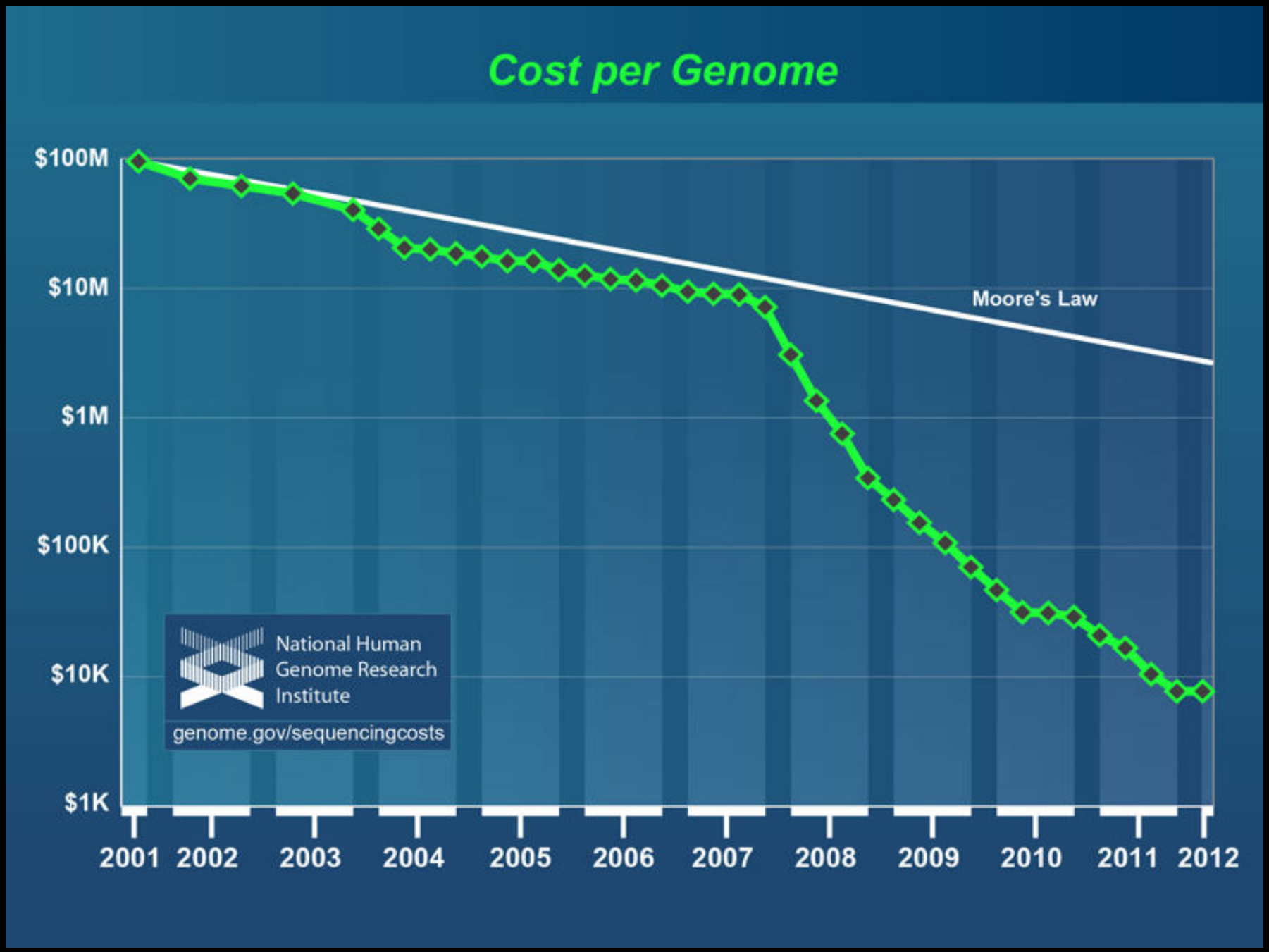
