Throughput SEQUENCING!

High-throughput sequencing is **transformative**


Scale
chr11:
20 kb
Basic Gene Annotation Set from ENCODE/GENCODE
Personal Genome Variants - Hgwdev only
Misha Angrist
George Church
Gates Jr
Gates Sr
Rosalynn Gill
M. Kriek
Greg Lucier
ENCODE Genome Institute of Singapore DNA Paired-End Ditags
K562 1k
K562 10k
K562 20k
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
Digital DNaseI Hypersensitivity Clusters from ENCODE
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
Transcription Factor ChIP-seq from ENCODE
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
Histone Modifications by ChIP-seq from ENCODE/University of Washington
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
ENCODE HudsonAlpha Methyl-seq
CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB

Resequencing

De novo genome sequencing

Direct RNA sequencing

Open Chromatin assays (DNase, FAIRE)

Transcription factors (ChIP-seq)

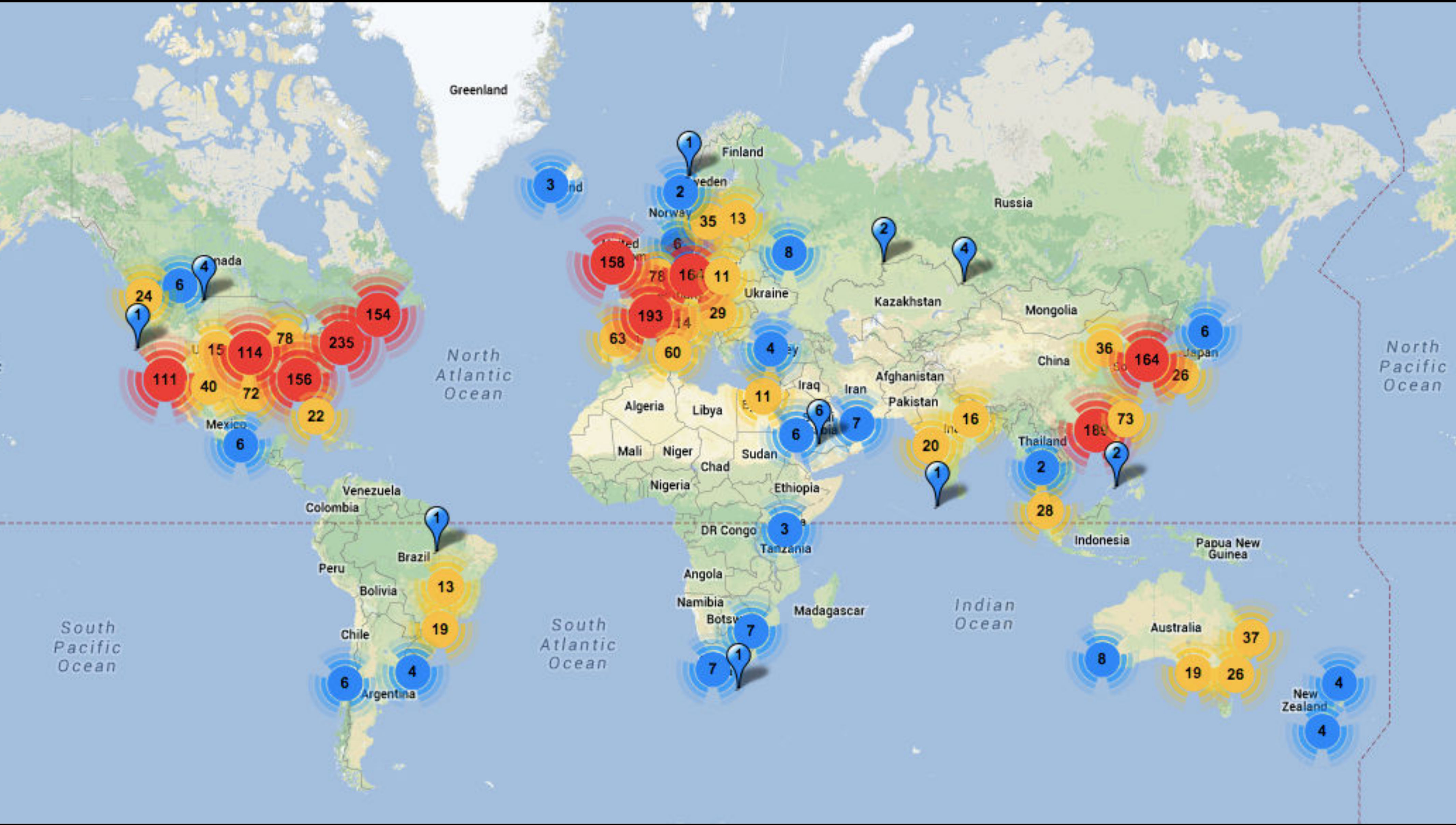
Histones variants (ChIP-seq, MNase-seq)

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

Methylation (Bisulfite-seq)

