Galaxy

GMOD Malaysia
Kuala Lumpur
26-28 Febrary 2014

Dave Clements
Johns Hopkins University
http://galaxyproject.org/

GMOD malaysia

Galaxy

amazon webservices™

National Evolutionary Synthesis Center
NESCent

The Agenda

Introduction to Galaxy

Hands-on Analysis

Community Resources

Galaxy on the Cloud

Done

Goal is to demonstrate how Galaxy can help you explore and learn options, perform analysis, and then share, repeat, and reproduce your analyses.

Not The Agenda

This workshop will *not* cover

- details of how tools are implemented, or
- new algorithm designs, or
- which assembler or mapper or peak caller or ... is best for you.

This workshop is *not* about learning how to do a specific type of analysis.

What is Galaxy?

- **A free (for everyone) web service**
- **Open source software**
- These options result in several **ways to use Galaxy**

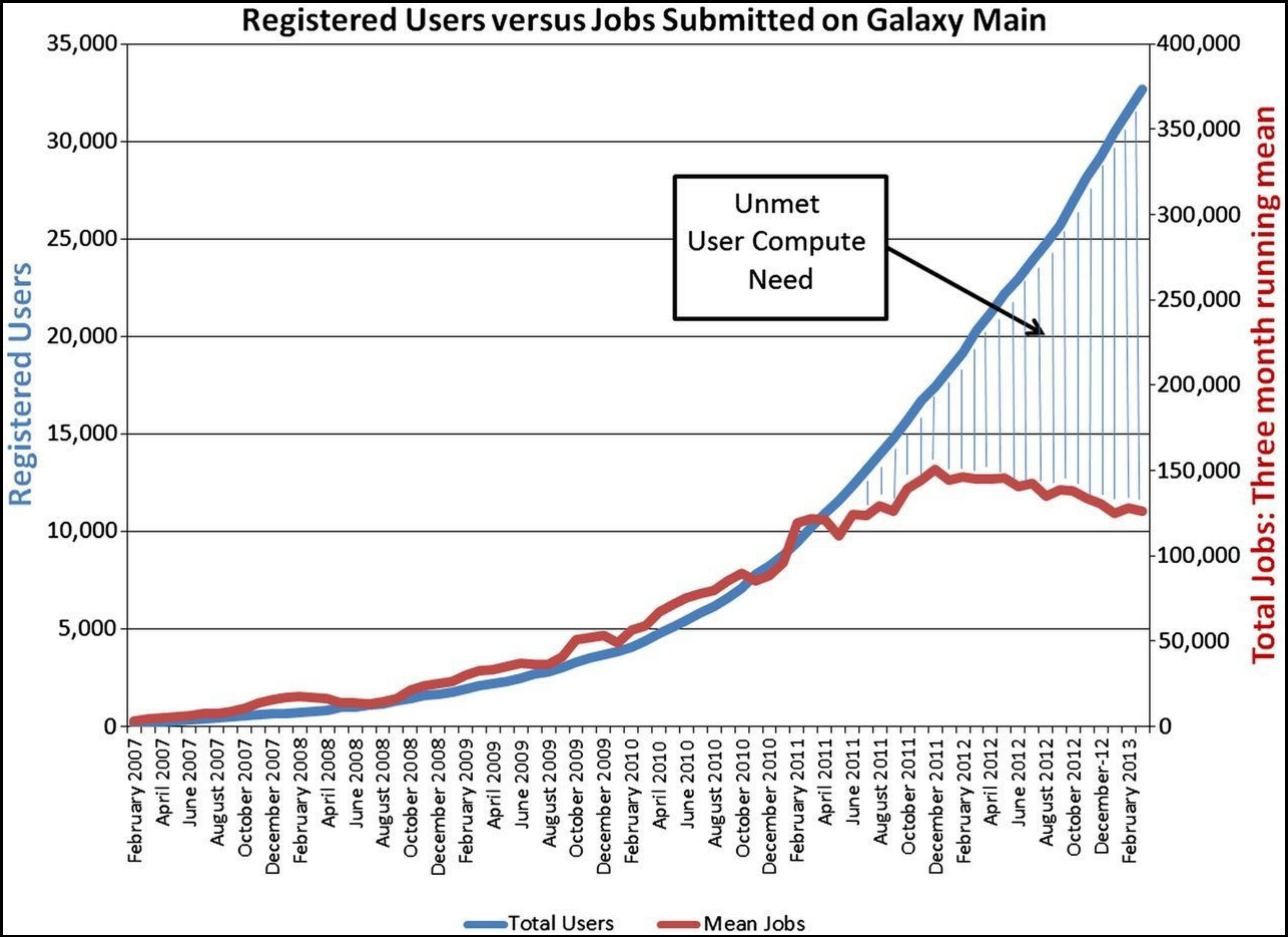
http://galaxyproject.org

Galaxy is available ...

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

http://usegalaxy.org

However, ***a centralized solution cannot support the different analysis needs of the entire world.***



Leveraging the national cyberinfrastructure for biomedical research
LeDuc, *et al. J Am Med Inform Assoc doi:10.1136/amiajnl-2013-002059*

Galaxy is available ...

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

Galaxy is available ...

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

- *On the Cloud*

  We are using this today.

amazon
web services™

http://aws.amazon.com/education

http://wiki.galaxyproject.org/Cloud

Galaxy is available ...

- As a free (for everyone) web service
- As open source software
- On the Cloud
- *With Commercial Support*

SLIPSTREAM APPLIANCE GALAXY EDITION
Galaxy made easy.

A ready-to-use appliance (BioTeam)

Cloud-based solutions (ABgenomica, AIS, Appistry, GenomeCloud)

Consulting & Customization (Arctix, BioTeam, Deena Bioinformatics)

Galaxy Project: Further reading & Resources

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://wiki.galaxyproject.org/Cloud

http://bit.ly/gxychoices

The Agenda

Introduction to Galaxy

Hands-on Analysis

Community Resources

Galaxy on the Cloud

Done

What is our path?

- Will walk through an NGS example.
- Will adjust content based on this audience's experience level
- Will get as far as we get.

http://cloud2.galaxyproject.org/

http://cloud3.galaxyproject.org/

Agenda

Introduction to Galaxy

Hands-on Analysis

- Quality Control

Community Resources

Galaxy on the cloud

Done

NGS Data Quality Control

- FASTQ format
- Examine quality in an RNA-Seq dataset
- Trim/filter as we see fit, hopefully without breaking anything.

Quality Control is not sexy.

It is vital.

What is FASTQ?

- Specifies sequence (FASTA) and quality scores (PHRED)
- Text format, 4 lines per entry

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- FASTQ is such a cool standard, there are 3 (or 5) of them!

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa        Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

http://en.wikipedia.org/wiki/FASTQ_format

NGS Data Quality Exercise

Create new history

⚙ (cog) → Create New

Get some data

Shared Data → Data Libraries

→ RNA-Seq Example*

→ Untrimmed FASTQ

→ Select MeOH_REP1_R1, MeOH_REP1_R2 and then Import to current history

UCDAVIS Bioinformatics Core
Genome Center

* RNA-Seq example datasets from the 2013 UC Davis Bioinformatics Short Course. http://bit.ly/ucdbsc2013

NGS Data Quality: Assessment tools

Options 1 & 2:

1. NGS QC and Manipulation → Compute Quality Statistics

   NGS QC and Manipulation → Draw quality score boxplot

   No control over how it is calculated or presented, statistics in text and graphic formats.

2. NGS QC and Manipulation → FastQ Summary Statistics,

   Graph / Display Data → Boxplot of quality statistics

   Lots of control over what the box plot looks like, statistics in text and graphic formats

NGS Data Quality: Assessment tools

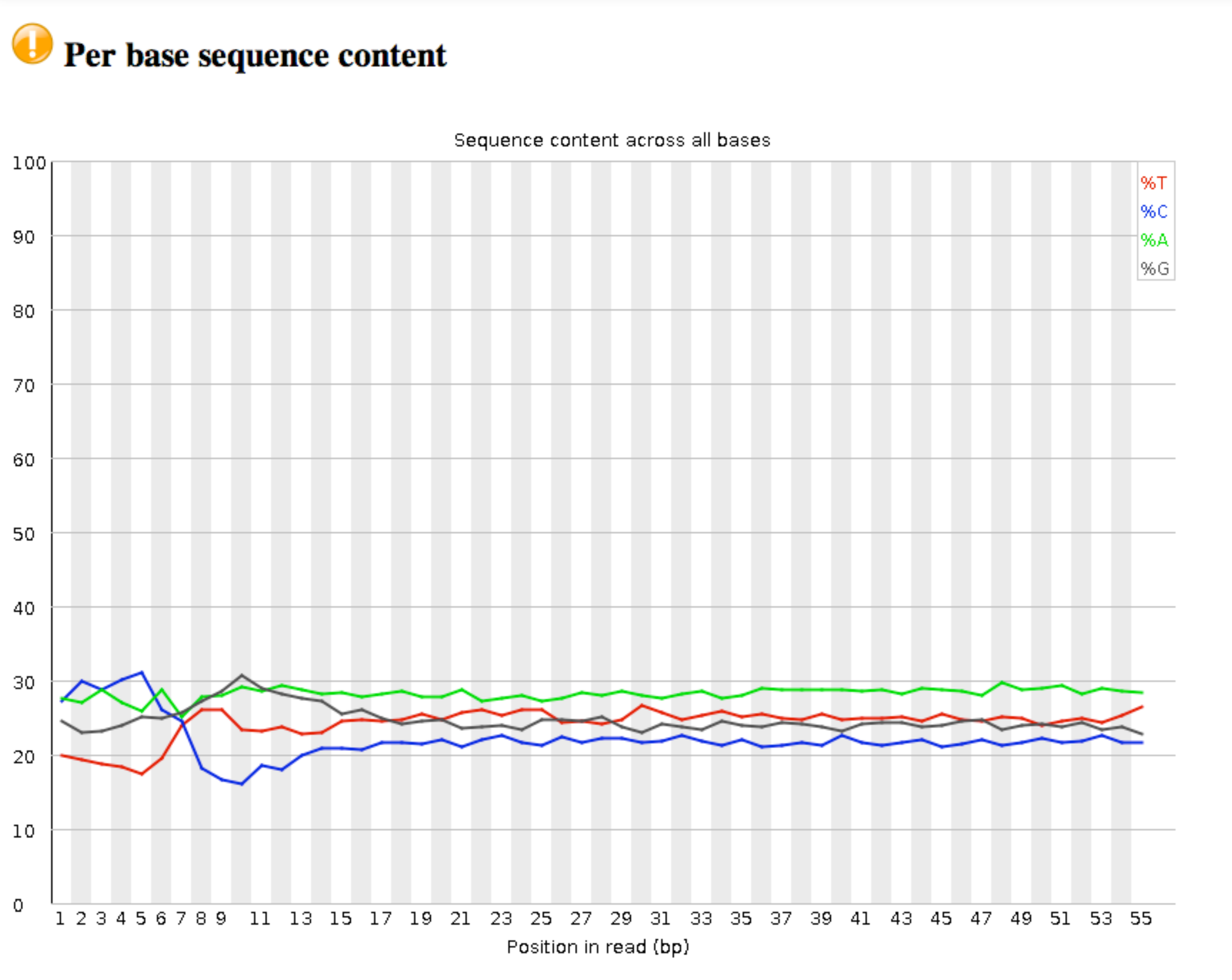
Option 3:

3. NGS QC and Manipulation → **FastQC**

- Gives you a lot a lot more information but little control over how it is calculated or presented.

http://bit.ly/FastQCBoxPlot

NGS Data Quality: Sequence bias at front of reads?

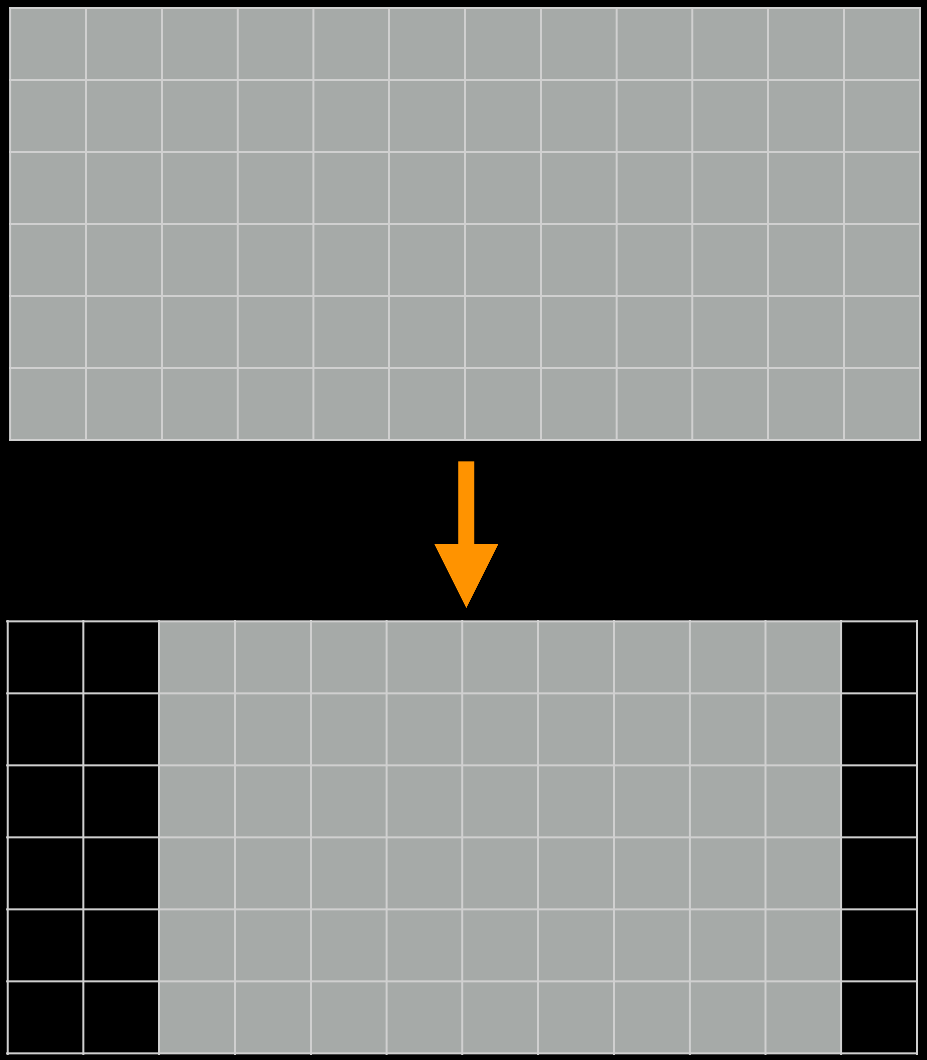


From a sequence specific bias that is caused by use of random hexamers in library preparation.