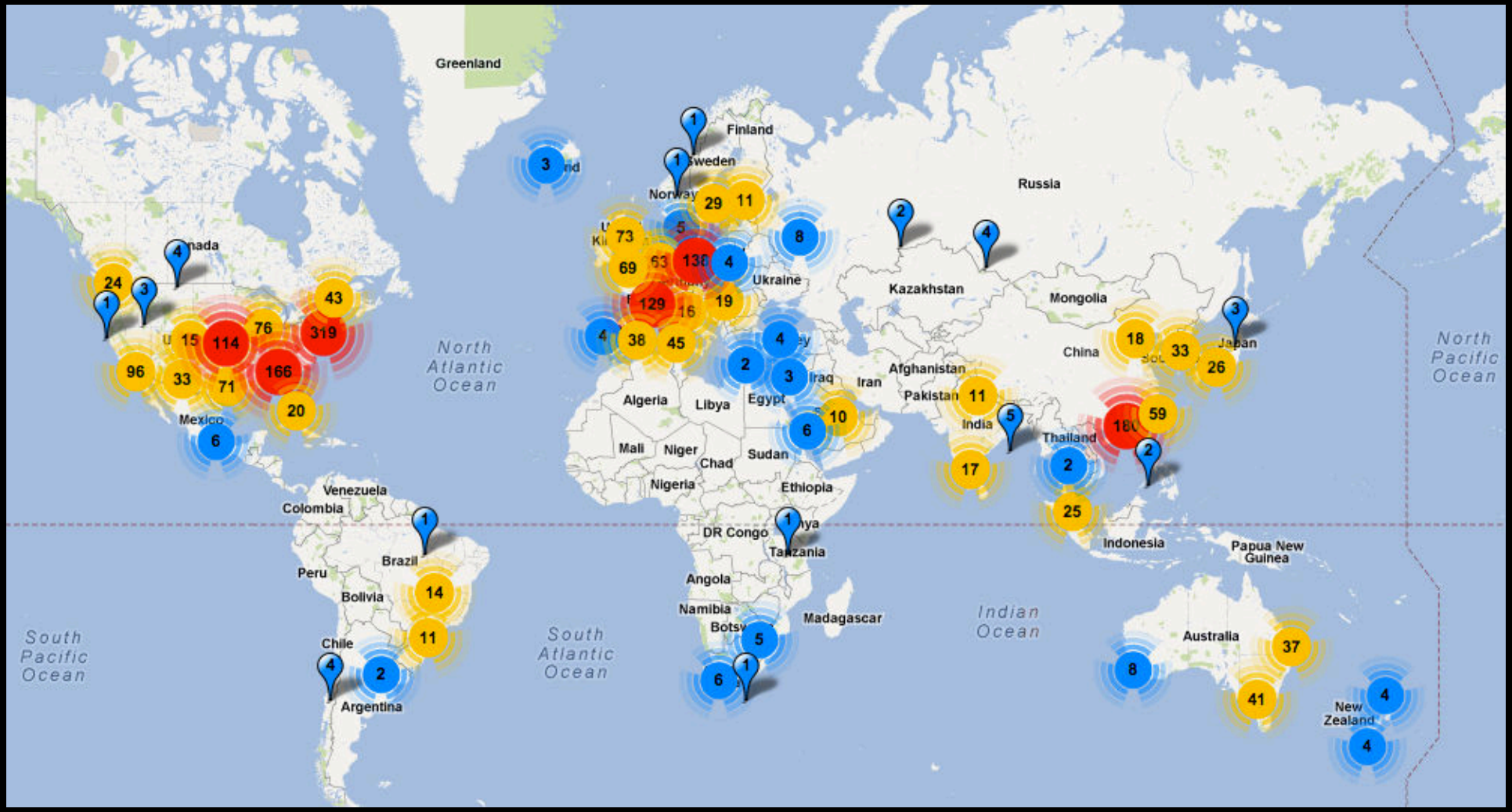
Lessons Learned from Galaxy