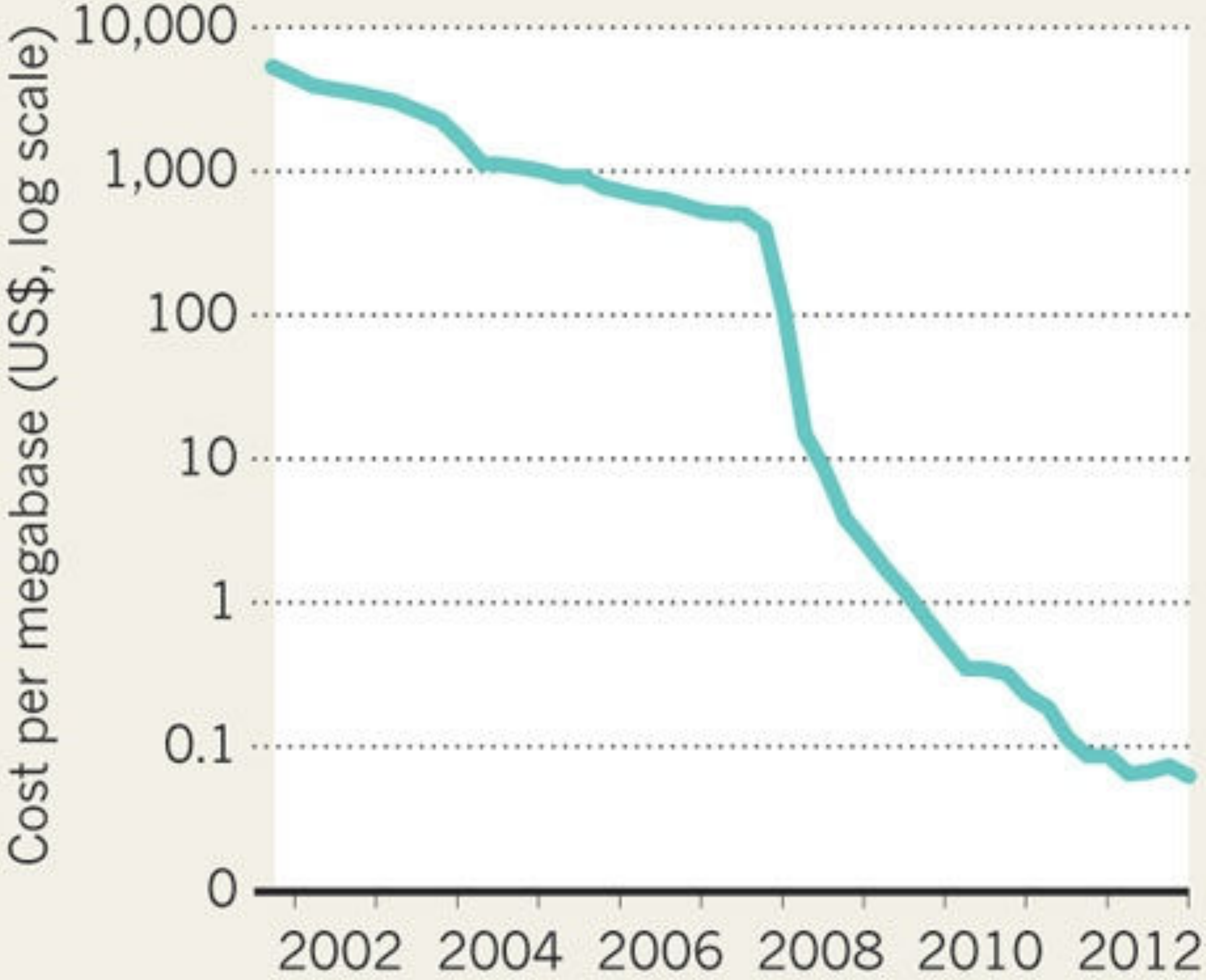
High-throughput sequencing is
democratizing

It is widely available...



(http://omicsmaps.com/)

...and practically free!



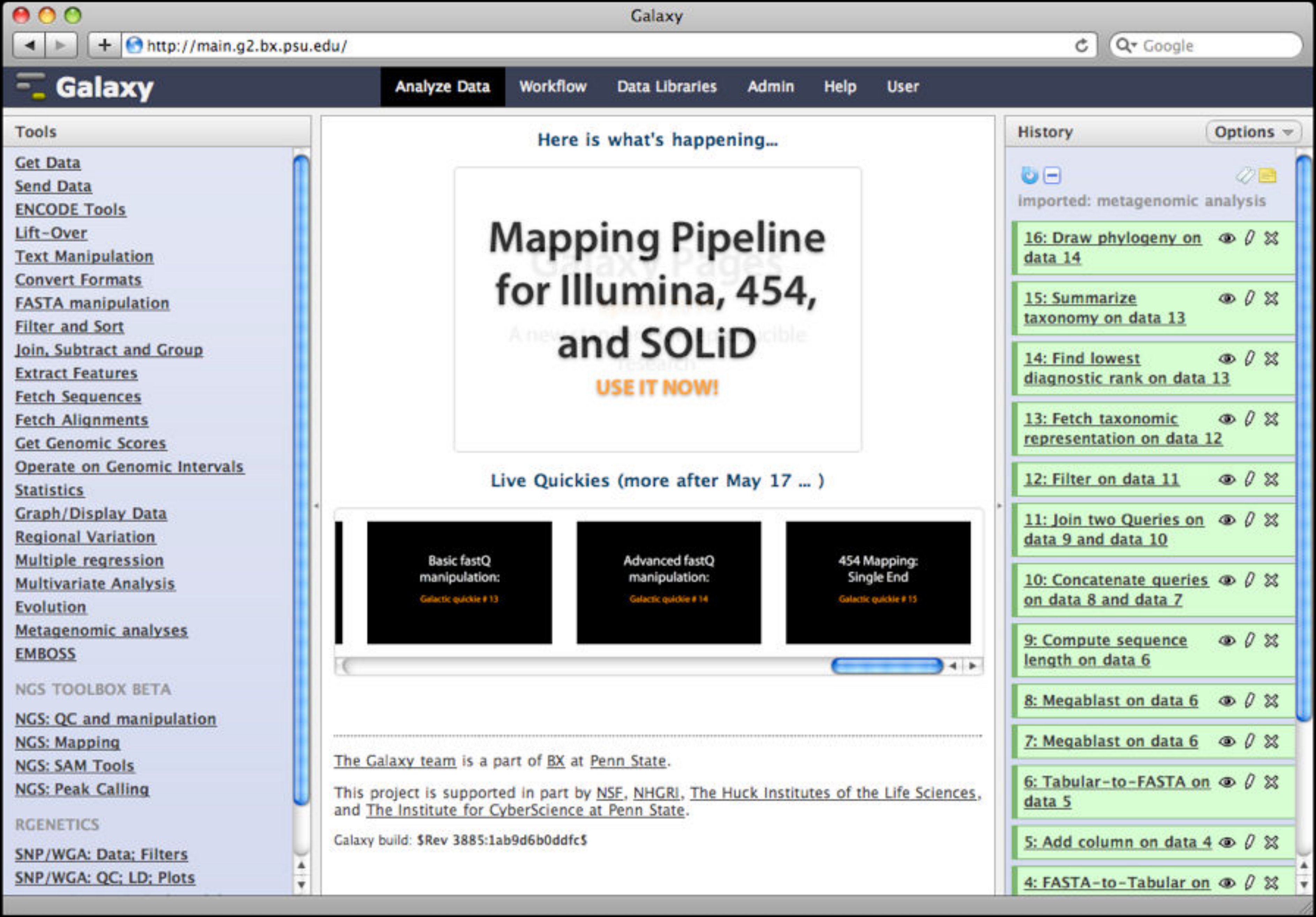
(NHGRI / *Nature* 497:546–547)

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



Here is what's happening...