Hansen, *et al.*, "Biases in Illumina transcriptome sequencing caused by random hexamer priming" *Nucleic Acids Research*, Volume 38, Issue 12 (2010)

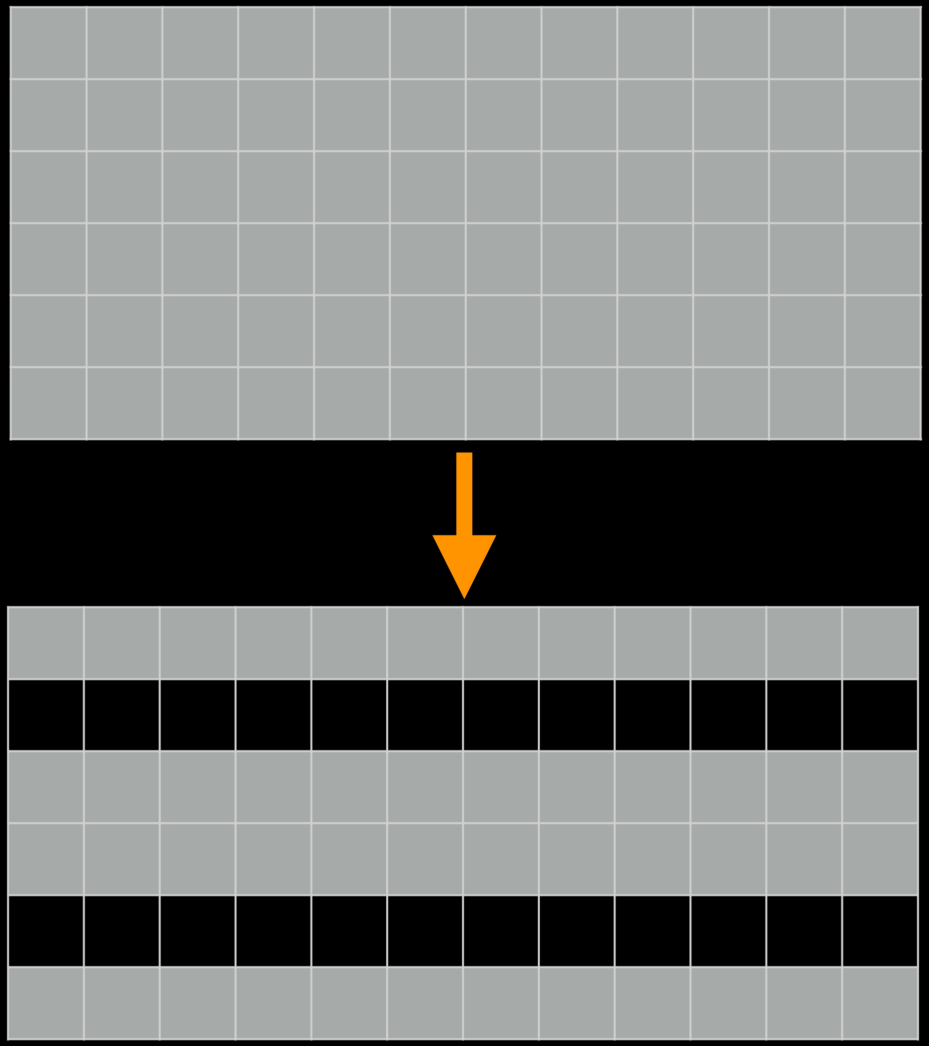
NGS Data Quality: Trim as we see fit

- Trim as we see fit: Option 1
  - NGS QC and Manipulation → **FASTQ Trimmer by column**
  - Trim same number of columns from every record
  - Can specify different trim for 5’ and 3’ ends



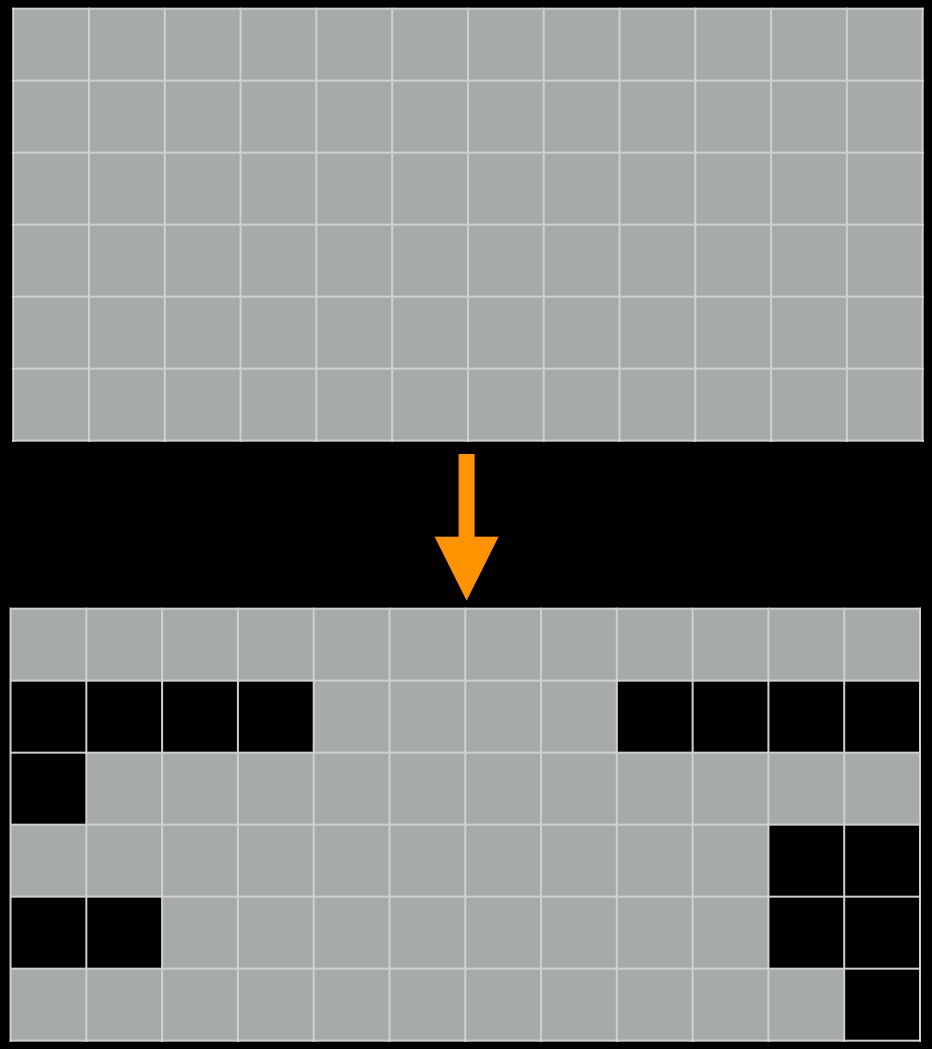
NGS Data Quality: Base Quality Trimming

- ~~Trim~~ Filter as we see fit: Option 2
  - NGS QC and Manipulation → **Filter FASTQ reads by quality score and length**
  - Keep or discard whole reads
  - Can have different thresholds for different regions of the reads.
  - Keeps original read length.

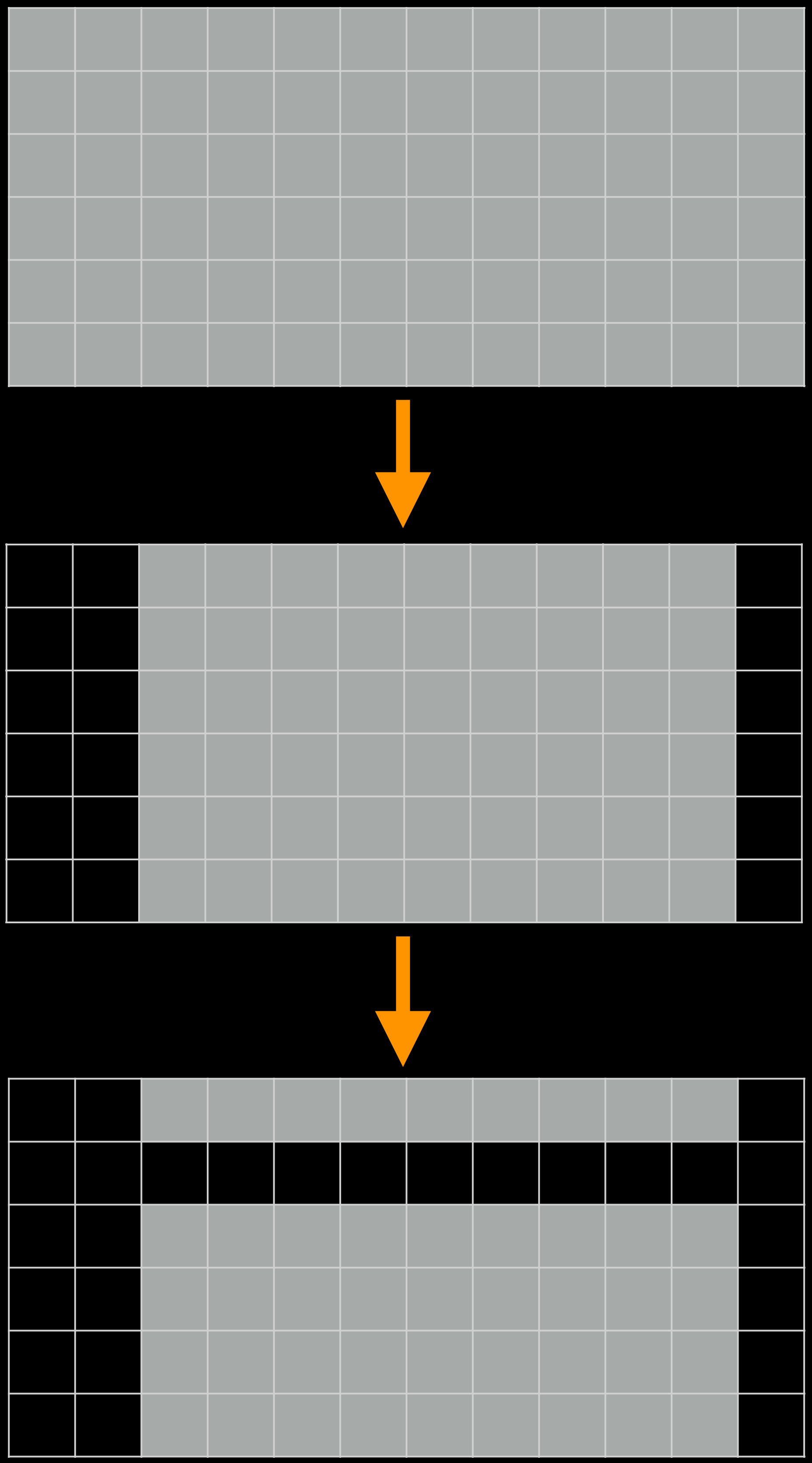


NGS Data Quality: Base Quality Trimming

- Trim as we see fit: Option 3
  - NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**
  - Trim from both ends, using sliding windows, until you hit a high-quality section.
  - Produces variable length reads



Options are not mutually exclusive



Option 1

(by column)

+

Option 2

(by entire row)

Trim? *As we see fit?*

- Introduced 3 options
  - One preserves original read length, two don’t
  - One preserves number of reads, two don’t
  - Two keep/make every read the same length, one does not
  - One preserves pairings, two don’t

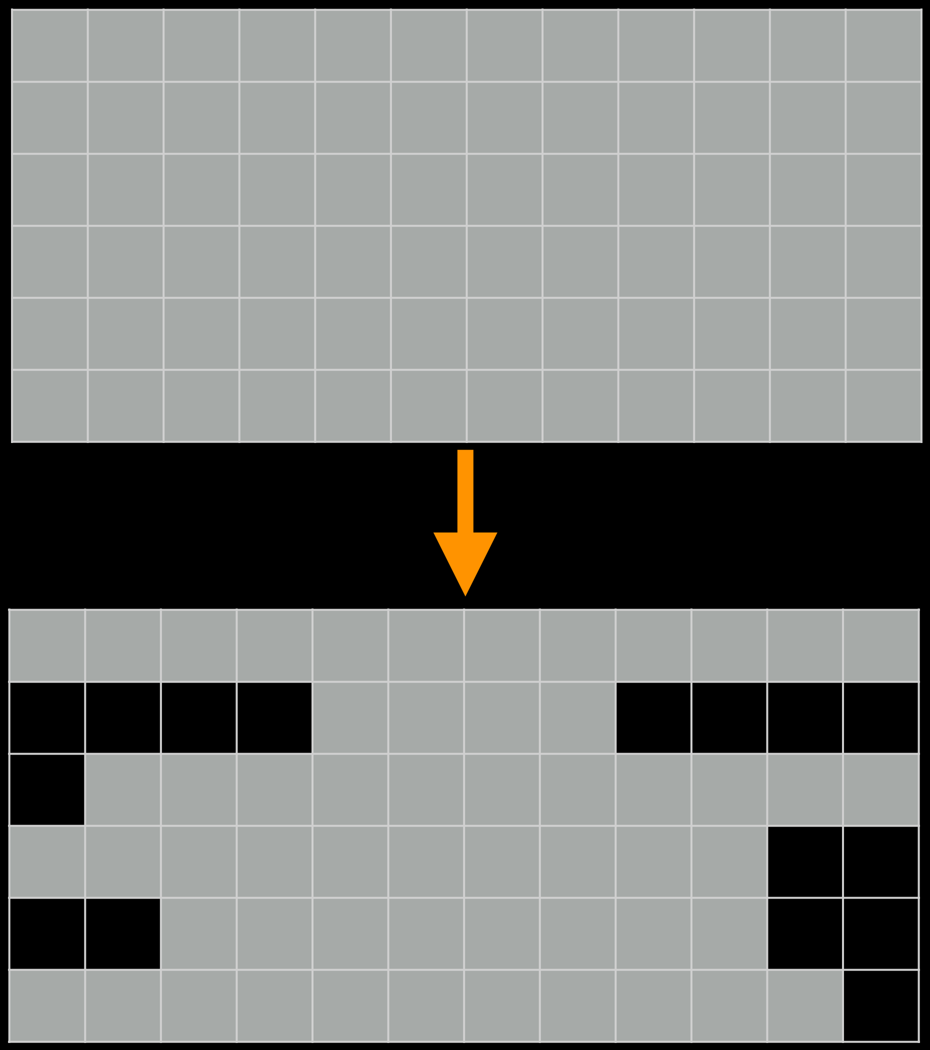
Trim? *As we see fit?*

- Choice depends on downstream tools
- Find out assumptions & requirements for downstream tools and make appropriate choice(s) now.
- How to do that?
  - Read the tool documentation
  - http://biostars.org/
  - http://seqanswers.com/
  - http://galaxyproject.org/search

SEQanswers
the next generation sequencing community

Biostar
BIOINFORMATICS EXPLAINED

NGS Data Quality: Base Quality Trimming



I really want to use Option 3:

- NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**

but ...

“Mixing paired- and single- end reads together is **not** supported.”

Tophat Manual

“If you are performing RNA-seq analysis, there is no need to filter the data to ensure exact pairs before running Tophat.”

Jen Jackson

Galaxy User Support Person Extraordinaire

“Dang.”

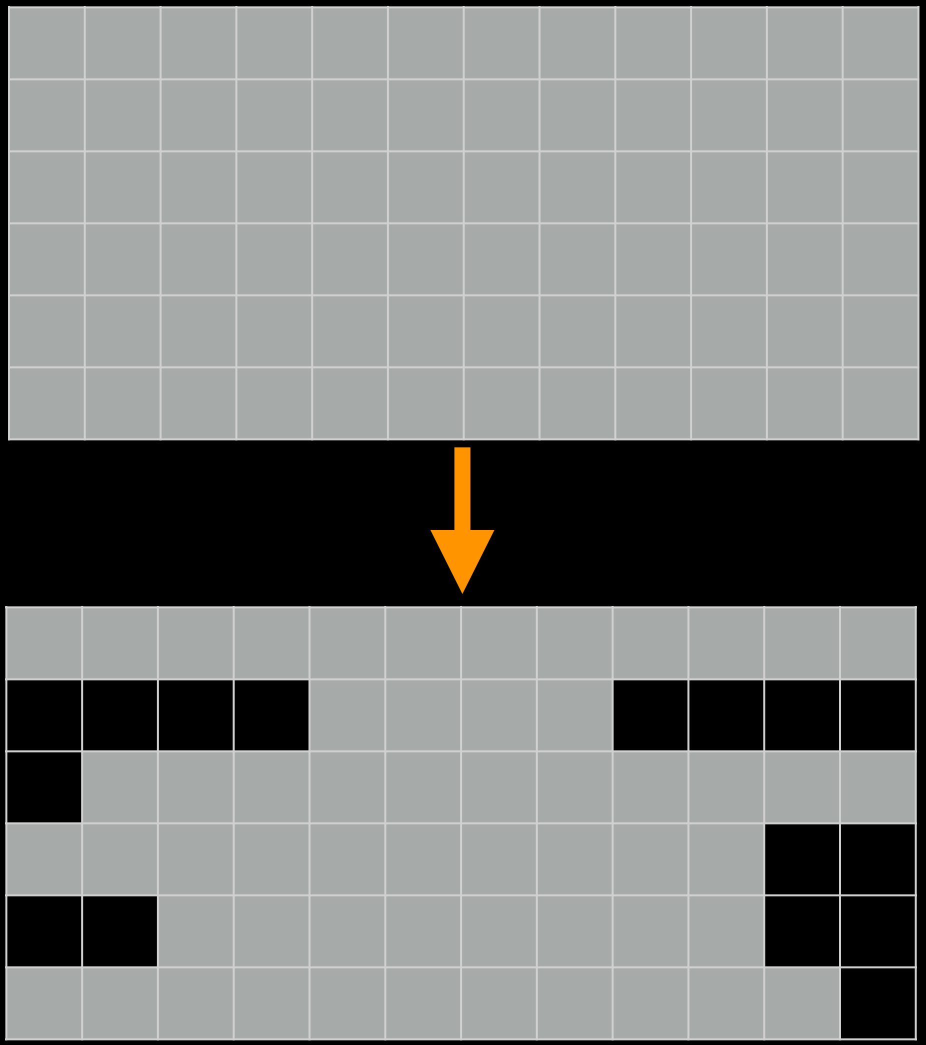
Dave C, mere mortal

Running Tophat on *no-longer-cleanly-paired* data *does map the reads*, but, it no longer keeps track of read pairs in the SAM/BAM file.