Jeremy Goecks, The Galaxy Team,
Anton Nekrutenko, and James Taylor



http://www.genome.gov/sequencingcosts/

World Sequencing Capacity > 15Pbp / year



http://omicsmaps.com

When Science becomes Computational

Scientists unfamiliar with computation

Reproducibility hindered by complexity: systems, scripts, tools, parameters

Collaboration and publishing difficult because current media do not support computational artifacts well

Galaxy

https://main.g2.bx.psu.edu

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 383.5 Gb

Tools

search 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
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:
1: ERR030882_1_brain.fastq
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Built-ins were indexed using default options

Select a reference genome:
Human (Homo sapiens): hg19 Full
If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Paired-end

RNA-Seq FASTQ file:
2: ERR030882_2_brain.fastq
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:
110

TopHat settings to use:
Commonly used
For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Execute

Tophat Overview

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).

Know what you are doing

There is no such thing (yet) as an automated gearshift in splice junction identification. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to **understand** the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.

Input formats

History

transcript expression

13: Cufflinks on data 8: gene expression

12: BodyMap-Brain 75bp SE mapped reads

11: Tophat for Illumina on data 4 and data 3: splice junctions

10: Tophat for Illumina on data 4 and data 3: deletions

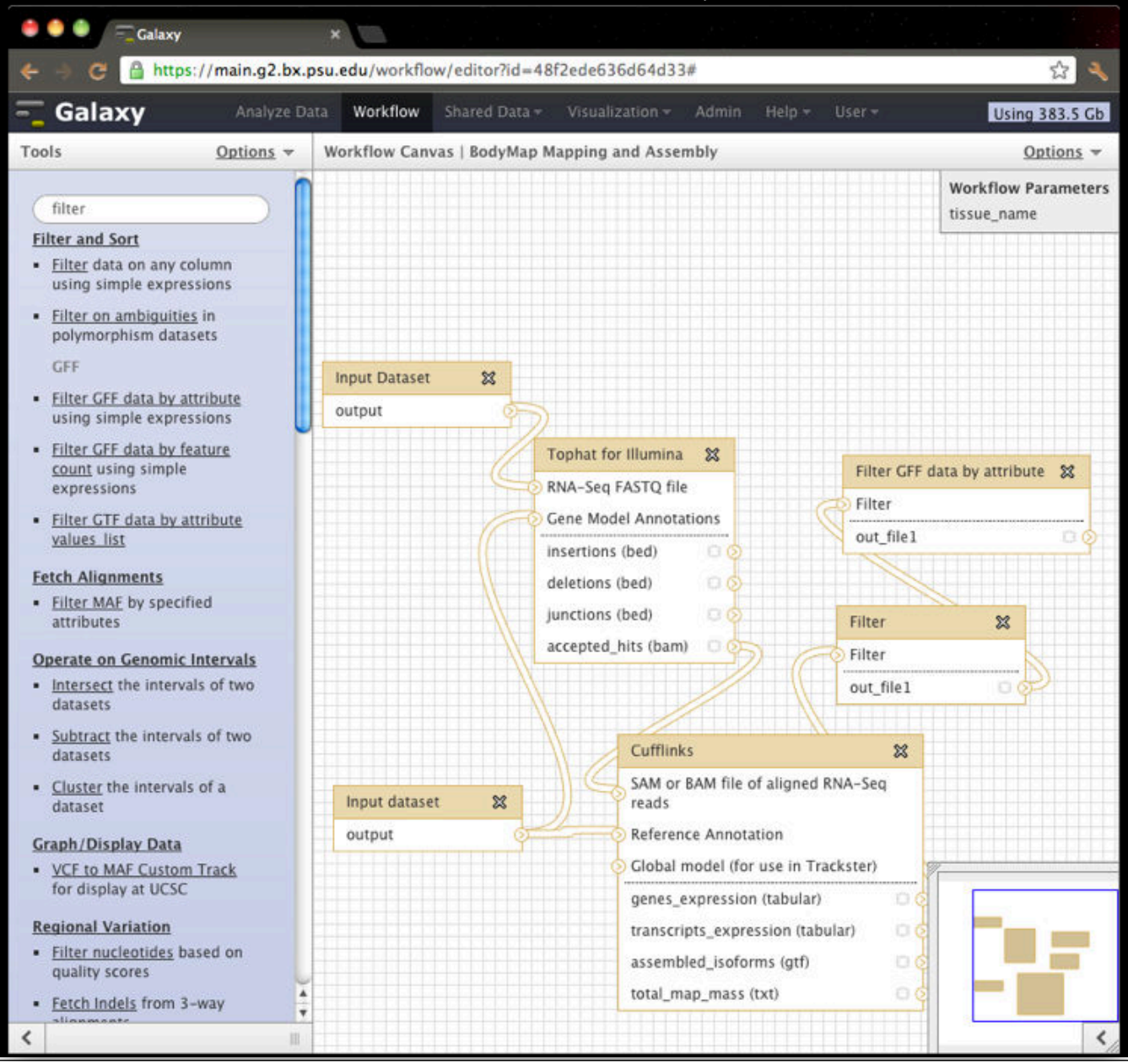
9: Tophat for Illumina on data 4 and data 3: insertions