Live Quickies (more after May 17 ...)

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

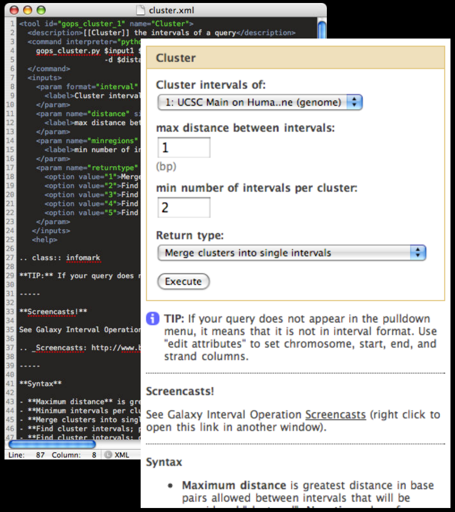
This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

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, ...



```
<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 interva
    </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
```

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

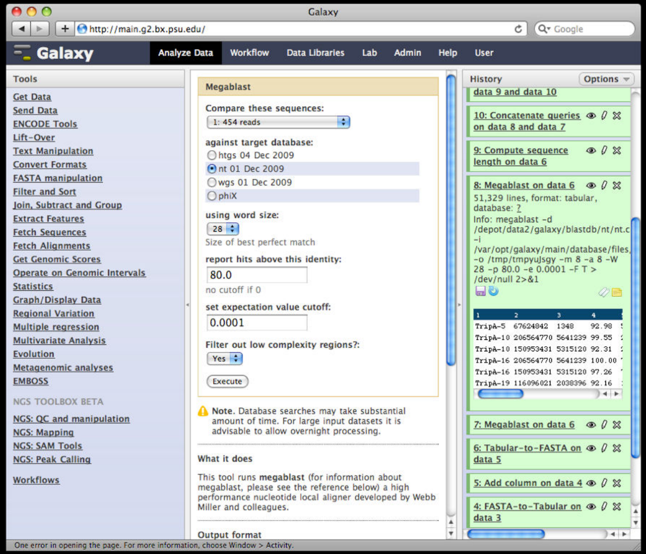
TIP: If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

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



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

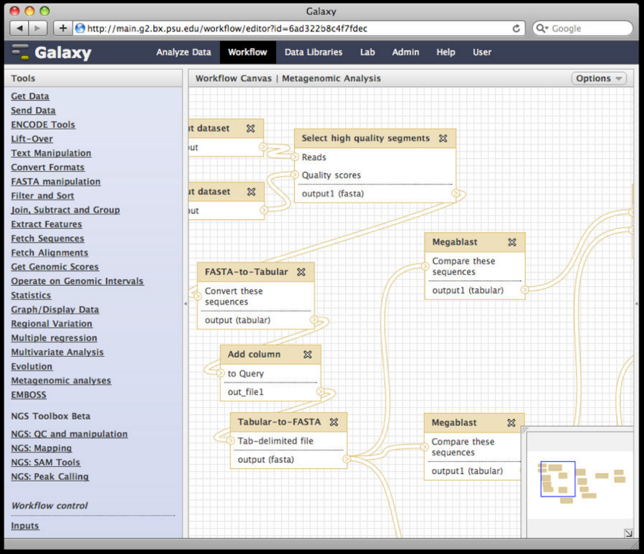
5: Add column on data 4

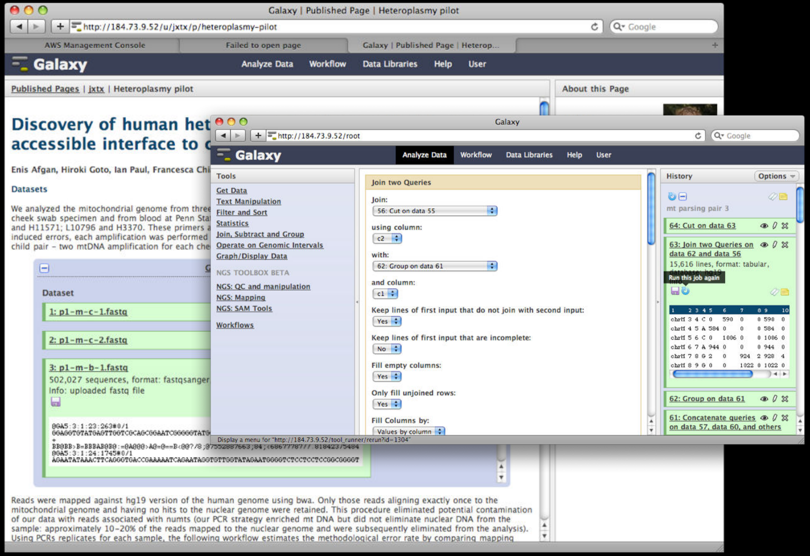
4: FASTA-to-Tabular on data 3

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

Describe analysis tool behavior abstractly

Analysis environment automatically and transparently tracks details





Galaxy | Published Page | Heteroplasmy pilot
http://184.73.9.52/u/jxtx/p/heteroplasmy-pilot
AWS Management Console
Failed to open page
Galaxy | Published Page | Heterop...
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Published Pages | jxtx | Heteroplasmy pilot
About this Page