Keeping paired ends paired: Options

- Don't bother.
- Run a workflow that removes any unpaired reads before mapping.
- Run the Picard Paired Read Mate Fixer after mapping reads.
- Use sliding windows for QC, but keep empty reads.

NGS Data Quality: Base Quality Trimming



I'll use Option 3 (*but with the special sauce):*

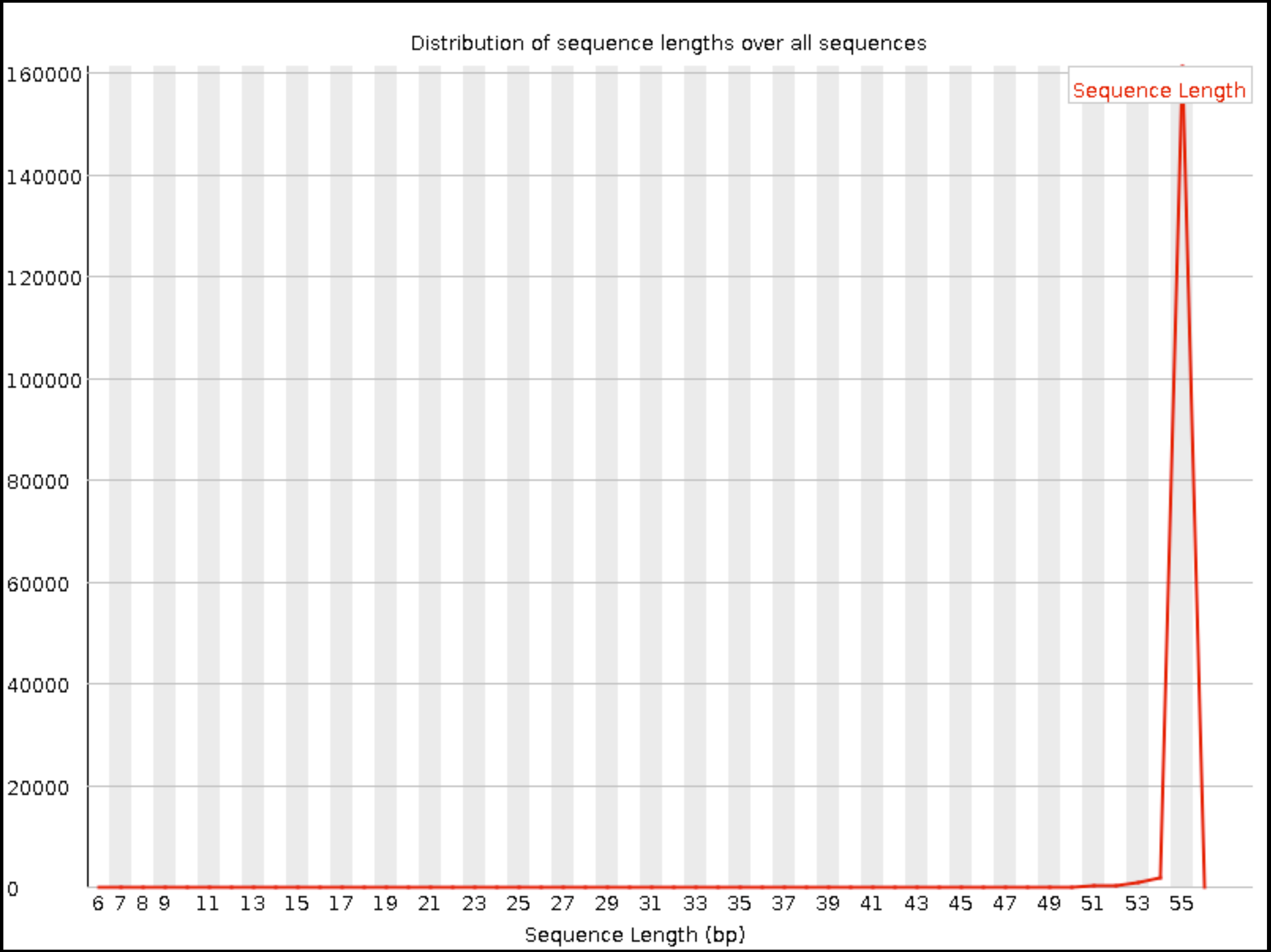
- NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**

  Check **"Keep reads with zero length"**

Run again:

- NGS QC and Manipulation → **FastQC** on trimmed dataset

NGS Data Quality: Base Quality Trimming



New Problem?

Now some reads are so short they are just noise and can't be meaningfully mapped

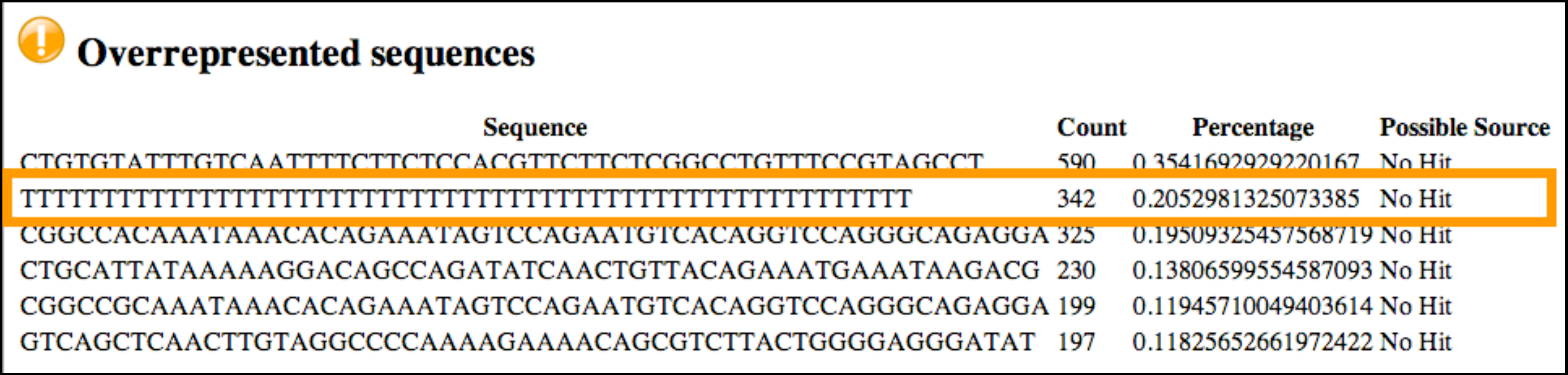
Option 2 can fix this (but break pairings).

Or, your mapper may have an option to ignore shorter reads

NGS Data Quality: Sequencing Artifacts

Repeat this process with MeOH Rep1 R2 (the reverse reads)

... and there's a problem in Overrepresented sequences:



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGTGTATTTGTCAATTTTCTTCTCCACGTTCTTCTCGGCCTGTTTCCGTAGCCT	590	0.3541692929220167	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	342	0.2052981325073385	No Hit
CGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	325	0.19509325457568719	No Hit
CTGCATTATAAAAAGGACAGCCAGATATCAACTGTTACAGAAATGAAATAAGACG	230	0.13806599554587093	No Hit
CGGCCGCAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	199	0.11945710049403614	No Hit
GTCAGCTCAACTTGTAGGCCCCAAAAGAAAACAGCGTCTTACTGGGGAGGGATAT	197	0.11825652661972422	No Hit

NGS QC and Manipulation → **Remove sequencing artifacts**

But this will break pairings.

NGS Data Quality: Done with 1st Replicate!

Now, only 3 (or 5) more to go!

Workflows:

Create a QC workflow that does all these steps

(Or, cheat and import the shared workflow.)

Load the MeOH_REP2, R3G_REP1, and R3G_REP2 replicates into your history, and

Run them through your workflow.

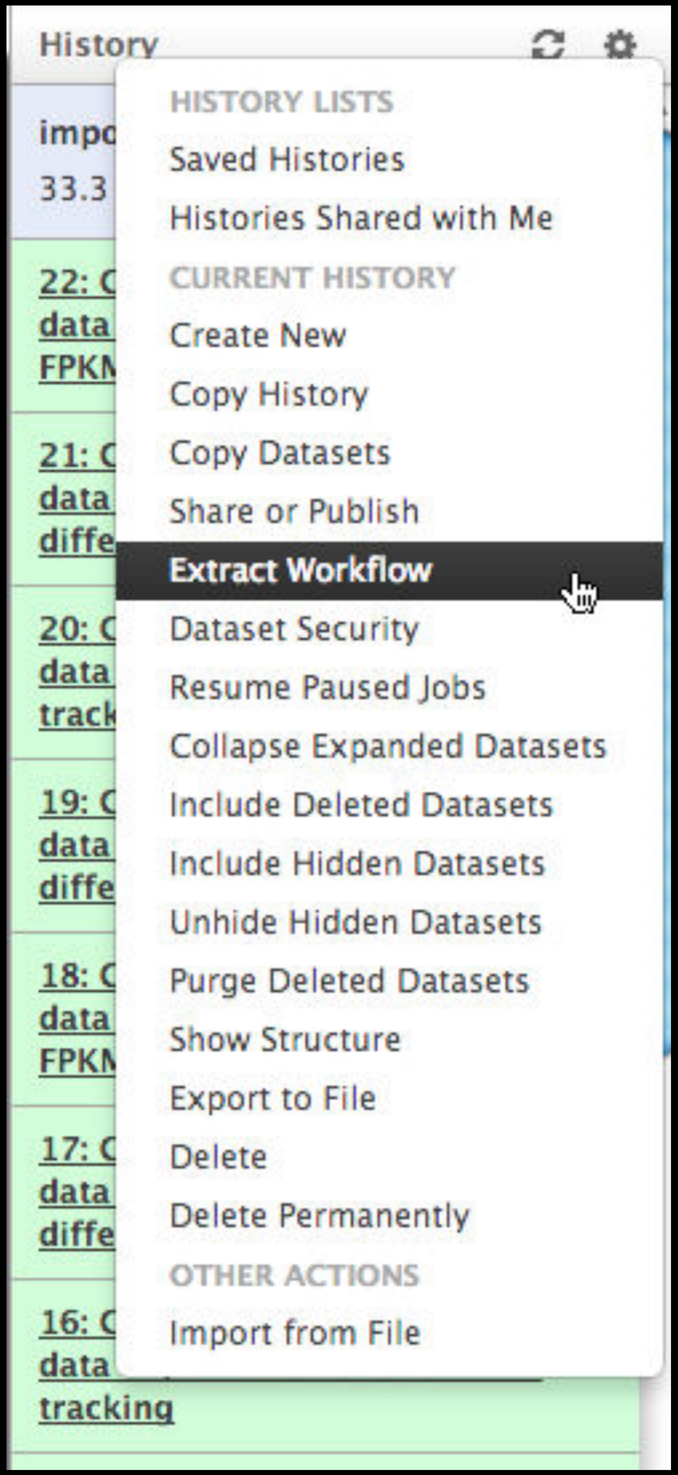
Create a Workflow from a History

Extract Workflow from history

Create a workflow from this history.

Edit it to make some things clearer.

⚙ (cog) → Extract Workflow



HISTORY LISTS
- Saved Histories
- Histories Shared with Me

CURRENT HISTORY
- Create New
- Copy History
- Copy Datasets
- Share or Publish
- **Extract Workflow**
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Include Deleted Datasets
- Include Hidden Datasets
- Unhide Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently

OTHER ACTIONS
- Import from File

NGS Data Quality: Further reading & Resources

FastQC Documenation

Read Quality Assessment & Improvement
by Joe Fass
From the UC Davis 2013 Bioinformatics Short Course

Manipulation of FASTQ data with Galaxy
by Blankenberg, *et al*.

Agenda

Introduction to Galaxy

Hands-on Analysis

- Mapping with TopHat

Community Resources

Galaxy on the cloud

Done

RNA-seq Exercise: Mapping with Tophat

Create a new history

Import all datasets from library:

RNA-Seq Example → Trimmed FASTQ

Get all datasets, and

RNA-Seq Example

Get genes_chr12.gtf

NGS: RNA Analysis → TopHat for Illumina

RNA-seq Exercise: Mapping with Tophat

- Tophat looks for best place(s) to map reads, and best places to insert introns
- *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq mapping here.*

Mapping with Tophat: mean inner distance

Expected distance between paired ends

- Has to be provided to you by sequencing core!
- We'll use 90* for mean inner distance
- We'll use 50 for standard deviation

✳ The library was constructed with the typical Illumina TruSeq protocol, which is supposed to have an average insert size of 200 bases. Our reads are 55 bases (R1) plus 55 bases (R2). So, the Inner Distance is estimated to be 200 - 55 - 55 = 90

From the 2013 UC Davis Bioinformatics Short Course

Mapping with Tophat: Use Existing Annotations?

You can bias Tophat towards known annotations

- Use Own Junctions → Yes
  - Use Gene Annotation → Yes
  - Gene Model Annotation → genes_chr12.gtf
- Use Raw Junctions → Yes (tab delimited file)
- Only look for supplied junctions → Yes

Mapping with Tophat: Make it quicker?

Warning: Here be dragons!

- Allow indel search → No
- Use Coverage Search → No (wee dragons)

TopHat generates its database of possible splice junctions from two sources of evidence. The first and strongest source of evidence for a splice junction is when two segments from the same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found *ab initio*. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million). This latter option will only report alignments across "GT-AG" introns

TopHat Manual

Mapping with Tophat: Max # of Alignments Allowed

Some reads align to more than one place equally well.

For such reads, how many should Tophat include?

If more than the specified number, Tophat will pick those with the best mapping score.

Tophat break ties randomly.

Tophat assigns equal fractional credit to all *n* mappings

Instructs TopHat to allow up to this many alignments to the reference for a given read, and choose the alignments based on their alignment scores if there are more than this number. The default is 20 for read mapping. Unless you use --report-secondary-alignments, TopHat will report the alignments with the best alignment score. If there are more alignments with the same score than this number, TopHat will randomly report only this many alignments. In case of using --report-secondary-alignments, TopHat will try to report alignments up to this option value, and TopHat may randomly output some of the alignments with the same score to meet this number.

TopHat Manual

Mapping with Tophat: Lets do it some more!