8: BodyMap-Brain 50bp PE mapped reads
3.9 Gb
format: bam, database: hg19
Info: TopHat v1.4.0
tophat -p 8 -r 110 -a 8 -m 0 -i 20 -I 500000 -g 40 -G /galaxy/main_pool/pool2/files/003/634/dataset_3634785.dat --library-type fr-unstranded --max-insertion-length 3 --max-deletion-length 3 --coverage-search --min-coverage-intron 20 --max-c
display at UCSC main
display at Ensembl Current
display with IGV web current local
display in IGB Local Web

Binary bam alignments file

7: Tophat for Illumina on data 2, data 4, and data 1: splice junctions

Galaxy Workflows

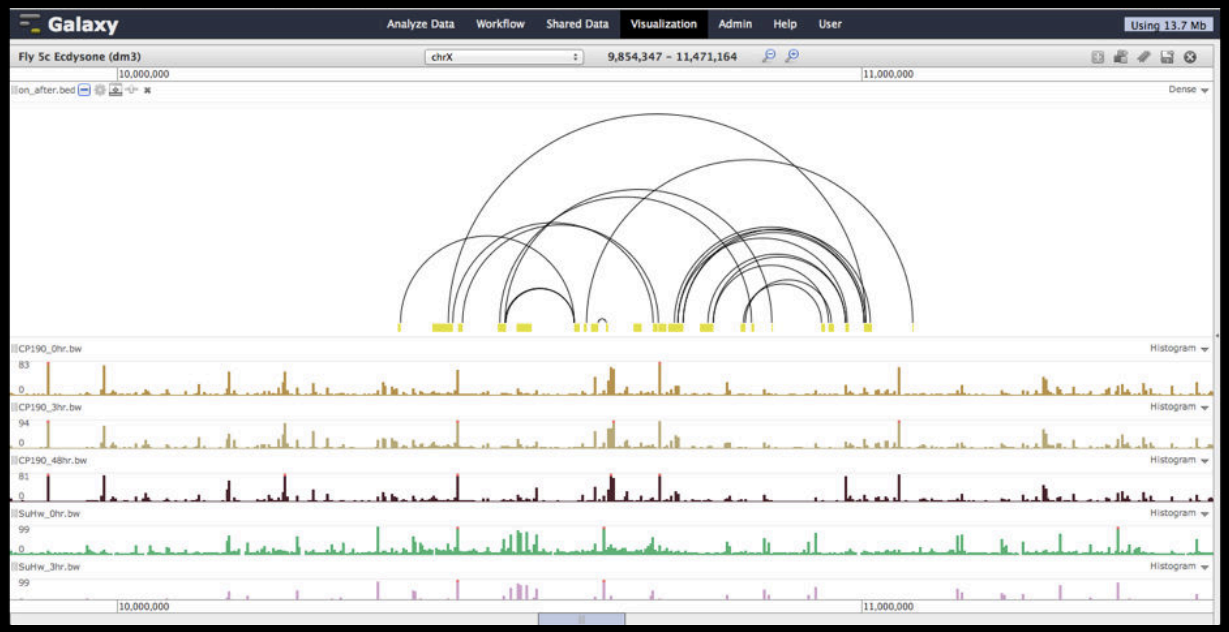


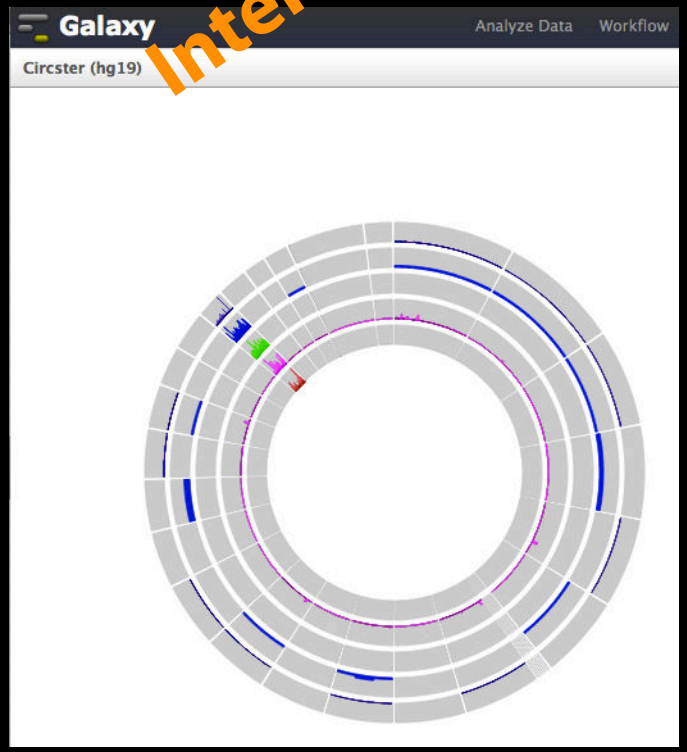
Workflows from scratch *or* extracted from existing analysis histories

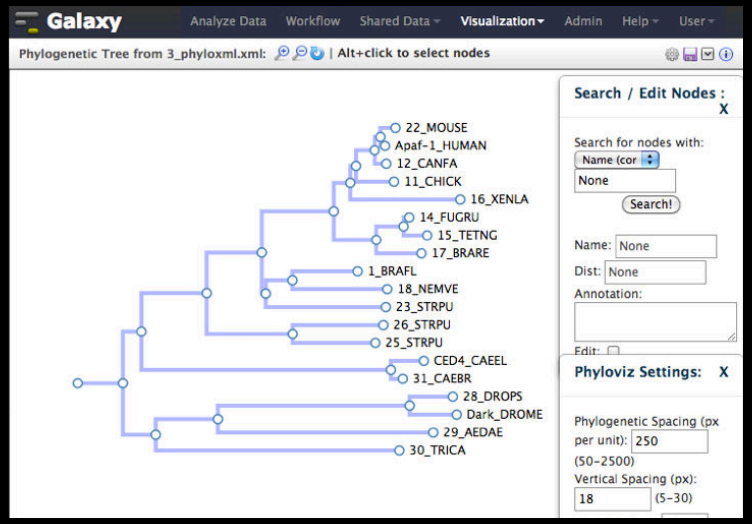
Facilitate reuse and provide precise reproducibility of a complex analysis

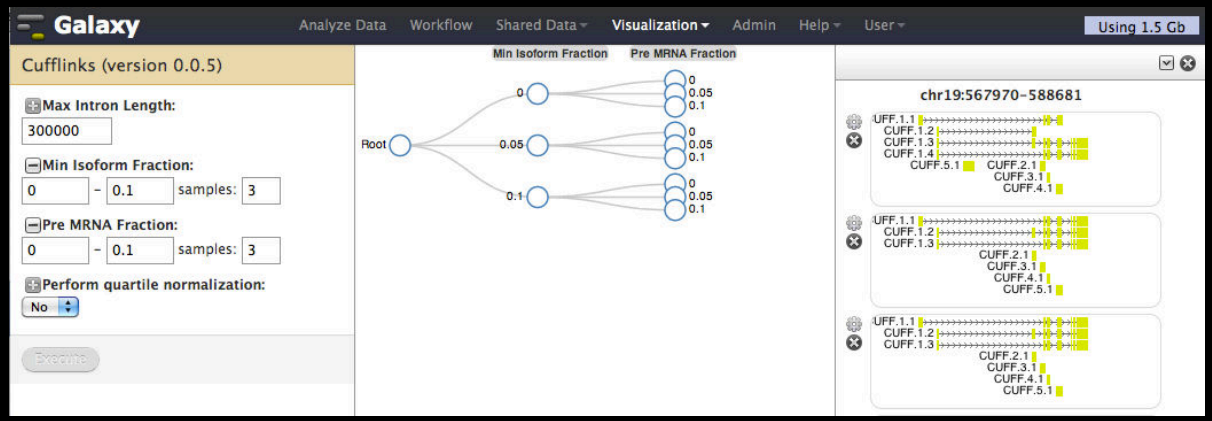
Galaxy Visualization

Interactivity









ec2-50-17-119-106.compute-1.amazonaws.com/root

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools | Options

- Get Data
- Send Data
- ENCODE Tools
- 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
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation

NGS TOOLBOX BETA

- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: Expression Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- Human Genome Variation
- EMBOSS

Welcome to Galaxy on the Cloud

History | Options

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

Connecting Scientists with HPC

When tools, workflows, or visualizations used, Galaxy uses HPC resources

- command line(s) created and submitted to computing cluster

A Web interface makes genomic analyses available to non-programmers

Galaxy Project: Fundamental Questions

When Biology (or any science) becomes dependent on computational methods, how to:

- make tools and methods accessible to scientists?
- ensure that analyses are reproducible?
- enable transparent communication and reuse of analyses?

Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and transparent computational biomedical research

What is Galaxy?

GUI for genomics

- analysis interface, tools and datasources
- data integration
- data libraries
- workflows
- visualization
- sharing
- publication

Customizable open-source software on various HPC resources

- public website
- local instance
- on the cloud
- tool shed/contributing

Galaxy Usage

Public site (http://usegalaxy.org)

- ~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month

Local instances

- U of Texas system, U of Minnesota, *U of Indiana*, U.S. JGI, Oxford U, NBIC (Netherlands), ...
- Public instances: http://wiki.g2.bx.psu.edu/PublicGalaxyServers

Citations

- > 300 for main Galaxy papers
- articles in *Science, Nature, Cell, Genome Research*, ...

Lessons Learned

Open, Extendable Frameworks

Advantages of the Web

Community

Do Science *and* Computing

Software Engineering Matters

Everything is a framework

- tools
- job runners
- storage
- tool shed
- visualizations
- sharing
- datatypes

Advantages (i.e. "motivations for software engineering in academia")

- developer friendly and amenable to community contributions
- adaptability (biology is incredibly diverse)
- amplification and novel recombination

gc_wrapper.xml

```xml
<tool id="fa_gc_content_1" name="Compute GC Content">
    <description>for each sequence in a file</description>
    <command> interpreter="python">gc_content.py $input $output</command>

    <inputs>
        <param format="fasta" name="input" type="data" label="Source file"/>
    </inputs>

    <outputs>
        <data format="tabular" name="output"/>
    </outputs>
</tool>
```