Discovery of human het
accessible interface to c

Enis Afgan, Hiroki Goto, Ian Paul, Francesca Chi

Datasets

We analyzed the mitochondrial genome from three cheek swab specimen and from blood at Penn Sta 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;@7552887663;84;<6867778777.818423/5484
@GA5:3:1:24:1745#0/1
AGAATATAAACTTCAGGGTGACCGAAAAATCAGAATAGGTGTTGGTATAGAATGGGGTCTCCTCCTCCGGCGGGGT
```

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

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

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
```

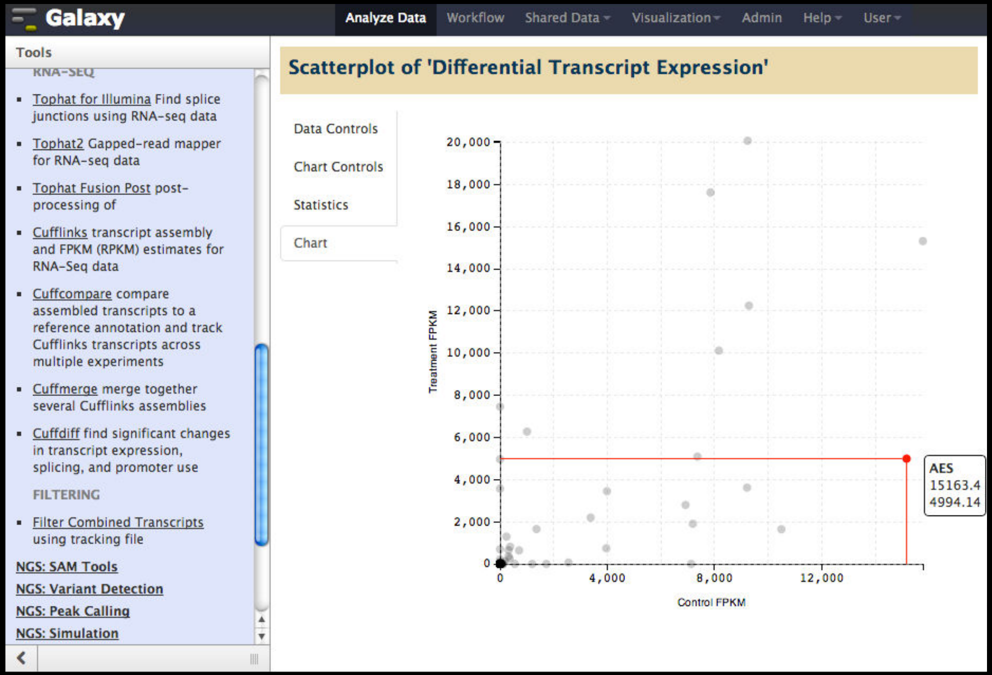
62: Group on data 61

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

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

Workflow system for complex analysis, constructed explicitly or automatically

Pervasive sharing, and publication of documents with integrated analysis



RNA-SEQ

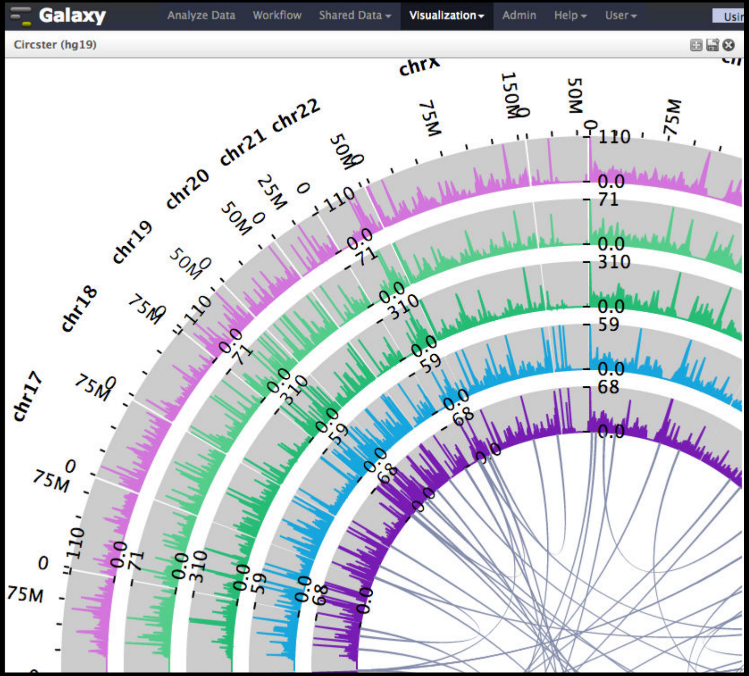
- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat2 Gapped-read mapper for RNA-seq data
- Tophat Fusion Post post-processing of
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track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