NGS: RNA Analysis → TopHat

for the remaining replicates

RNA-Seq Mapping With Tophat: Resources

RNA-Seq Concepts, Terminology, and Work Flows
by Monica Britton

Aligning PE RNA-Seq Reads to a Genome
by Monica Britton

both from the UC Davis 2013 Bioinformatics Short Course

RNA-Seq Analysis with Galaxy
by Jeroen F.J. Laros, Wibowo Arindrarto, Leon Mei

from the GCC2013 Training Day

RNA-Seq Analysis with Galaxy
by Curtis Hendrickson, David Crossman, Jeremy Goecks

from the GCC2012 Training Day

Tophat Manual

Agenda

Introduction to Galaxy

Hands-on Analysis

Community Resources

Galaxy on the cloud

Done

Galaxy Resources and Community: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (47 posts in 2013, 3400+ members)

Galaxy-User

Questions about using Galaxy and usegalaxy.org

High volume (1328 posts in 2013, 2600+ members)

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (5200 posts in 2013, 900+ members)

Community: Public Galaxy Instances

http://bit.ly/gxyServers

Interested in:

ChIP-chip and ChIP-seq?

- ✓ Cistrome

Statistical Analysis?

- ✓ Genomic Hyperbrowser

Protein synthesis?

- ✓ GWIPS-viz

de novo assembly?

- ✓ CBIIT Galaxy

Reasoning with ontologies?

- ✓ OPPL Galaxy

Repeats!

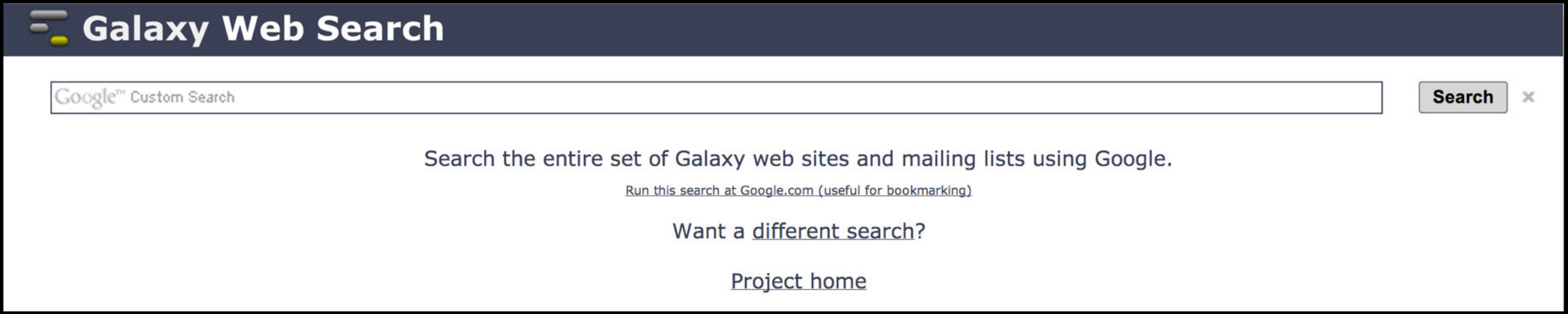
- ✓ RepeatExplorer

Everything?

- ✓ Andromeda

Over 50 public Galaxy servers

Unified Search: http://galaxyproject.org/search



Galaxy Web Search

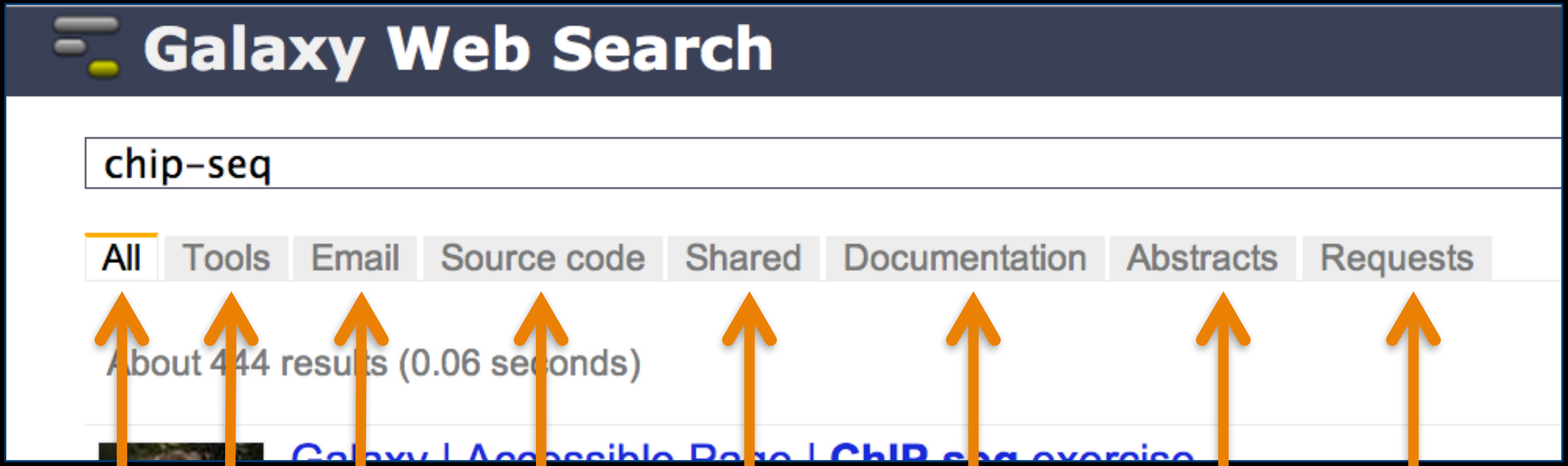
Google™ Custom Search

Search the entire set of Galaxy web sites and mailing lists using Google.

Run this search at Google.com (useful for bookmarking)

Want a different search?

Project home



Galaxy Web Search

chip-seq

About 444 results (0.06 seconds)

Galaxy | Accessible Page | ChIP-seq exercise

Find

Everything on …

Tools for …

Email about …

Source code for …

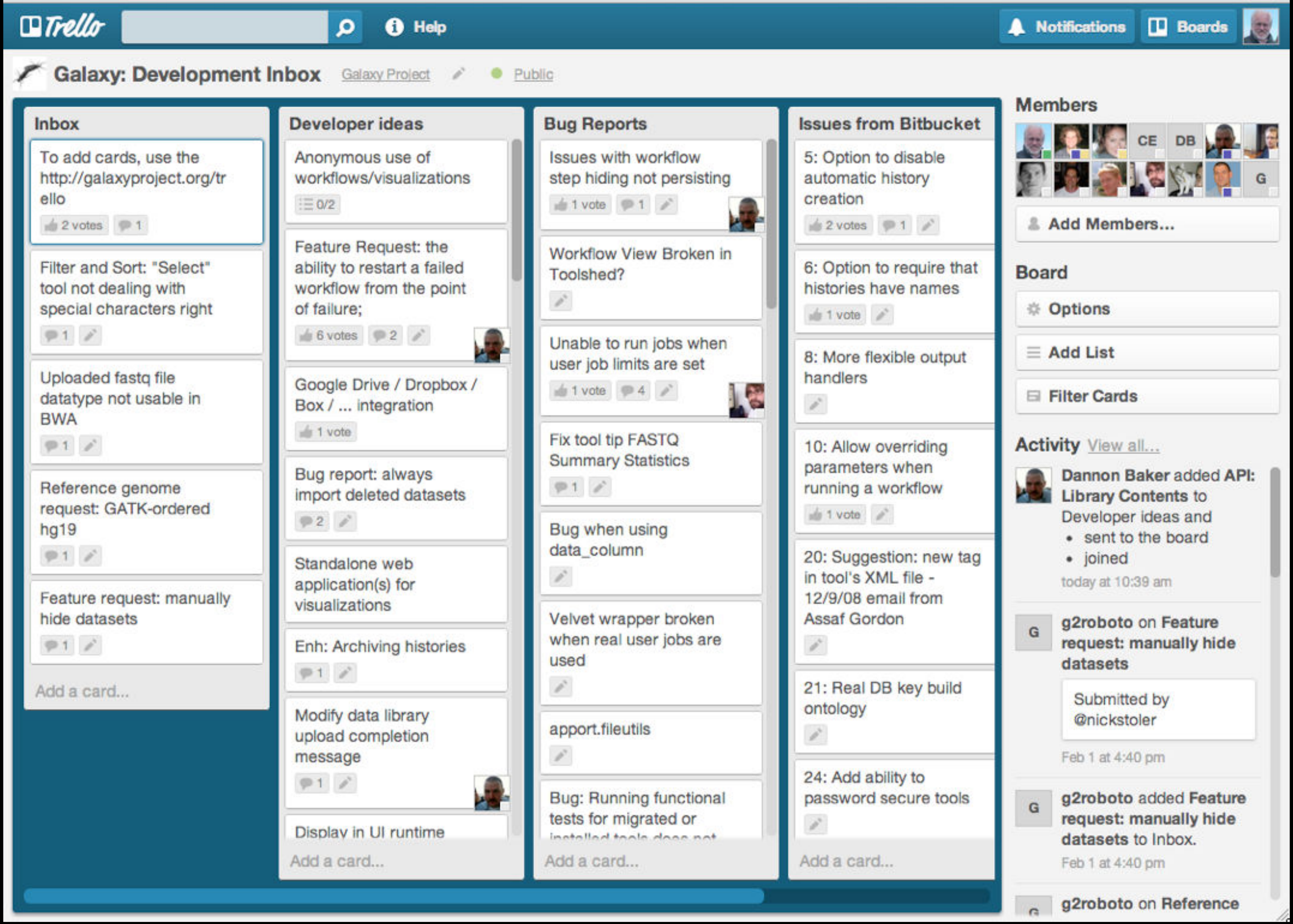
Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

Community can create, vote and comment on issues



Galaxy: Development Inbox Galaxy Project · Public

Inbox

- To add cards, use the http://galaxyproject.org/trello — 2 votes, 1
- Filter and Sort: "Select" tool not dealing with special characters right — 1
- Uploaded fastq file datatype not usable in BWA — 1
- Reference genome request: GATK-ordered hg19 — 1
- Feature request: manually hide datasets — 1

Add a card...

Developer ideas

- Anonymous use of workflows/visualizations — 0/2
- Feature Request: the ability to restart a failed workflow from the point of failure; — 6 votes, 2
- Google Drive / Dropbox / Box / ... integration — 1 vote
- Bug report: always import deleted datasets — 2
- Standalone web application(s) for visualizations
- Enh: Archiving histories — 1
- Modify data library upload completion message — 1
- Display in UI runtime

Add a card...

Bug Reports

- Issues with workflow step hiding not persisting — 1 vote, 1
- Workflow View Broken in Toolshed?
- Unable to run jobs when user job limits are set — 1 vote, 4
- Fix tool tip FASTQ Summary Statistics — 1
- Bug when using data_column
- Velvet wrapper broken when real user jobs are used
- apport.fileutils
- Bug: Running functional tests for migrated or installed tools does not

Add a card...

Issues from Bitbucket

- 5: Option to disable automatic history creation — 2 votes, 1
- 6: Option to require that histories have names — 1 vote
- 8: More flexible output handlers
- 10: Allow overriding parameters when running a workflow — 1 vote
- 20: Suggestion: new tag in tool's XML file - 12/9/08 email from Assaf Gordon
- 21: Real DB key build ontology
- 24: Add ability to password secure tools

Add a card...

Members

Add Members...

Board

- Options
- Add List
- Filter Cards

Activity View all...

Dannon Baker added **API: Library Contents** to Developer ideas and

- sent to the board
- joined

today at 10:39 am

g2roboto on **Feature request: manually hide datasets**

Submitted by @nickstoler

Feb 1 at 4:40 pm

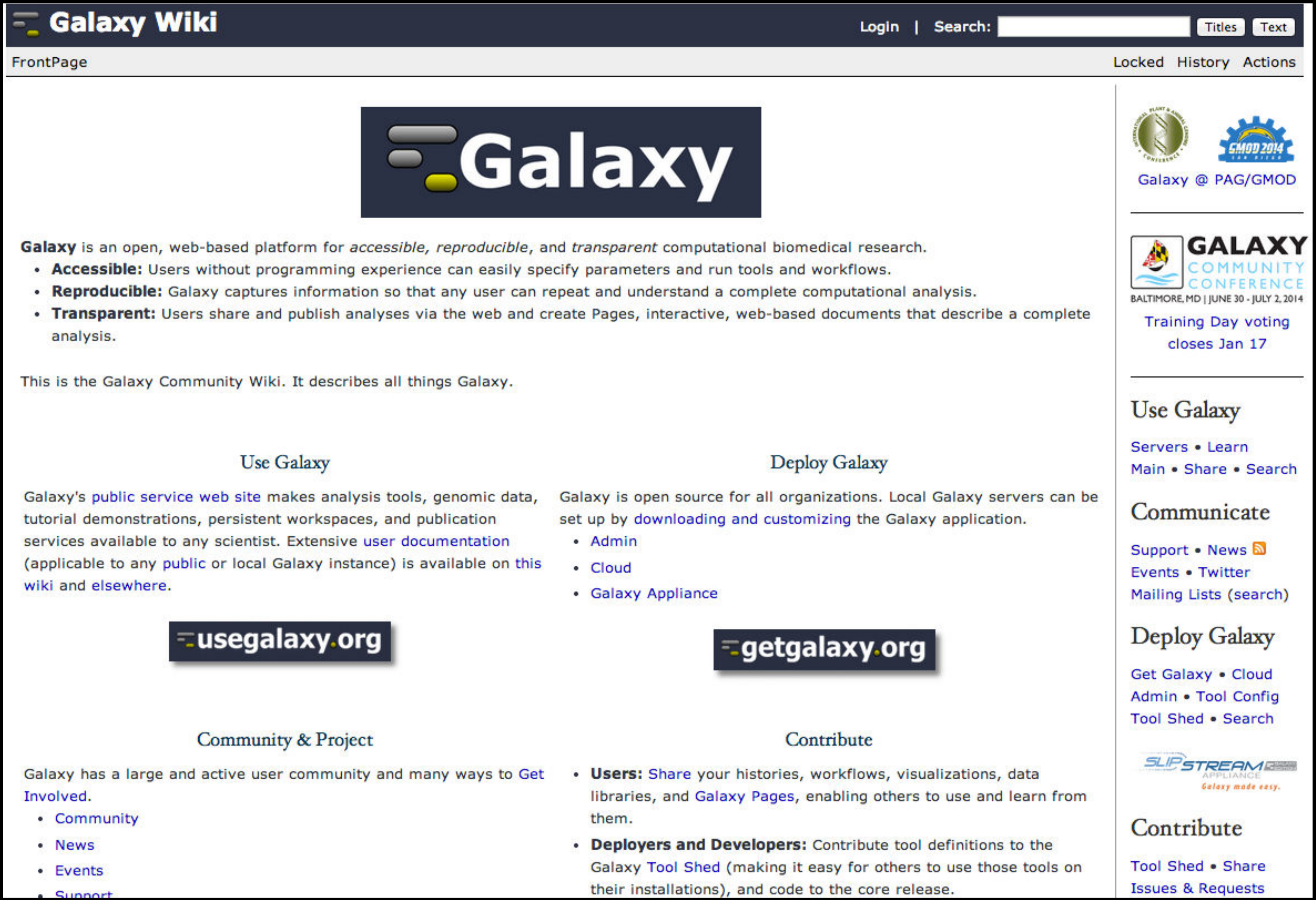
g2roboto added **Feature request: manually hide datasets** to Inbox.

Feb 1 at 4:40 pm

g2roboto on **Reference**

http://bit.ly/gxyissues

http://wiki.galaxyproject.org



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

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

This is the Galaxy Community Wiki. It describes all things Galaxy.

Use Galaxy

Galaxy's public service web site makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive user documentation (applicable to any public or local Galaxy instance) is available on this wiki and elsewhere.

usegalaxy.org

Deploy Galaxy

Galaxy is open source for all organizations. Local Galaxy servers can be set up by downloading and customizing the Galaxy application.

- Admin
- Cloud
- Galaxy Appliance

getgalaxy.org

Community & Project

Galaxy has a large and active user community and many ways to Get Involved.

- Community
- News
- Events
- Support

Contribute

- **Users:** Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.
- **Deployers and Developers:** Contribute tool definitions to the Galaxy Tool Shed (making it easy for others to use those tools on their installations), and code to the core release.