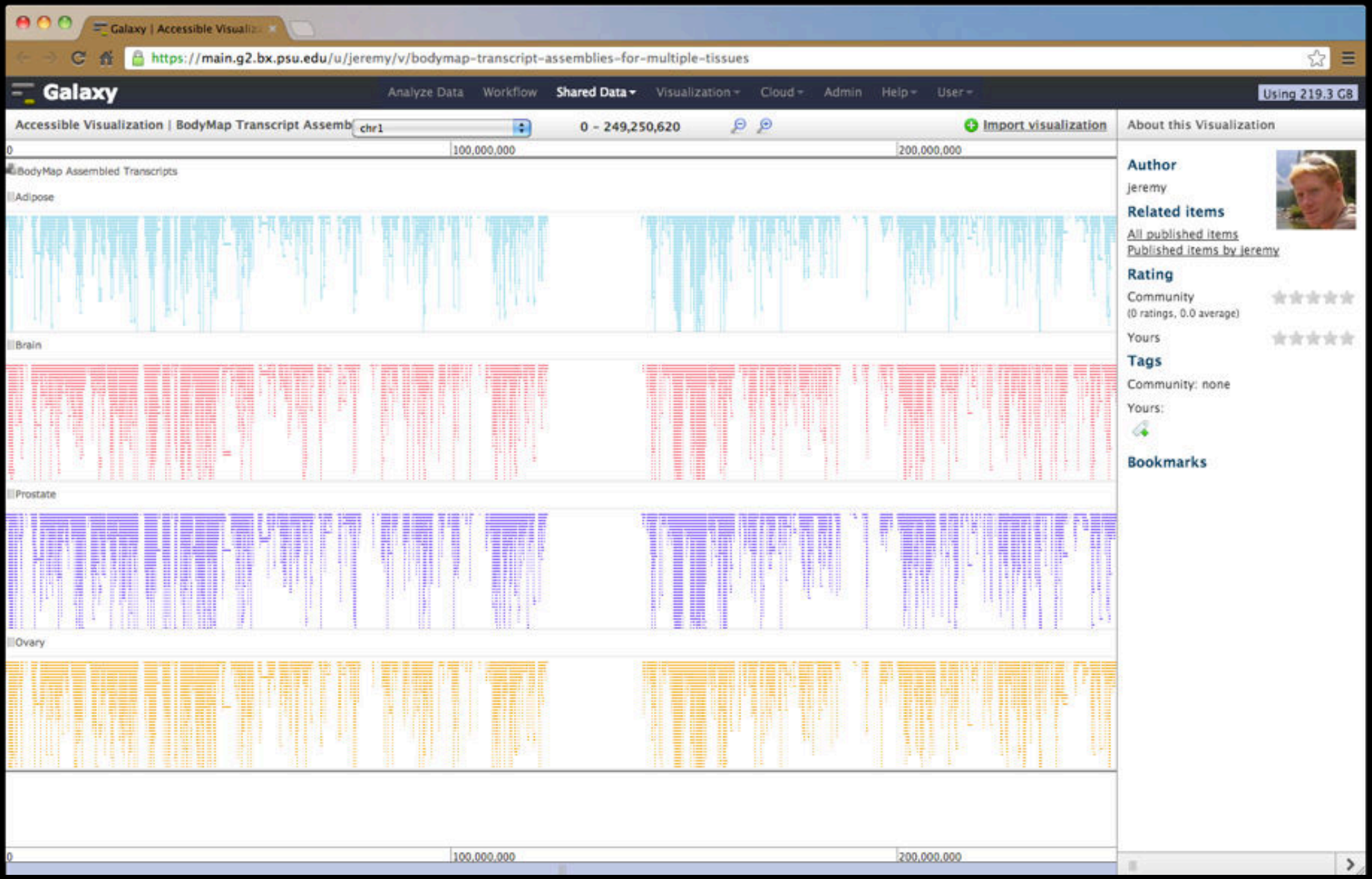
The Web is Amazing

Remote computing is necessary for big data

Rapid development of Web technologies

- HTML5 & JavaScript frameworks

Sharing via Web browser simple: requires only a URL


Galaxy | Accessible Visualiz
https://main.g2.bx.psu.edu/u/jeremy/v/bodymap-transcript-assemblies-for-multiple-tissues
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Admin
Help
User
Using 219.3 GB
Accessible Visualization | BodyMap Transcript Assemb
chr1
0 - 249,250,620
Import visualization
About this Visualization
0
100,000,000
200,000,000
BodyMap Assembled Transcripts
Adipose
Brain
Prostate
Ovary
Author
jeremy
Related items
All published items
Published items by jeremy
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community: none
Yours:
Bookmarks

It Takes a (Global) Village to Build Galaxy

Galaxy bridges two communities: users and developers

- outside investment critical to success

Mailing lists and (soon) Web forums for community usage

Galaxy Community Conference (~200 attendees)

Toolshed for user contributions

- started with tools
- eventually workflows, visualizations, etc.

Galaxy Tool Shed

toolshed.g2.bx.psu.edu

Galaxy Tool Shed

Repositories Help User

2132 valid tools on Oct 06, 2012

Search

- Search for valid tools
- Search for workflows

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Categories

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	22
Computational chemistry	Tools for use in computational chemistry	4
Convert Formats	Tools for converting data formats	29
Data Source	Tools for retrieving data from external data sources	12
Fasta Manipulation	Tools for manipulating fasta data	24
Genomic Interval Operations	Tools for operating on genomic intervals	20
Graphics	Tools producing images	14
Metagenomics	Tools enabling the study of metagenomes	6
Micro-array Analysis	Tools for performing micro-array analysis	0
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	40
Ontology Manipulation	Tools for manipulating ontologies	5
Proteomics	Tools enabling the study of proteins	2
SAM	Tools for manipulating alignments in the SAM format	19
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	109
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	16
Statistics	Tools for generating statistics	26
Systems Biology	Systems biology tools	0
Text Manipulation	Tools for manipulating data	24
Tool Generators	Tools that make or help make new tools	1
Visualization	Tools for visualizing data	23
Web Services	Tools enabling access to web services	1

Science is Fun!