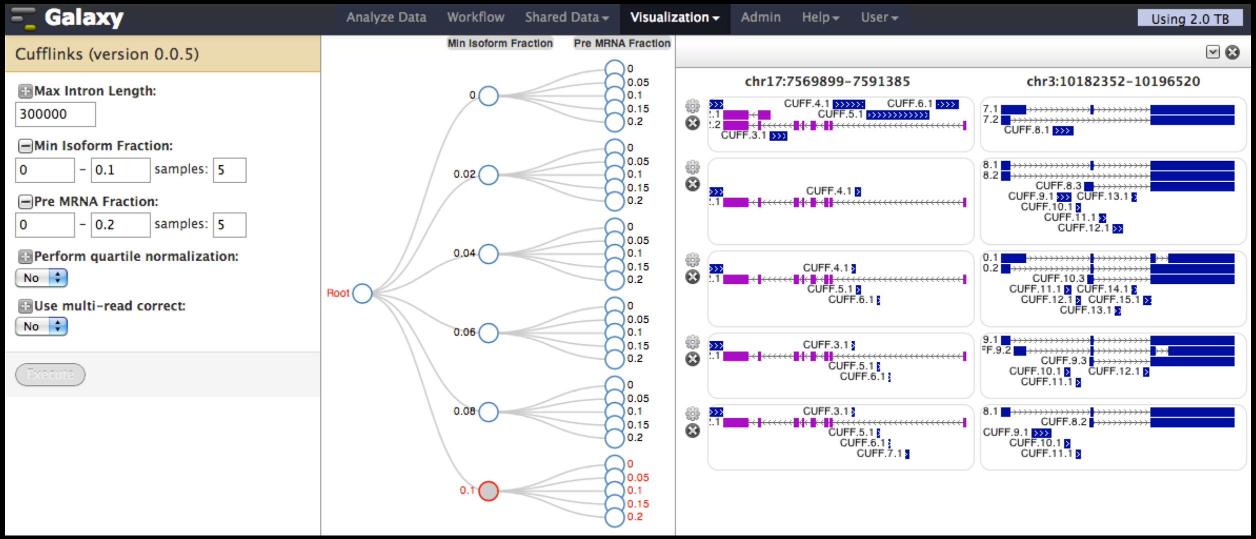
FILTERING

- 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

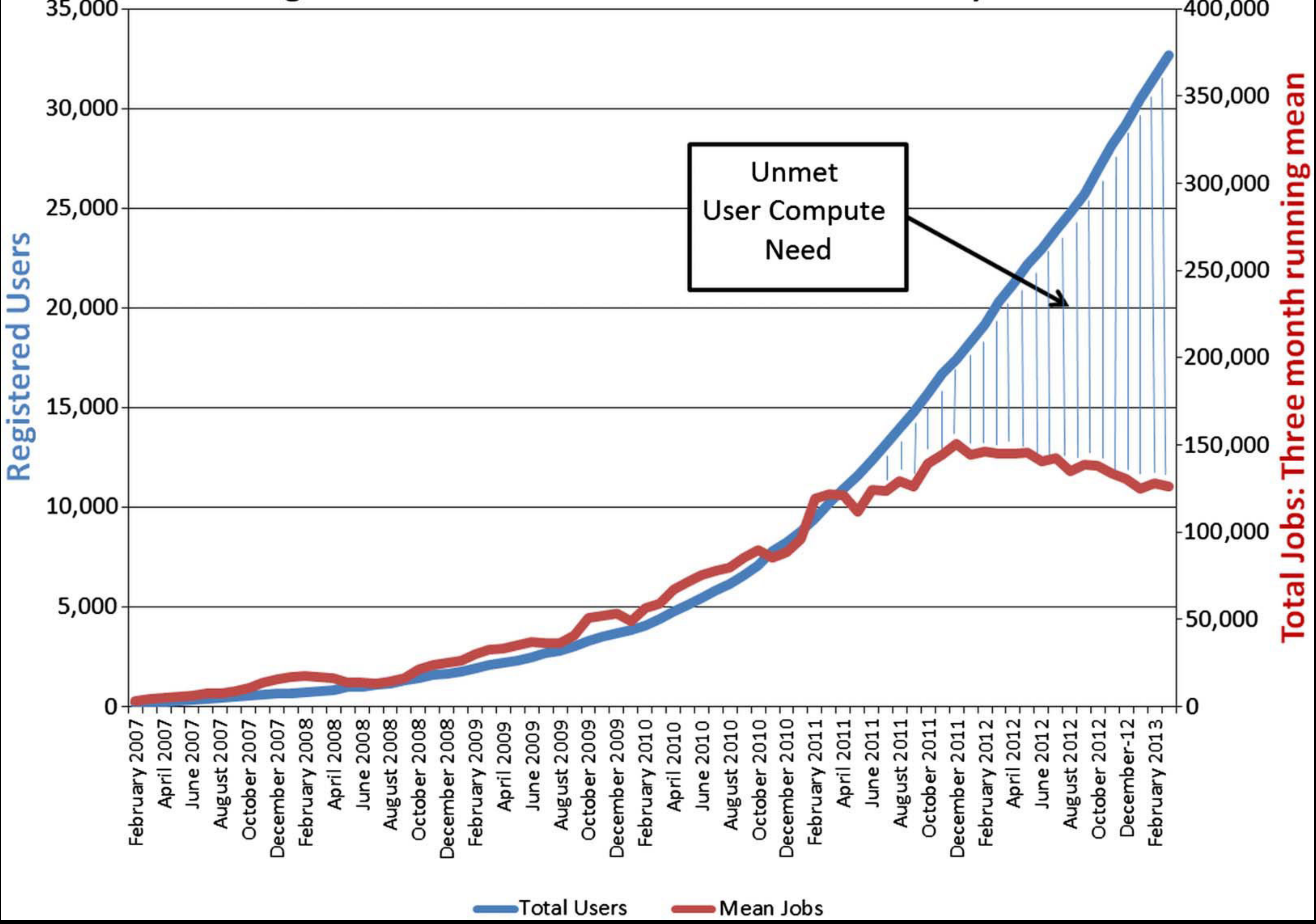
Use multi-read correct: No