Galaxy @ PAG/GMOD

GALAXY COMMUNITY CONFERENCE
BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Training Day voting closes Jan 17

Use Galaxy

Servers • Learn
Main • Share • Search

Communicate

Support • News
Events • Twitter
Mailing Lists (search)

Deploy Galaxy

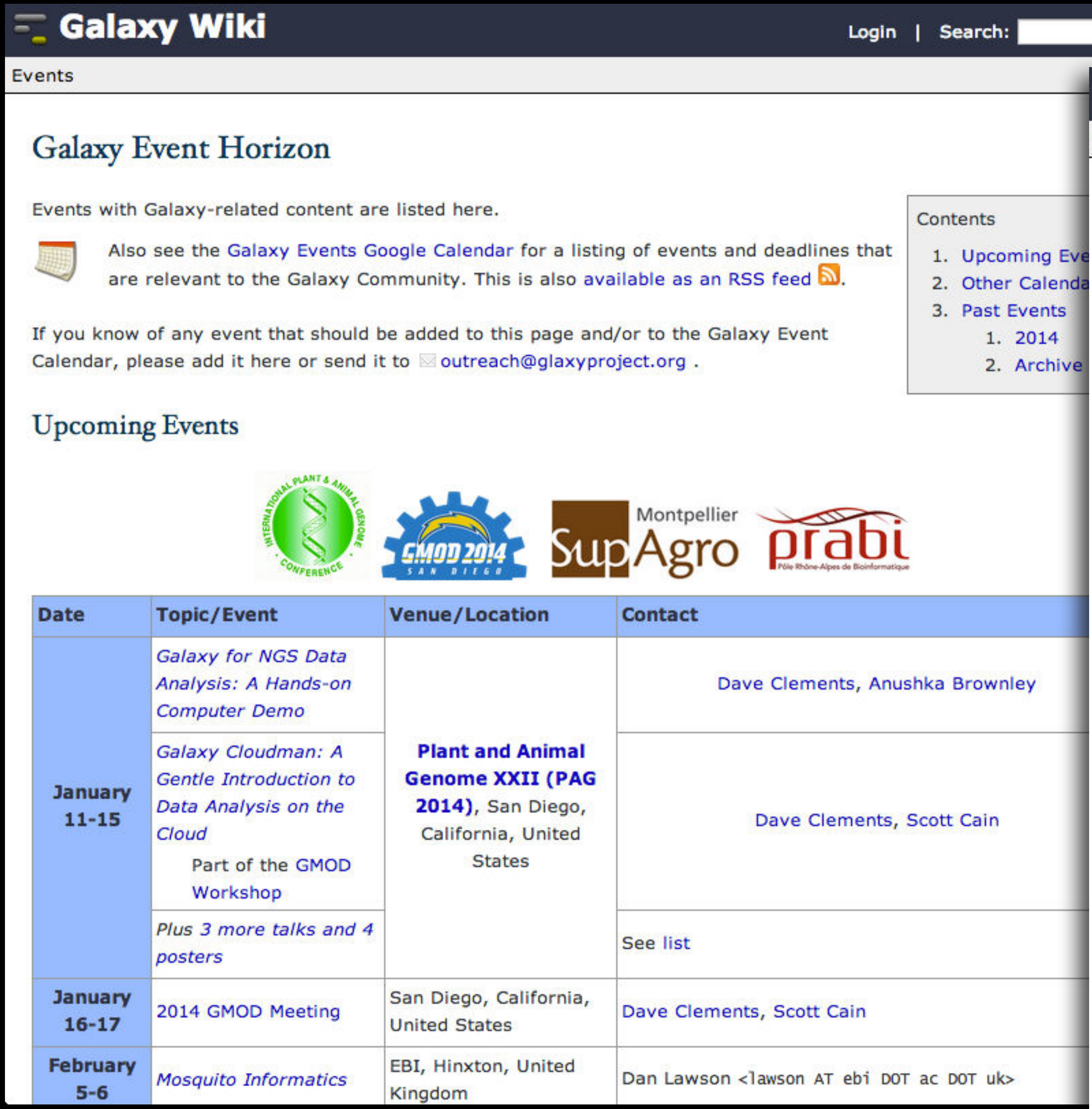
Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

SLIPSTREAM APPLIANCE
Galaxy made easy.

Contribute

Tool Shed • Share
Issues & Requests

Events



Galaxy Event Horizon

Events with Galaxy-related content are listed here.

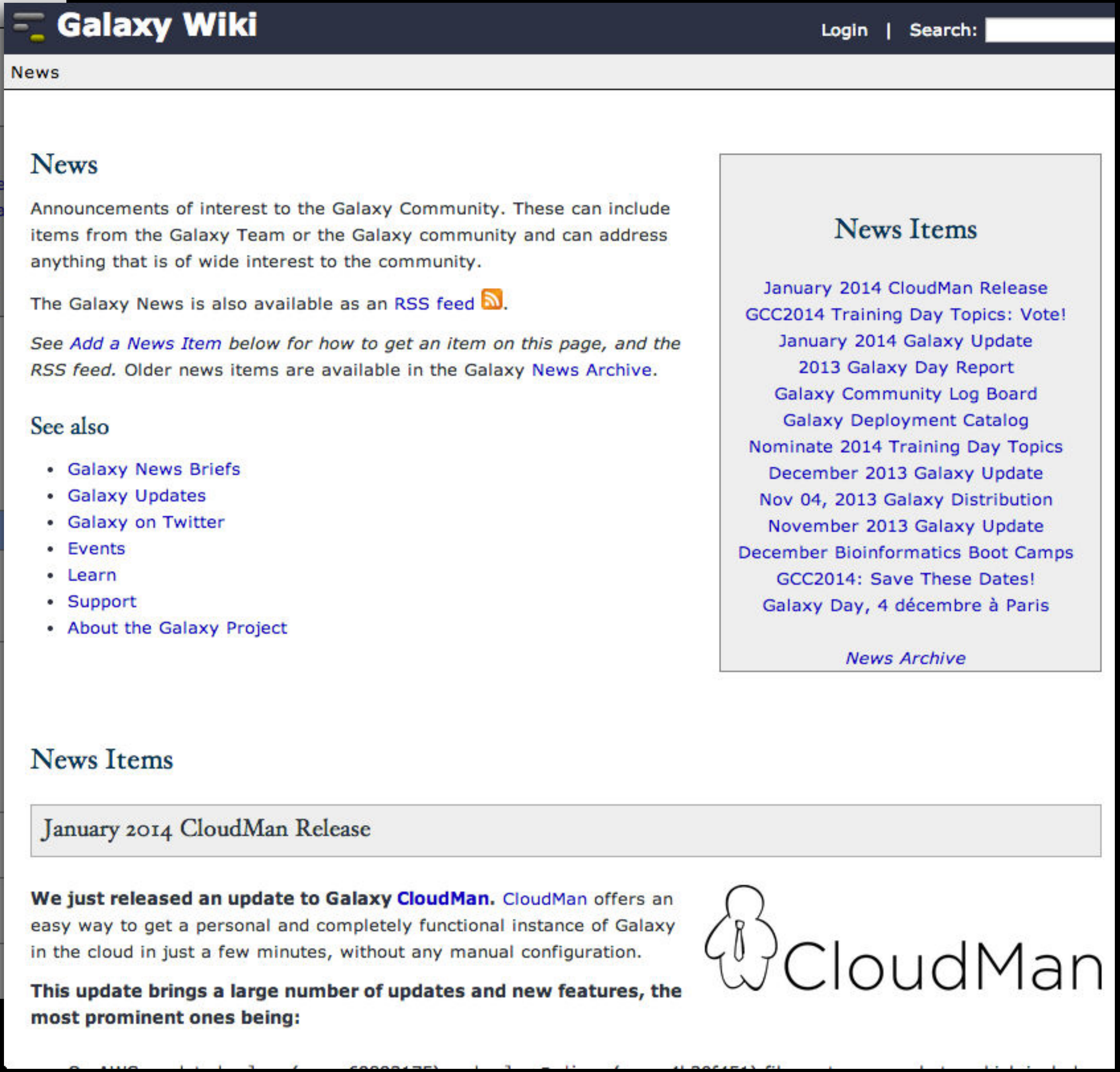
Also see the Galaxy Events Google Calendar for a listing of events and deadlines that are relevant to the Galaxy Community. This is also available as an RSS feed.

If you know of any event that should be added to this page and/or to the Galaxy Event Calendar, please add it here or send it to outreach@glaxyproject.org .

Upcoming Events

Date	Topic/Event	Venue/Location	Contact
January 11-15	*Galaxy for NGS Data Analysis: A Hands-on Computer Demo*	**Plant and Animal Genome XXII (PAG 2014)**, San Diego, California, United States	Dave Clements, Anushka Brownley
	Galaxy Cloudman: A Gentle Introduction to Data Analysis on the Cloud Part of the GMOD Workshop		Dave Clements, Scott Cain
	Plus 3 more talks and 4 posters		See list
January 16-17	2014 GMOD Meeting	San Diego, California, United States	Dave Clements, Scott Cain
February 5-6	*Mosquito Informatics*	EBI, Hinxton, United Kingdom	Dan Lawson <lawson AT ebi DOT ac DOT uk>

News



News

Announcements of interest to the Galaxy Community. These can include items from the Galaxy Team or the Galaxy community and can address anything that is of wide interest to the community.

The Galaxy News is also available as an RSS feed.

See Add a News Item below for how to get an item on this page, and the RSS feed. Older news items are available in the Galaxy News Archive.

See also

- Galaxy News Briefs
- Galaxy Updates
- Galaxy on Twitter
- Events
- Learn
- Support
- About the Galaxy Project

News Items

January 2014 CloudMan Release
GCC2014 Training Day Topics: Vote!
January 2014 Galaxy Update
2013 Galaxy Day Report
Galaxy Community Log Board
Galaxy Deployment Catalog
Nominate 2014 Training Day Topics
December 2013 Galaxy Update
Nov 04, 2013 Galaxy Distribution
November 2013 Galaxy Update
December Bioinformatics Boot Camps
GCC2014: Save These Dates!
Galaxy Day, 4 décembre à Paris

News Archive

News Items

January 2014 CloudMan Release

We just released an update to Galaxy CloudMan. CloudMan offers an easy way to get a personal and completely functional instance of Galaxy in the cloud in just a few minutes, without any manual configuration.

This update brings a large number of updates and new features, the most prominent ones being:

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

http://bit.ly/gcc2014

Galaxy 2011
Community Conference
25-26 May Lunteren, The Netherlands

Galaxy
Community Conference 2012
July 25-27
UIC Forum
University of Illinois at Chicago
http://galaxyproject.org/GCC2012
UIC

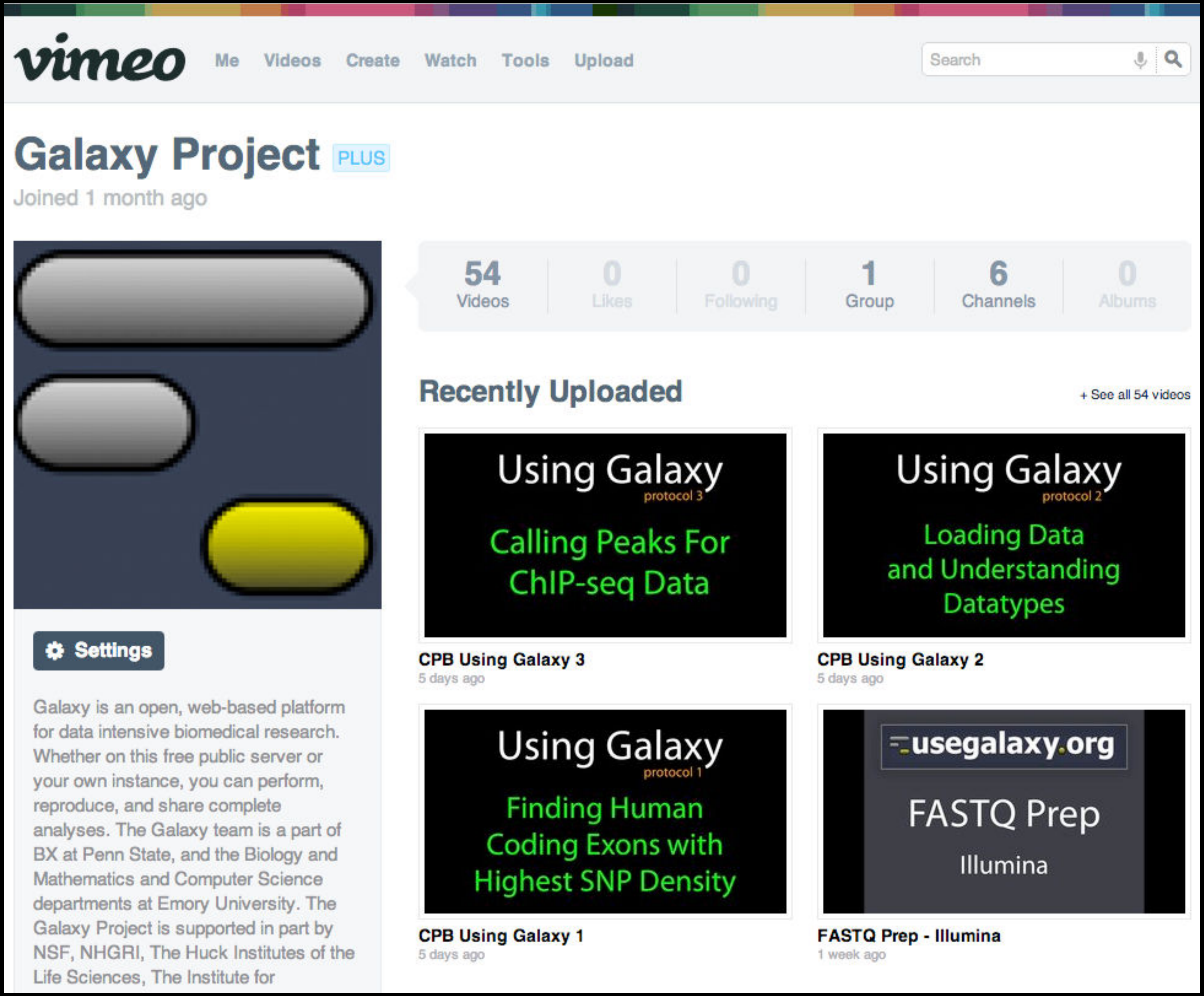
Galaxy
Community Conference
30 June - 2 July
2013
OSLO
UiO : University of Oslo