Doing science uncovers real computing challenges

Galaxy research model

1. find challenge while doing science
2. invent new computing to address challenge
3. demonstrate usefulness of new computing via biology investigation

Software

Highly accessed | Open Access

Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences

Jeremy Goecks[1], **Anton Nekrutenko**[2*], **James Taylor**[1*] and **The Galaxy Team**

* Corresponding authors: Anton Nekrutenko anton@bx.psu.edu - James Taylor james.taylor@emory.edu

▸ Author Affiliations

For all author emails, please log on.

Genome Biology 2010, **11**:R86 doi:10.1186/gb-2010-11-8-r86
Published: 25 August 2010

Harnessing cloud computing with Galaxy Cloud

Enis Afgan, Dannon Baker, Nate Coraor, Hiroki Goto, Ian M Paul, Kateryna D Makova, Anton Nekrutenko & James Taylor

Affiliations | **Corresponding authors**

Nature Biotechnology **29**, 972–974 (2011) | doi:10.1038/nbt.2028
Published online 08 November 2011

CSH PRESS GENOME RESEARCH

– Gen
– Qua
– RNA

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CON

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

Back to the Future

Virtual and cloud resources are the future, but there are challenges

Saving/restoring/archiving data when Galaxy instances are transient

- often costs are in storage, not compute

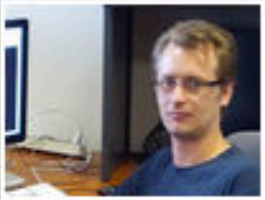
Finding shared and published items across "universe of Galaxies"

Enis Afgan
IRB

Guru Ananda
Penn State

Dannon Baker
Emory

Dan Blankenberg
Penn State



Dave Bouvier
Penn State

Dave Clements
Emory

Nate Coraor
Penn State

Carl Eberhard
Emory

Jeremy Goecks
Emory

Jennifer Hillman Jackson
Penn State

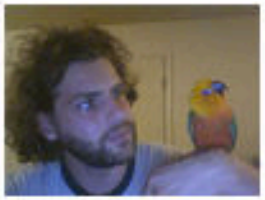
Greg von Kuster
Penn State

Ross Lazarus
Harvard & BakerIDI

Rémi Marenco
Emory

Scott McManus
Emory

Anton Nekrutenko
Penn State

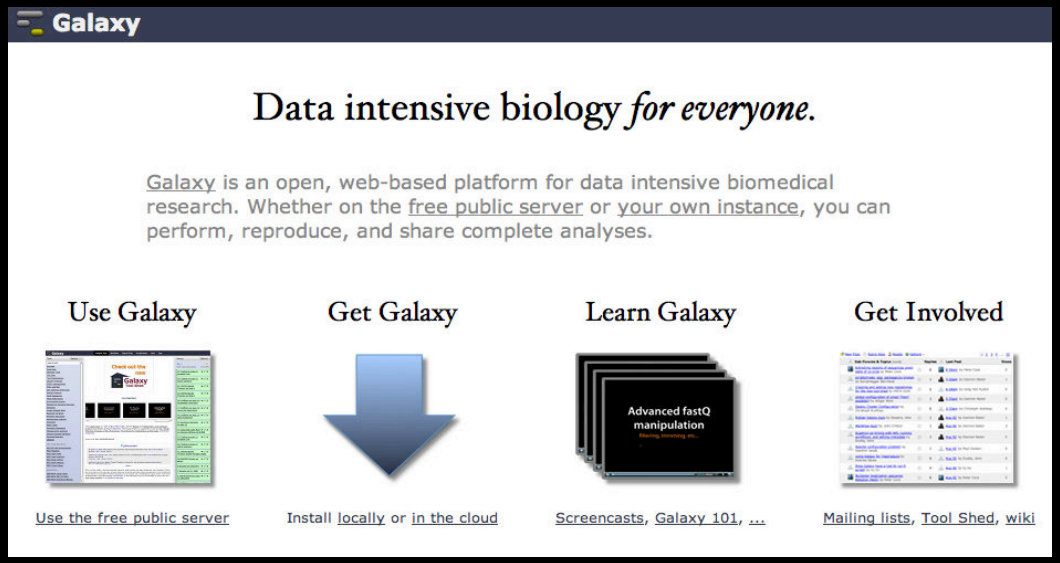


James Taylor
Emory

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

Thanks! Questions?

http://galaxyproject.org



Galaxy publications: http://galaxyproject.org/wiki/Citing

jeremy.goecks@emory.edu