Visualization and visual analytics

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

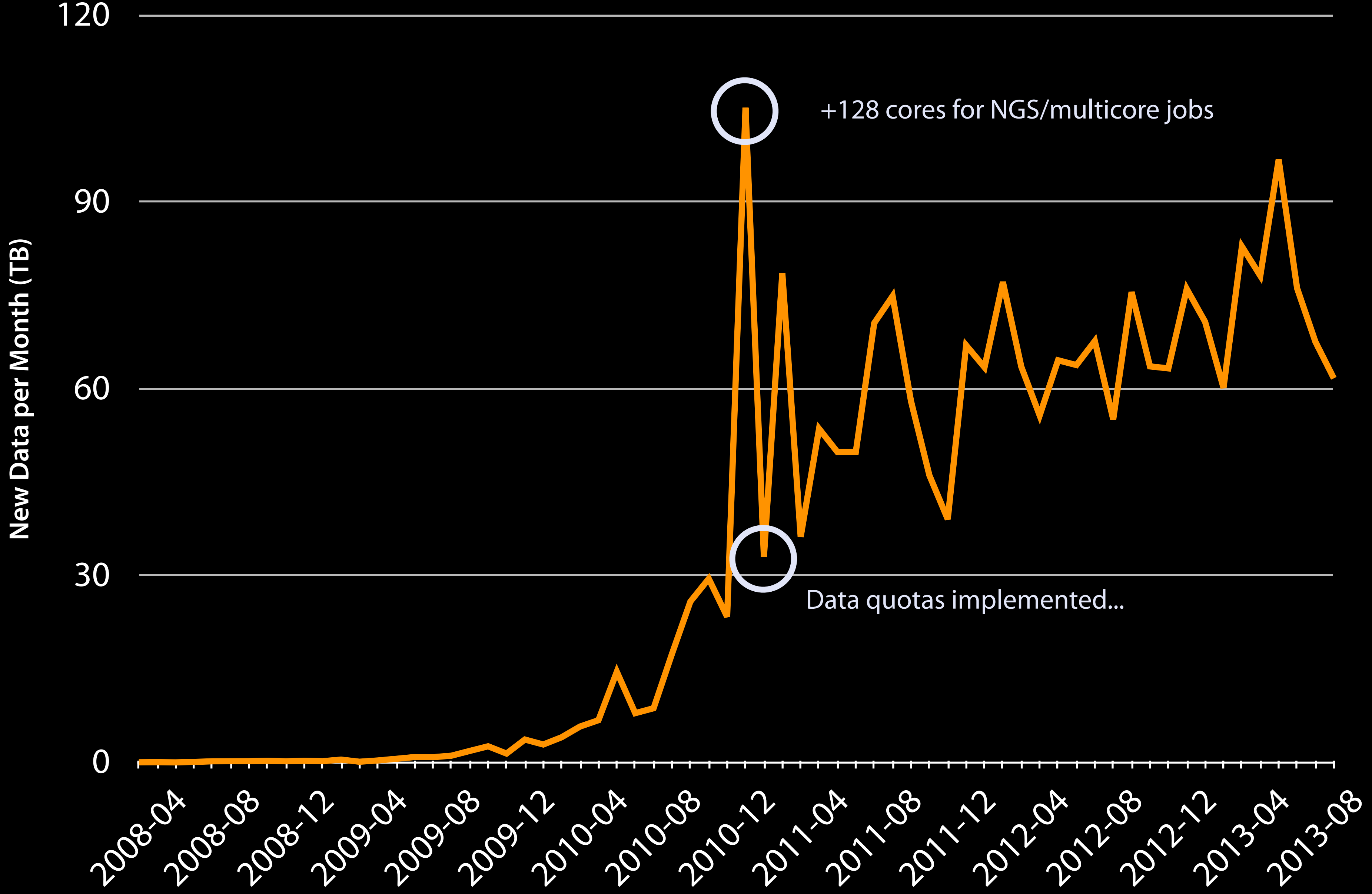
How can we possibly sustain this?