Galaxy Australasia Workshop 2014

24-25 March Melbourne

http://bit.ly/gaw2014

Galaxy Resources & Community: Videos



Galaxy Project PLUS

Joined 1 month ago

54	0	0	1	6	0
Videos	Likes	Following	Group	Channels	Albums

Recently Uploaded

+ See all 54 videos

CPB Using Galaxy 3
5 days ago

CPB Using Galaxy 2
5 days ago

CPB Using Galaxy 1
5 days ago

FASTQ Prep - Illumina
1 week ago

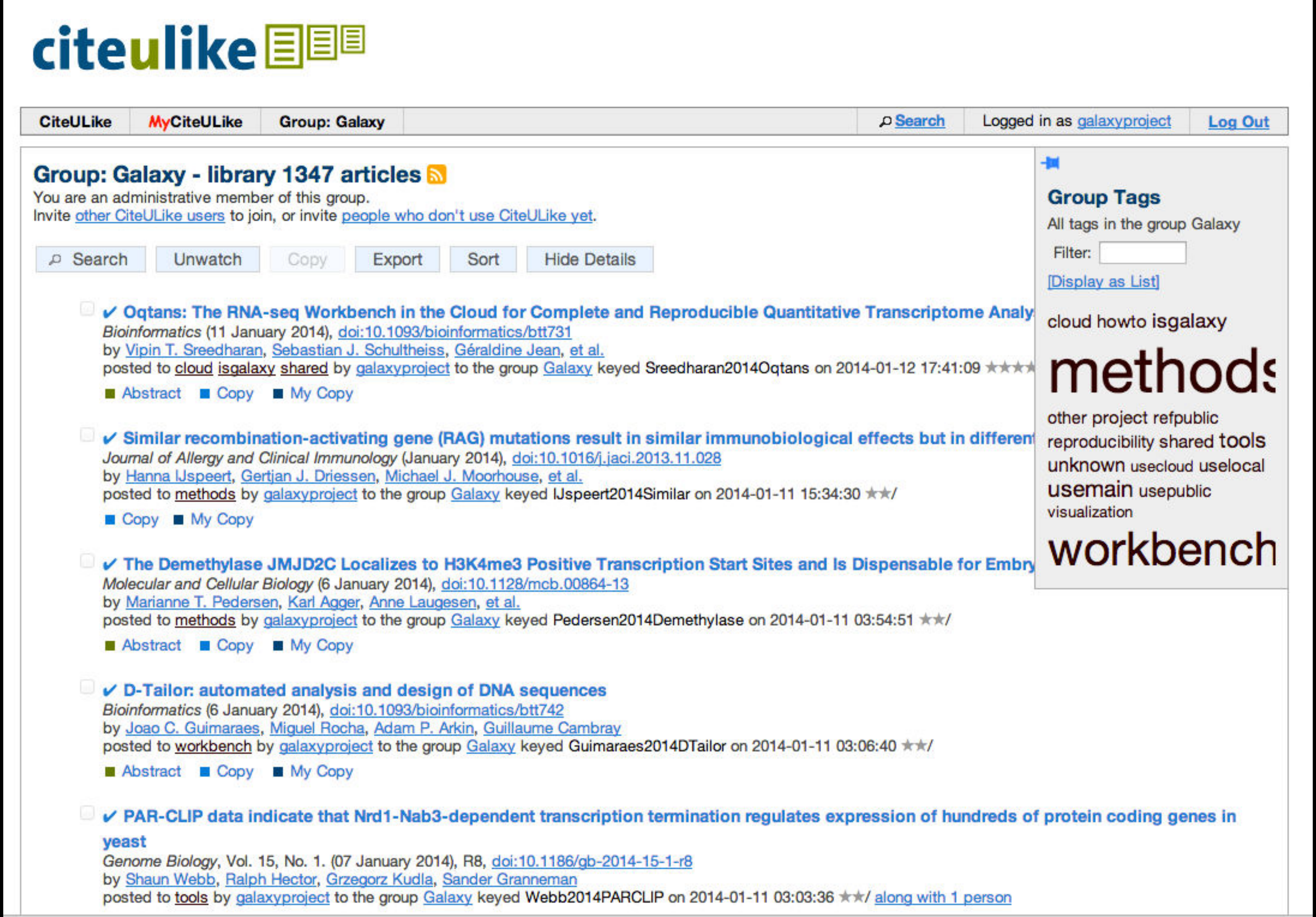
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for

"How to" screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject

Galaxy Resources & Community: CiteULike Group



Group: Galaxy - library 1347 articles
You are an administrative member of this group.
Invite other CiteULike users to join, or invite people who don't use CiteULike yet.

✔ **Oqtans: The RNA-seq Workbench in the Cloud for Complete and Reproducible Quantitative Transcriptome Analy**
Bioinformatics (11 January 2014), doi:10.1093/bioinformatics/btt731
by Vipin T. Sreedharan, Sebastian J. Schultheiss, Géraldine Jean, et al.
posted to cloud isgalaxy shared by galaxyproject to the group Galaxy keyed Sreedharan2014Oqtans on 2014-01-12 17:41:09 ★★★★
■ Abstract ■ Copy ■ My Copy

✔ **Similar recombination-activating gene (RAG) mutations result in similar immunobiological effects but in different**
Journal of Allergy and Clinical Immunology (January 2014), doi:10.1016/j.jaci.2013.11.028
by Hanna IJspeert, Gertjan J. Driessen, Michael J. Moorhouse, et al.
posted to methods by galaxyproject to the group Galaxy keyed IJspeert2014Similar on 2014-01-11 15:34:30 ★★/
■ Copy ■ My Copy

✔ **The Demethylase JMJD2C Localizes to H3K4me3 Positive Transcription Start Sites and Is Dispensable for Embry**
Molecular and Cellular Biology (6 January 2014), doi:10.1128/mcb.00864-13
by Marianne T. Pedersen, Karl Agger, Anne Laugesen, et al.
posted to methods by galaxyproject to the group Galaxy keyed Pedersen2014Demethylase on 2014-01-11 03:54:51 ★★/
■ Abstract ■ Copy ■ My Copy

✔ **D-Tailor: automated analysis and design of DNA sequences**
Bioinformatics (6 January 2014), doi:10.1093/bioinformatics/btt742
by Joao C. Guimaraes, Miguel Rocha, Adam P. Arkin, Guillaume Cambray
posted to workbench by galaxyproject to the group Galaxy keyed Guimaraes2014DTailor on 2014-01-11 03:06:40 ★★/
■ Abstract ■ Copy ■ My Copy

✔ **PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast**
Genome Biology, Vol. 15, No. 1. (07 January 2014), R8, doi:10.1186/gb-2014-15-1-r8
by Shaun Webb, Ralph Hector, Grzegorz Kudla, Sander Granneman
posted to tools by galaxyproject to the group Galaxy keyed Webb2014PARCLIP on 2014-01-11 03:03:36 ★★/ along with 1 person

Group Tags
All tags in the group Galaxy
Filter:
[Display as List]

cloud howto isgalaxy methods other project refpublic reproducibility shared tools unknown usecloud uselocal usemain usepublic visualization workbench

(Now)
Over
1400
papers

17
different
tags

http://bit.ly/gxycul

Agenda

Introduction to Galaxy

Hands-on Analysis

Community Resources

Galaxy on the cloud

Done

Galaxy is available ...

- As a free (for everyone) web service
- As open source software
- *On the Cloud*

amazon web services™

openstack™ CLOUD SOFTWARE

OpenNebula.org The Open Source Toolkit for Cloud Computing

http://wiki.galaxyproject.org/Cloud

AWS in Education Grants Program

amazon
web services™

http://aws.amazon.com/education

What is our path?

Today we will:

- Launch our own Galaxy server on AWS
- Make the server dynamically scalable in response to demand.
- Run some basic analysis on it.
- Make it go away.

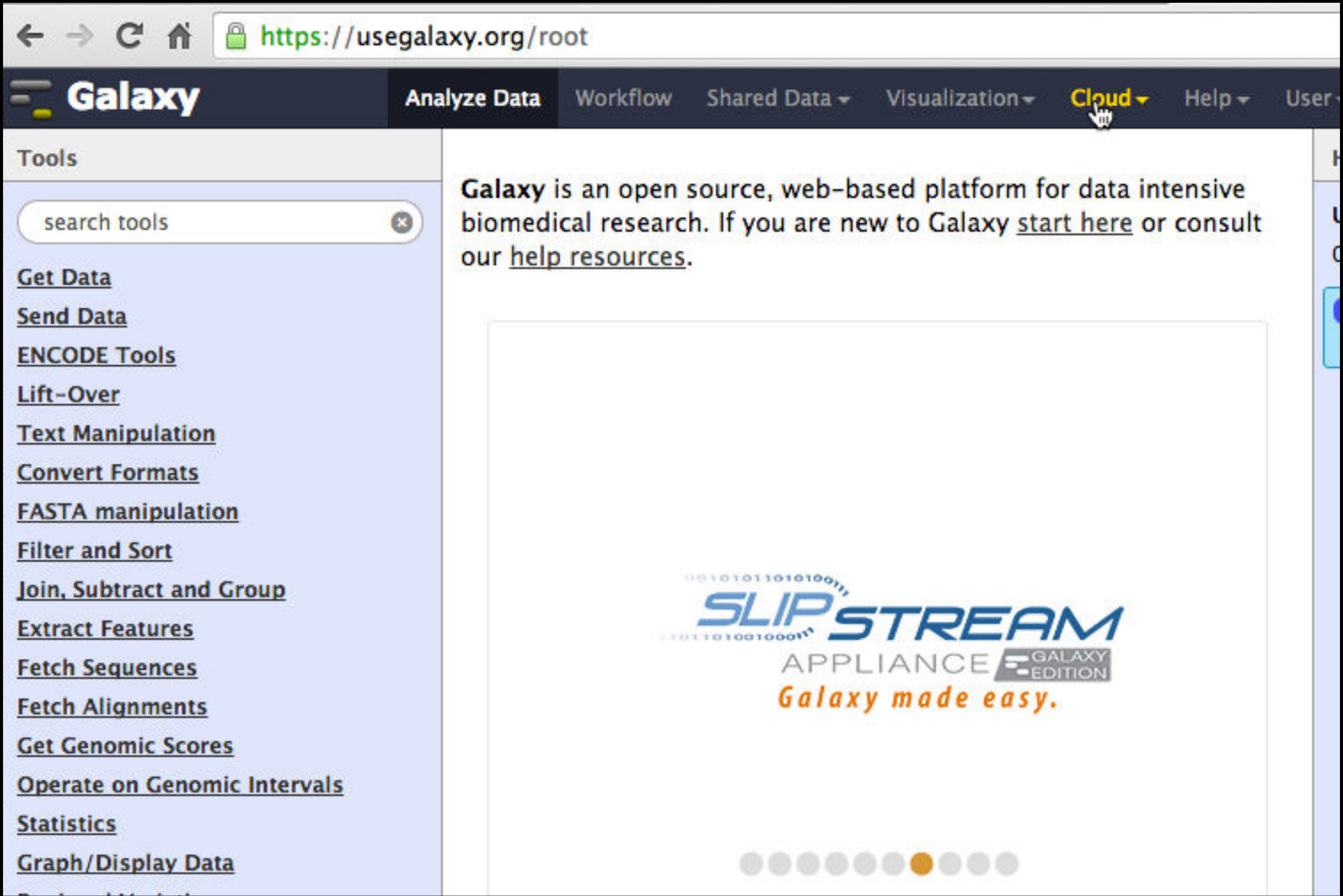
Full Disclosure

To use AWS you must create an AWS account with a credit card associated with it.

You must also have created a key pair.

We will use the IAM account for this workshop.

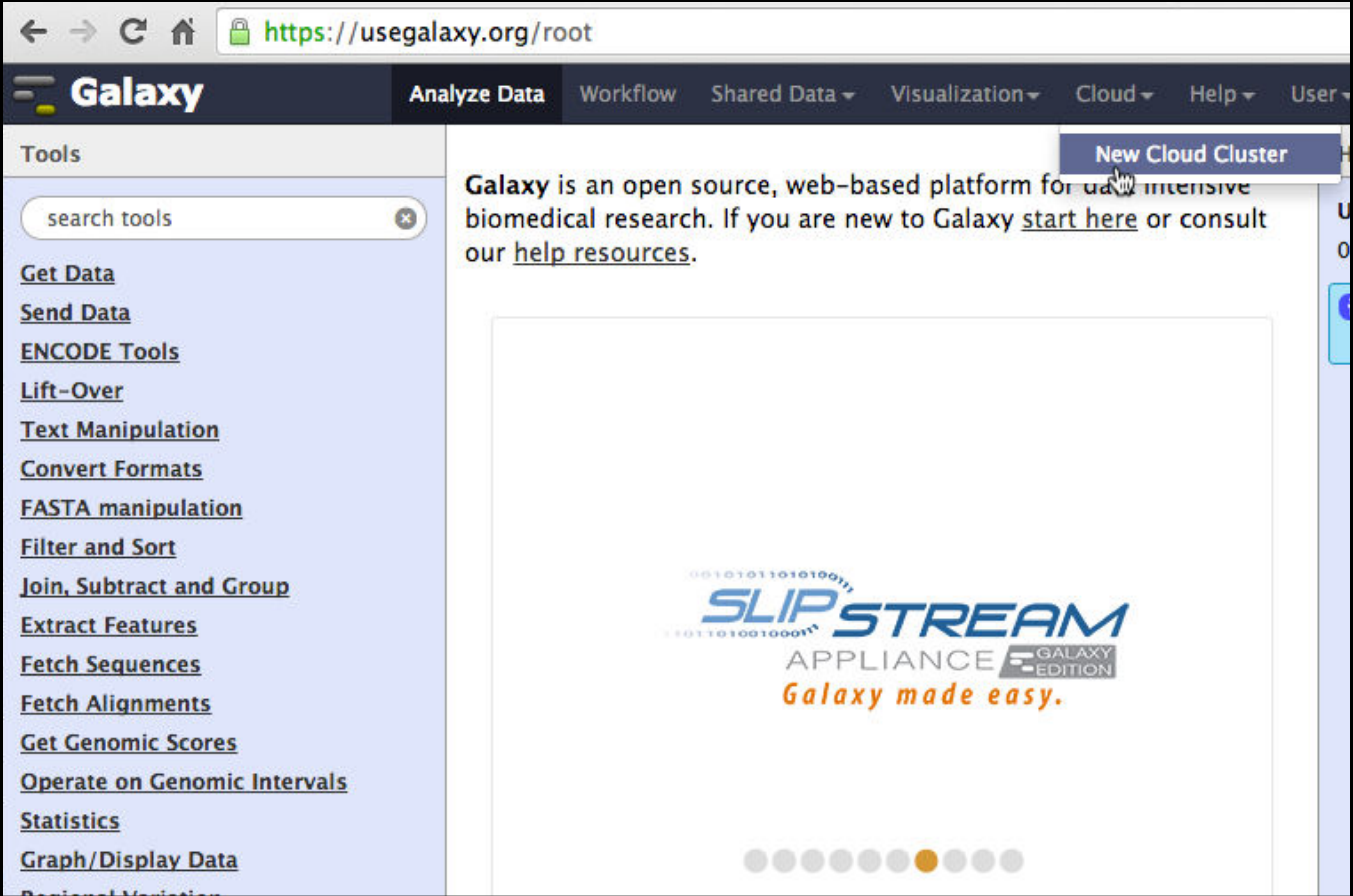
CloudLaunch



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

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

CloudLaunch

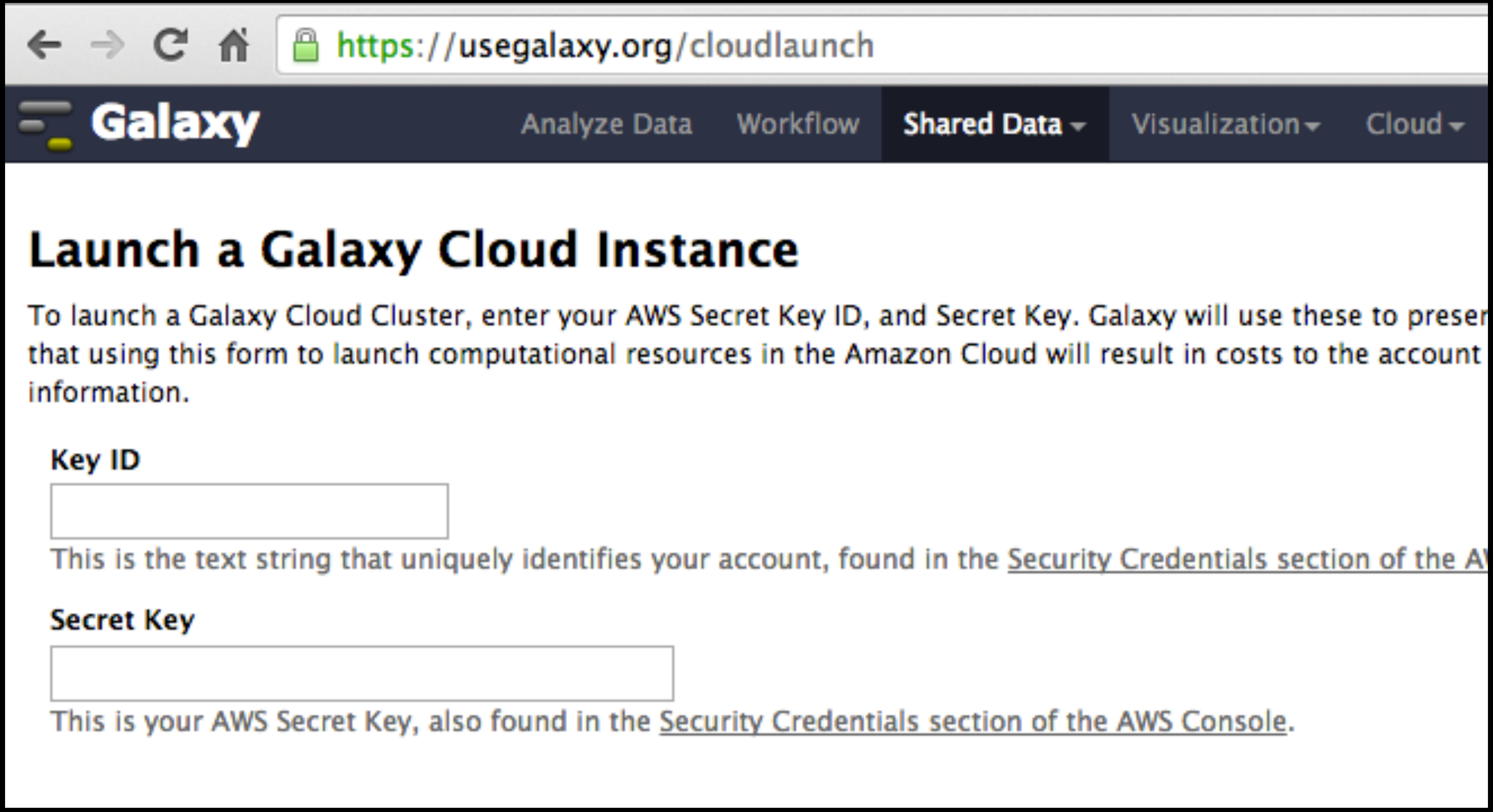


https://usegalaxy.org/root

- 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

Galaxy is an open source, web-based platform for da[illegible] intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