Registered Users versus Jobs Submitted on Galaxy Main



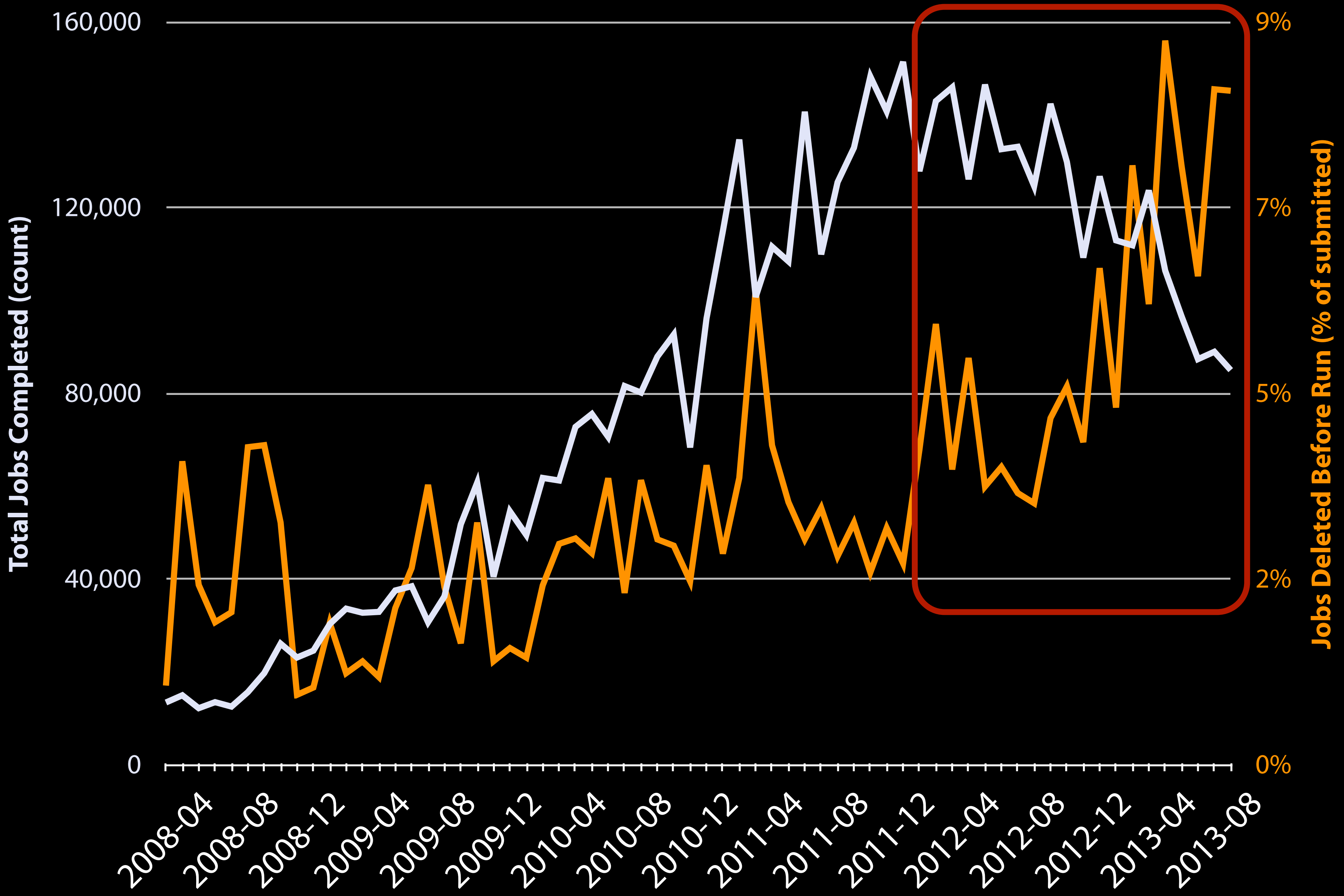
Unmet User Compute Need

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 — not talking about this today)

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

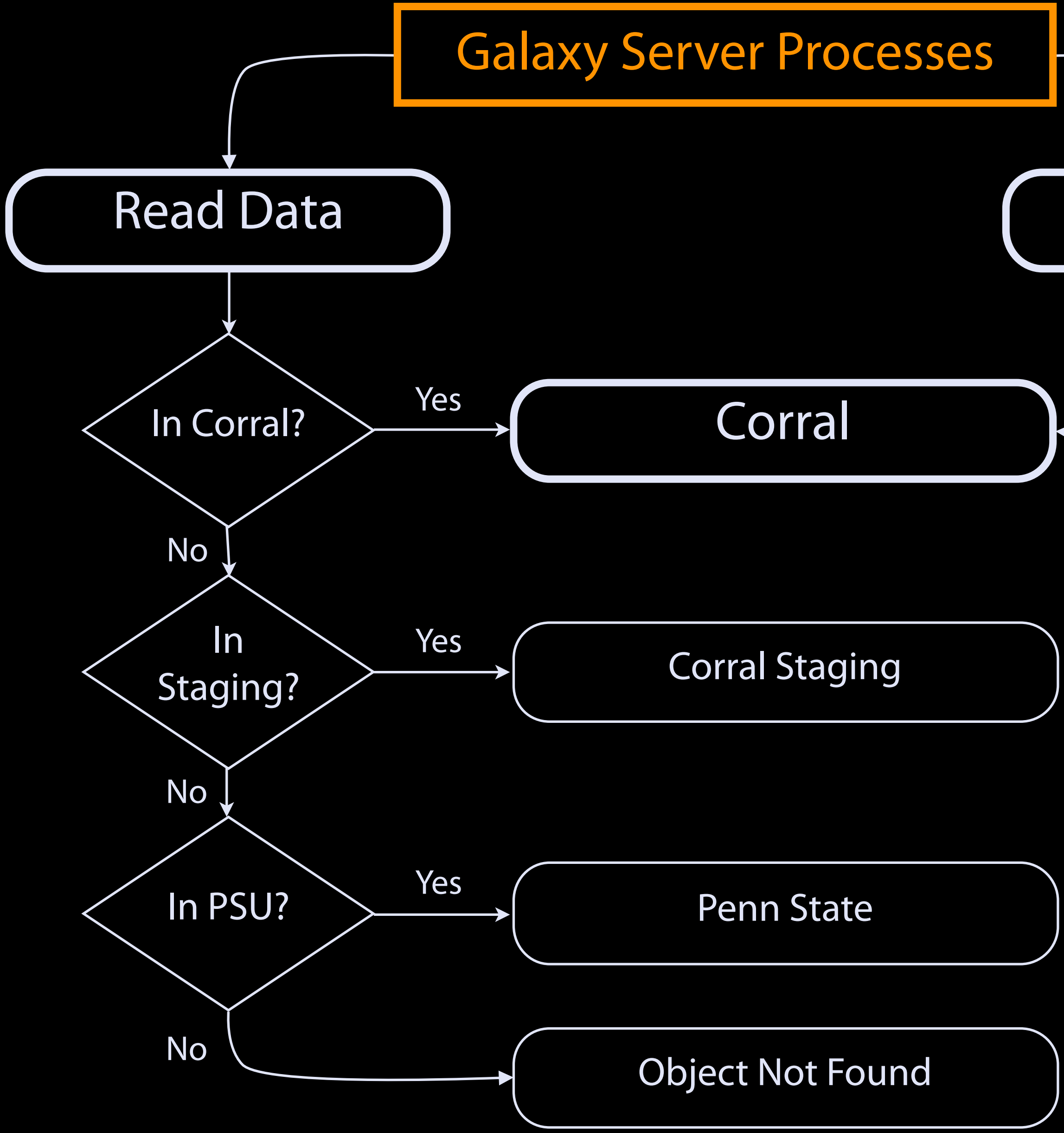
The Texas Advanced Computing Center (TACC) has already built substantial infrastructure in the context of the iPlant project

(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



Nate Coraor

Expanding to more XSEDE resources

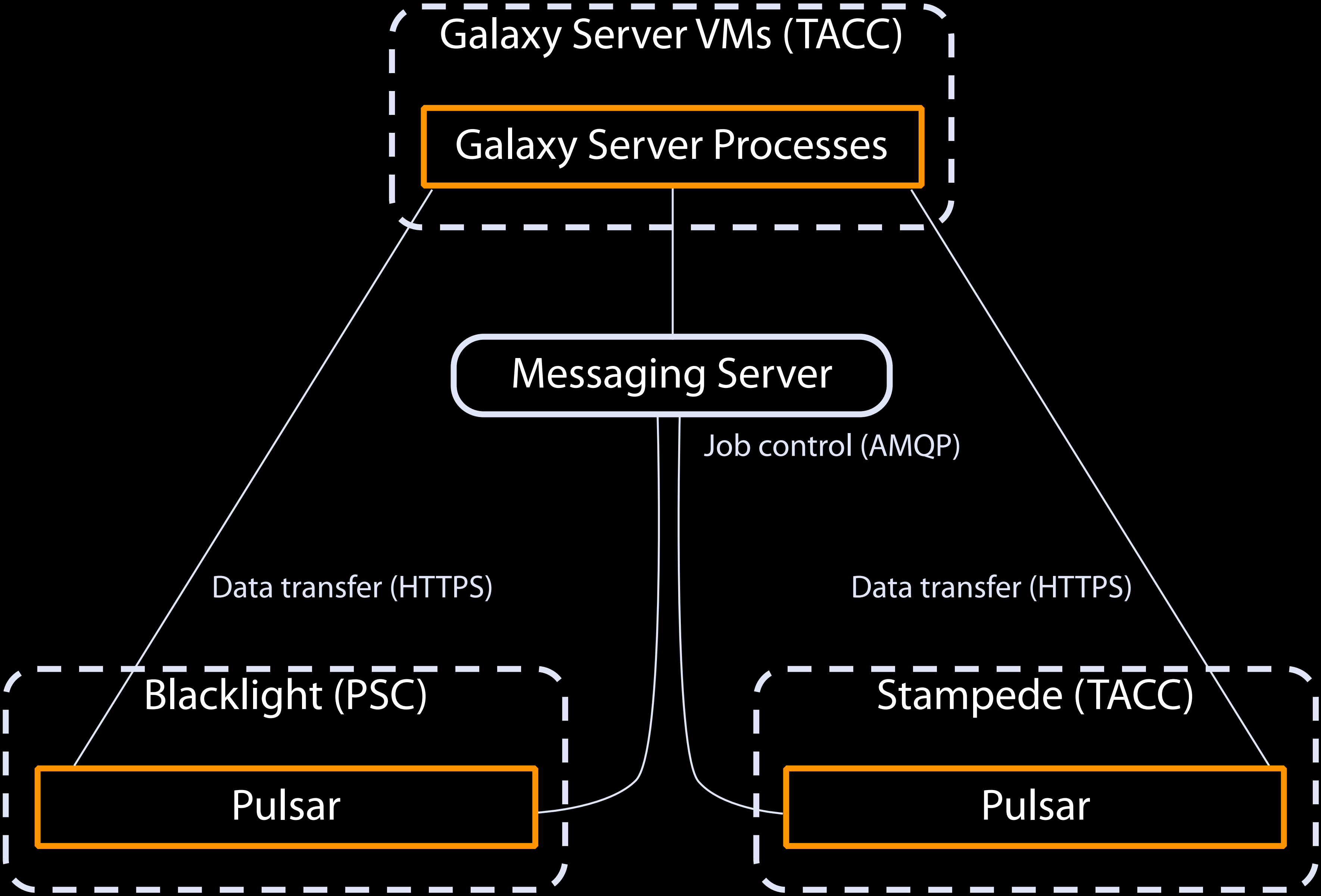
Galaxy can already run jobs on almost any batch system, but most XSEDE resources do not provide direct access for job submission…

Pulsar

Galaxy job runner that can
run almost anywhere

No shared filesystem, stages all necessary
Galaxy components

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

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

Nate Coraor and John Chilton

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

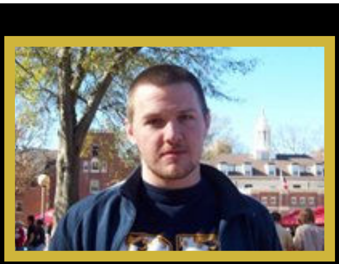
- Texas Advanced Computing Center
  - Dan Stanzione
  - Matt Vaughn
  - Chris Jordan
  - Mike Packard
  - Nathaniel Mendoza
- iPlant Collaborative
  - Stephen Goff
- Pittsburgh Supercomputing Center
  - Philip Blood
  - Kathy Benninger
  - Robert Budden
  - Jared Yanovich
  - Josephine Palencia
  - J.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

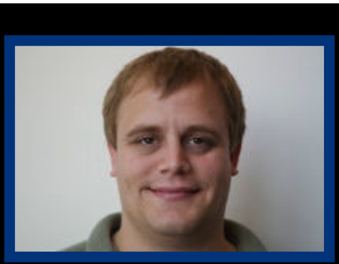
... and the Galaxy Team and community

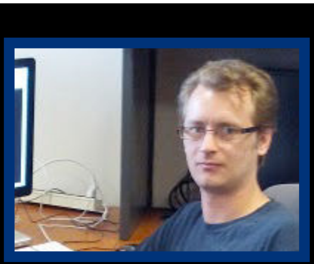
Galaxy is supported in part by NSF, NHGRI, Pennsylvania Department of Public Health, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University

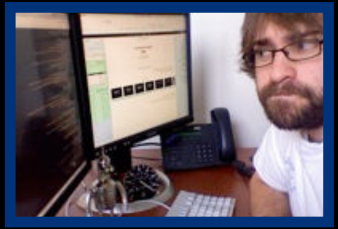
Engineering

Enis Afgan


Dannon Baker


Dan Blankenberg

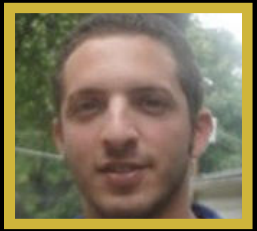

Dave Bouvier


Nate Coraor

Martin Čech

John Chilton

Carl Eberhard


Sam Guerler

Nick Stoler

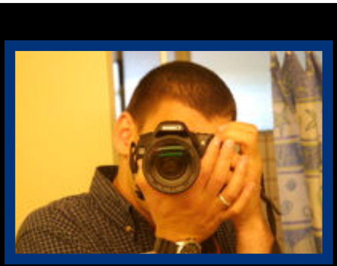
Support and outreach

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