CloudLaunch



Launch a Galaxy Cloud Instance

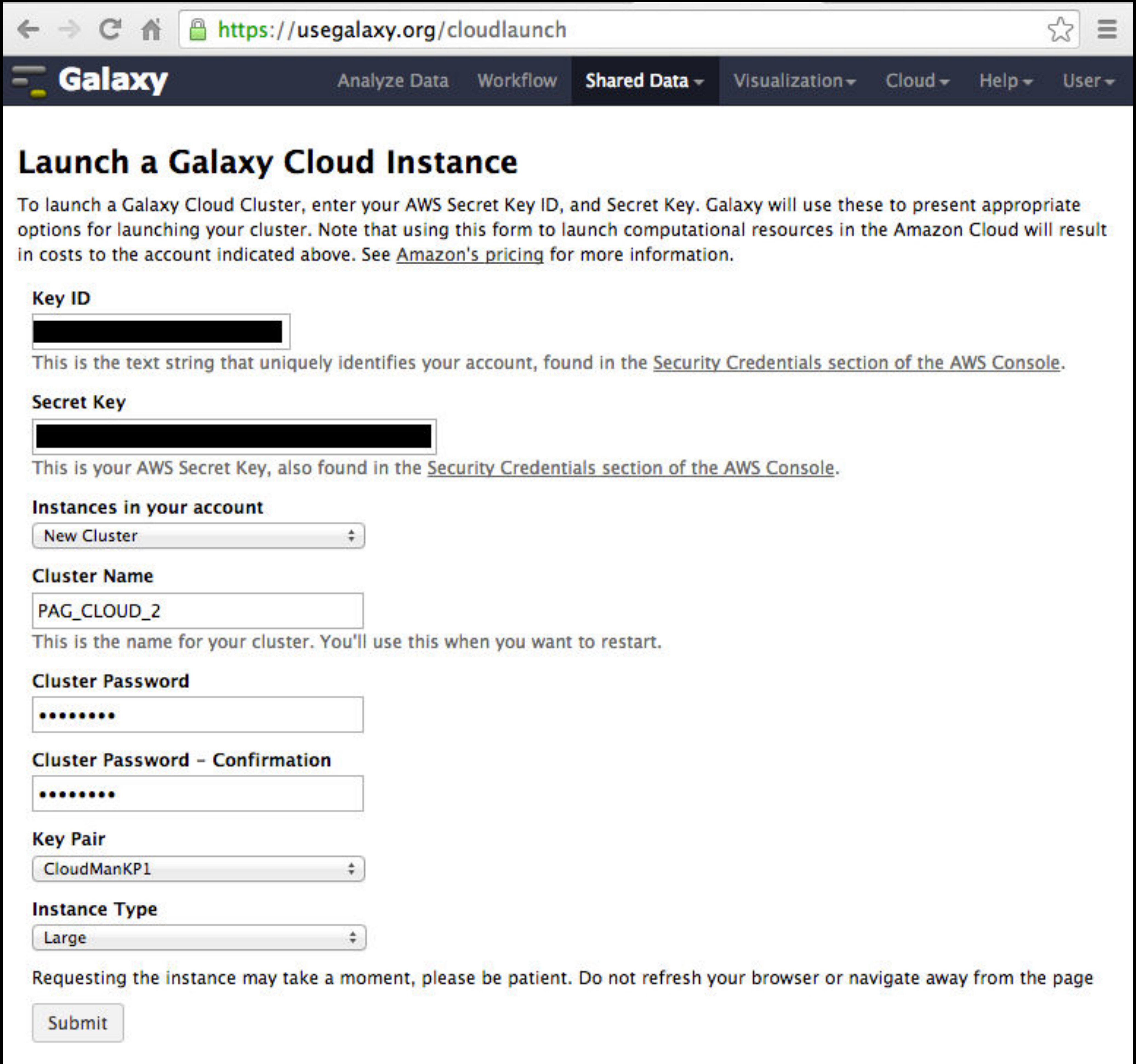
To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to preser
that using this form to launch computational resources in the Amazon Cloud will result in costs to the account
information.

Key ID

This is the text string that uniquely identifies your account, found in the Security Credentials section of the A

Secret Key

This is your AWS Secret Key, also found in the Security Credentials section of the AWS Console.



Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See Amazon's pricing for more information.

Key ID

[redacted]

This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Console.

Secret Key

[redacted]

This is your AWS Secret Key, also found in the Security Credentials section of the AWS Console.

Instances in your account

New Cluster

Cluster Name

PAG_CLOUD_2

This is the name for your cluster. You'll use this when you want to restart.

Cluster Password

••••••••

Cluster Password - Confirmation

••••••••

Key Pair

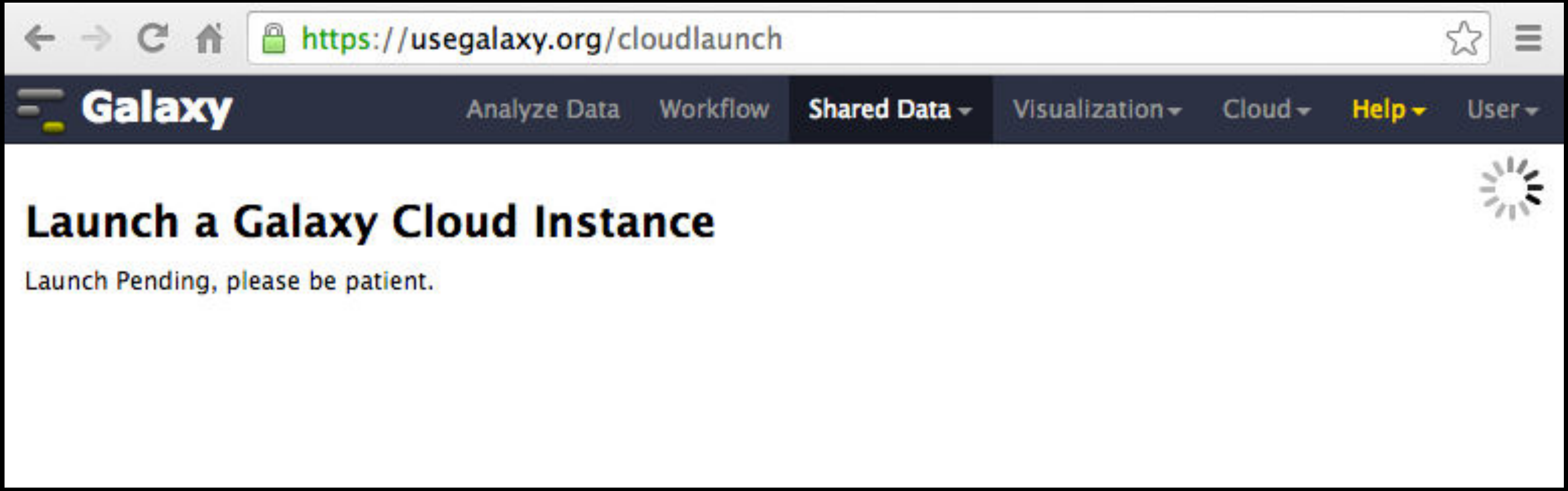
CloudManKP1

Instance Type

Large

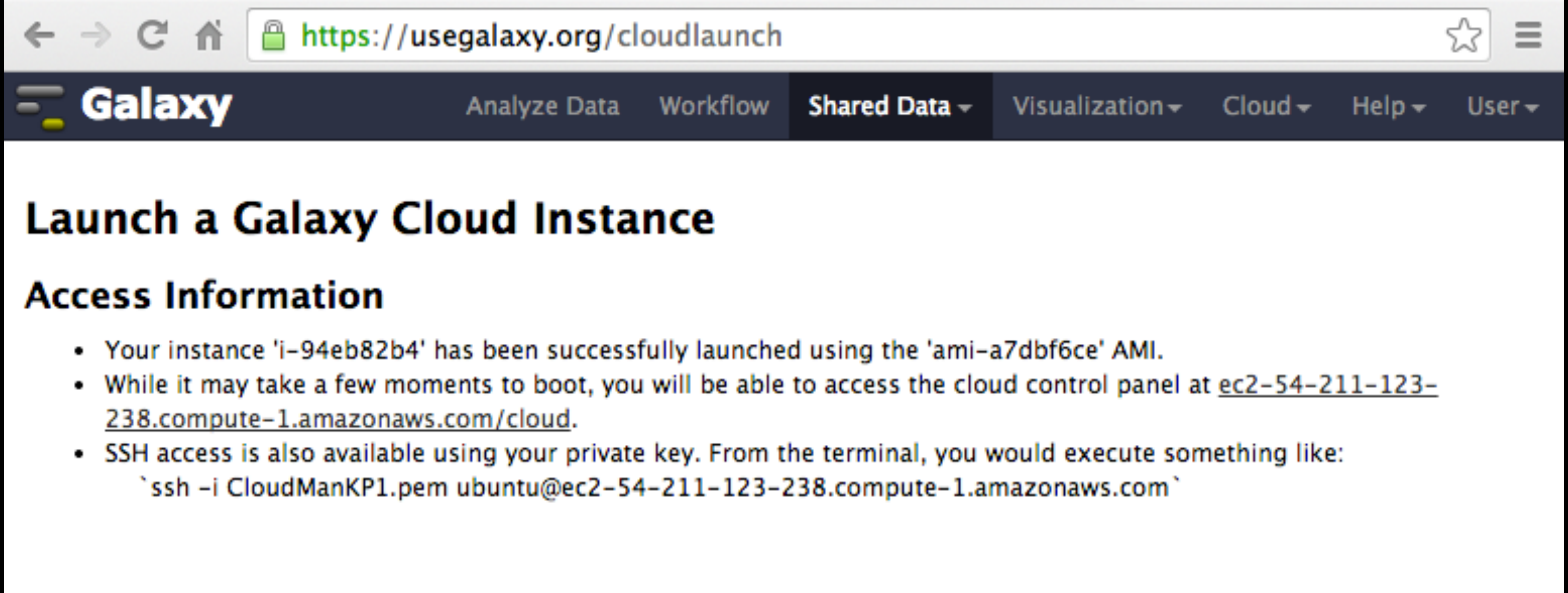
Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page

CloudLaunch



Launch a Galaxy Cloud Instance

Launch Pending, please be patient.

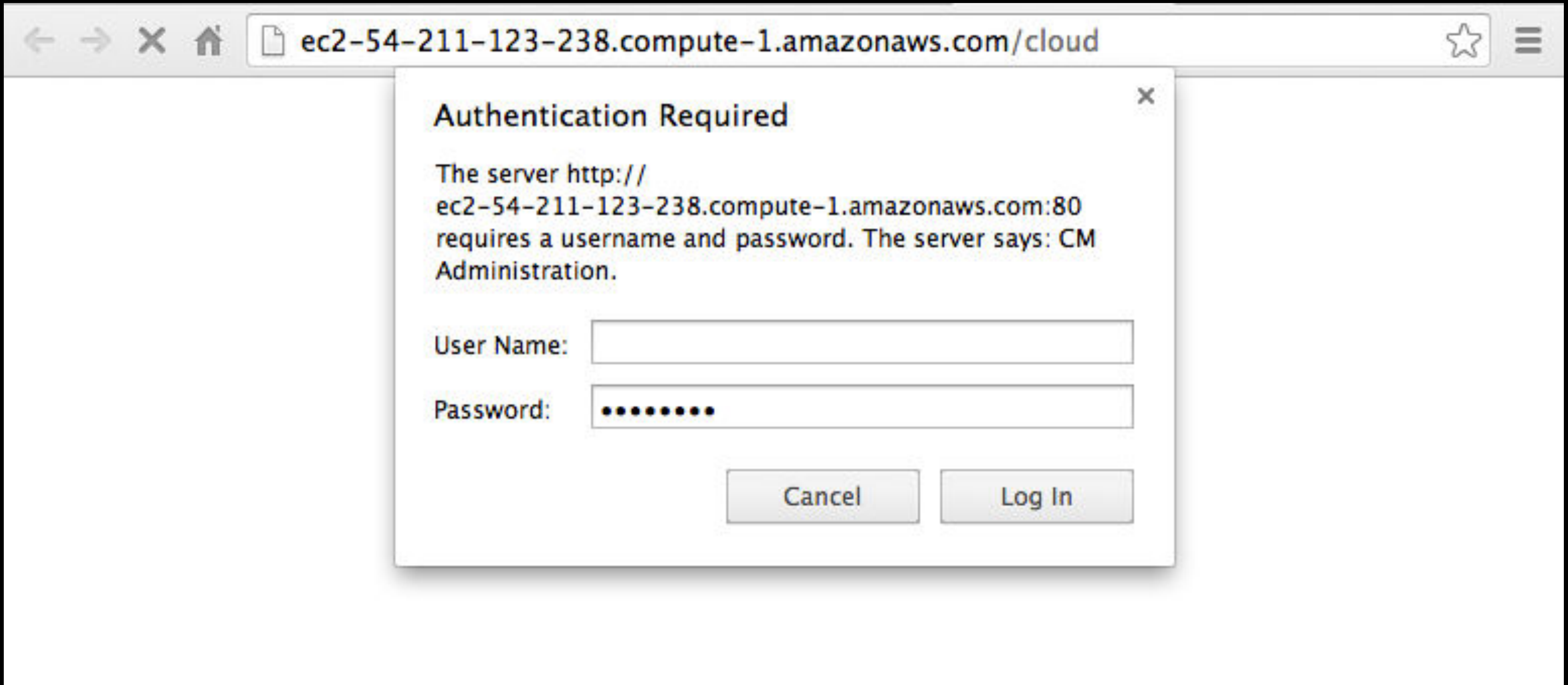


Launch a Galaxy Cloud Instance

Access Information

- Your instance 'i-94eb82b4' has been successfully launched using the 'ami-a7dbf6ce' AMI.
- While it may take a few moments to boot, you will be able to access the cloud control panel at ec2-54-211-123-238.compute-1.amazonaws.com/cloud.
- SSH access is also available using your private key. From the terminal, you would execute something like:
  `ssh -i CloudManKP1.pem ubuntu@ec2-54-211-123-238.compute-1.amazonaws.com`

CloudLaunch



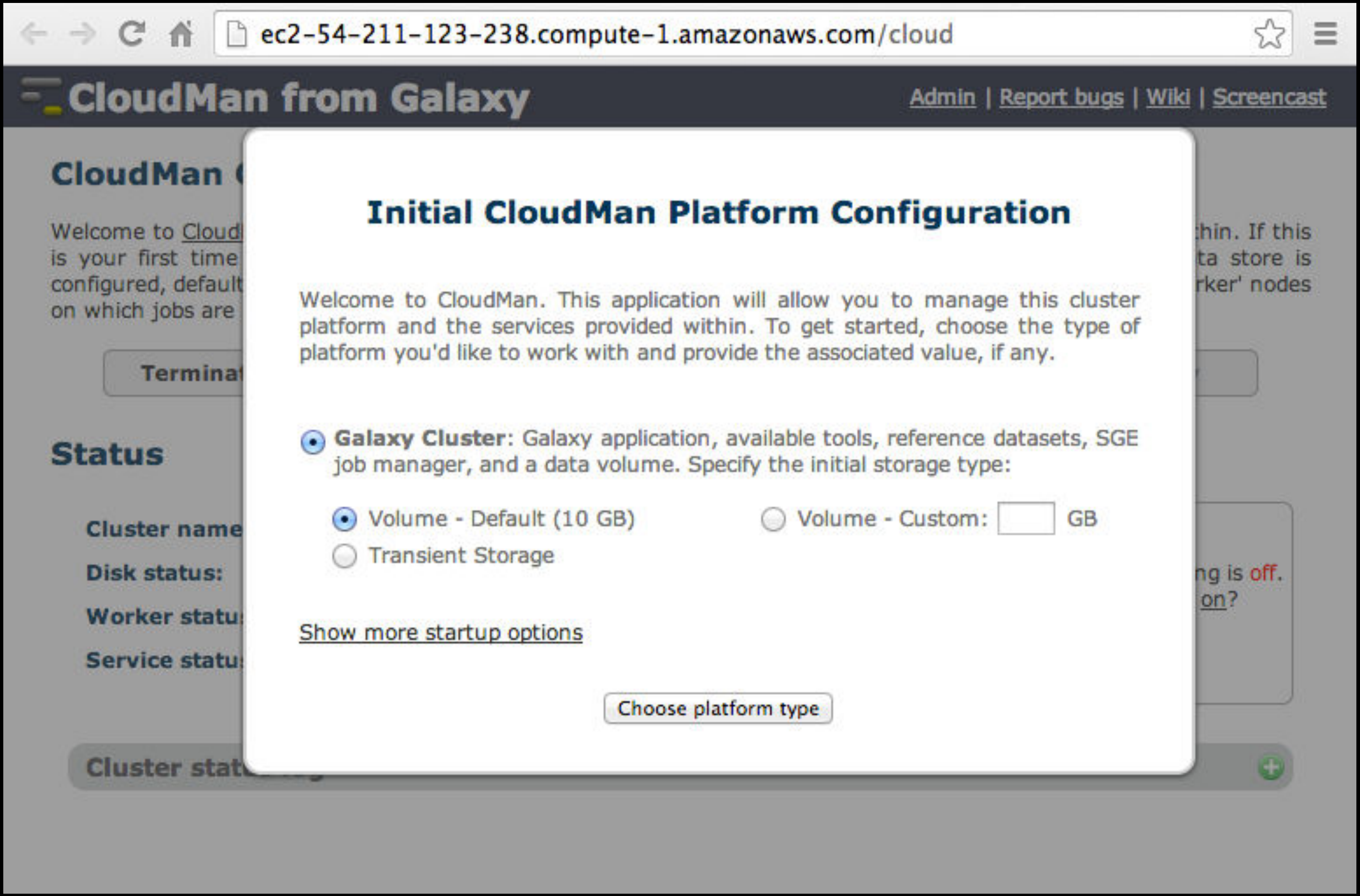
Authentication Required

The server http://ec2-54-211-123-238.compute-1.amazonaws.com:80 requires a username and password. The server says: CM Administration.

User Name:

Password: •••••••••

CloudLaunch



ec2-54-211-123-238.compute-1.amazonaws.com/cloud

CloudMan from Galaxy

CloudMan (

Welcome to Cloud… is your first time… configured, default… on which jobs are… …hin. If this …ta store is …rker' nodes

Status

Cluster name…
Disk status:
Worker statu…
Service statu…

…ng is off.
…on?

Cluster stat…

Initial CloudMan Platform Configuration

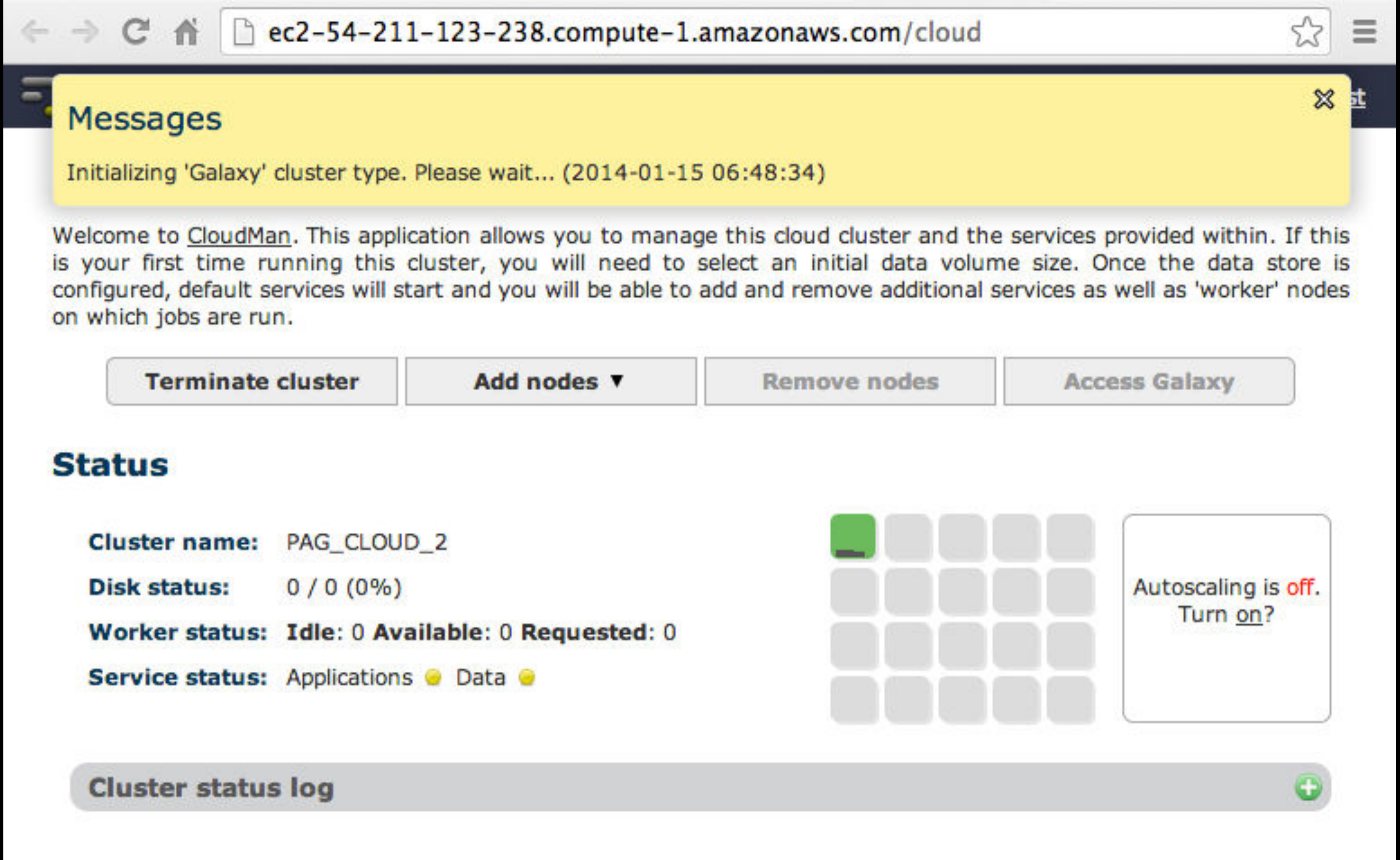
Welcome to CloudMan. This application will allow you to manage this cluster platform and the services provided within. To get started, choose the type of platform you'd like to work with and provide the associated value, if any.

● **Galaxy Cluster**: Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage type:

● Volume - Default (10 GB) ○ Volume - Custom: ____ GB
○ Transient Storage

Show more startup options

CloudLaunch



Messages

Initializing 'Galaxy' cluster type. Please wait... (2014-01-15 06:48:34)

Welcome to CloudMan. This application allows you to manage this cloud cluster and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

Cluster name: PAG_CLOUD_2

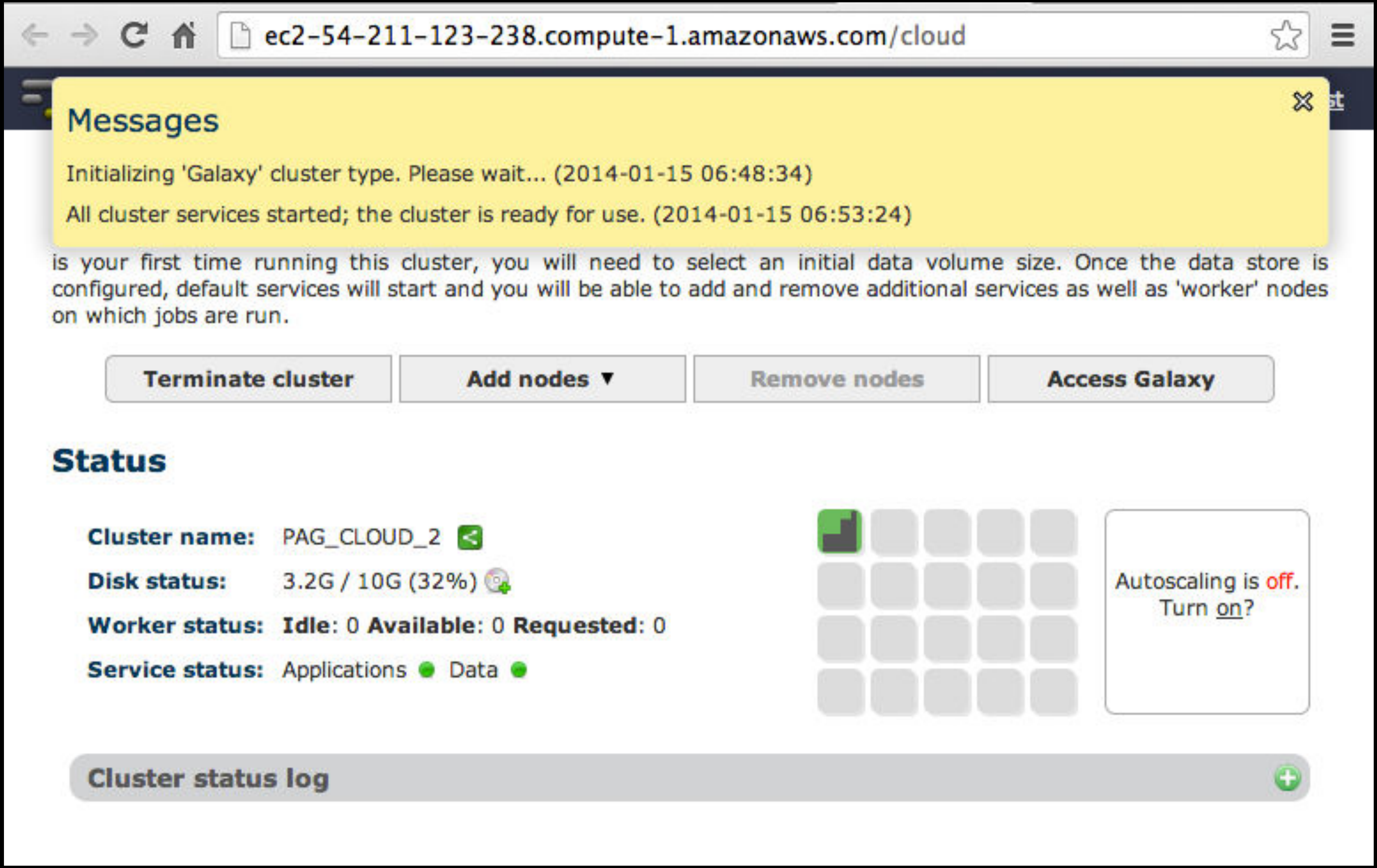
Disk status: 0 / 0 (0%)

Worker status: **Idle**: 0 **Available**: 0 **Requested**: 0

Service status: Applications ● Data ●

Autoscaling is off. Turn on?

Cloud Launched



Messages

Initializing 'Galaxy' cluster type. Please wait... (2014-01-15 06:48:34)

All cluster services started; the cluster is ready for use. (2014-01-15 06:53:24)

is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

Cluster name: PAG_CLOUD_2

Disk status: 3.2G / 10G (32%)

Worker status: **Idle**: 0 **Available**: 0 **Requested**: 0

Service status: Applications ● Data ●

Autoscaling is off. Turn on?

Cool things to do

- Create a login
- Become an admin
- Set up autoscaling
- Run ~ Galaxy 101
    - http://usegalaxy.org/galaxy101
- Shut it down

Basic Analysis

Which genes have most overlapping Repeats?

http://cloud2.galaxyproject.org/

http://cloud3.galaxyproject.org/

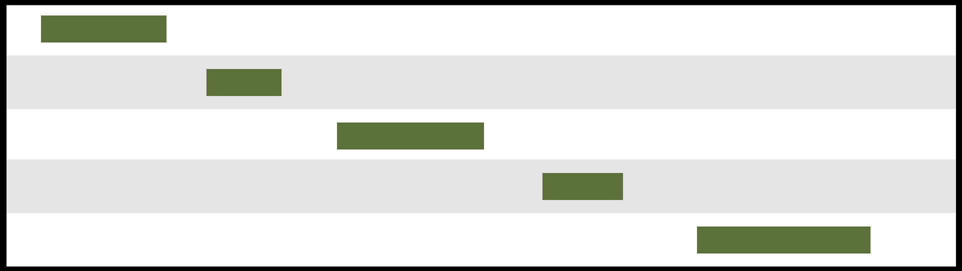
(~ http://usegalaxy.org/galaxy101)

Genes & Repeats: A General Plan

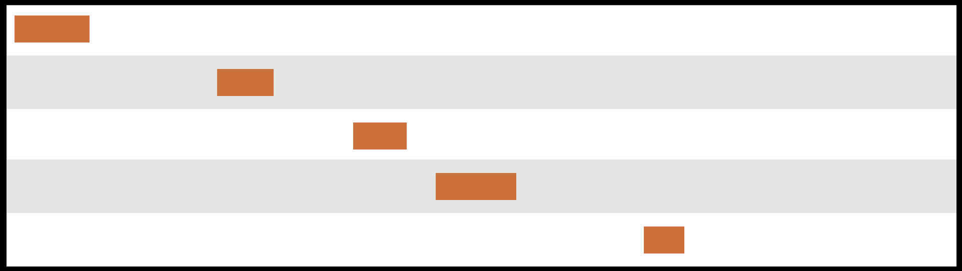
- Get some data
  - Get Data → UCSC Table Browser
- Identify which genes/exons have Repeats
- Count Repeats per exon
- Visualize, save, download, ... exons with most Repeats

http://cloud2.galaxyproject.org/
http://cloud3.galaxyproject.org/

(~ http://usegalaxy.org/galaxy101)

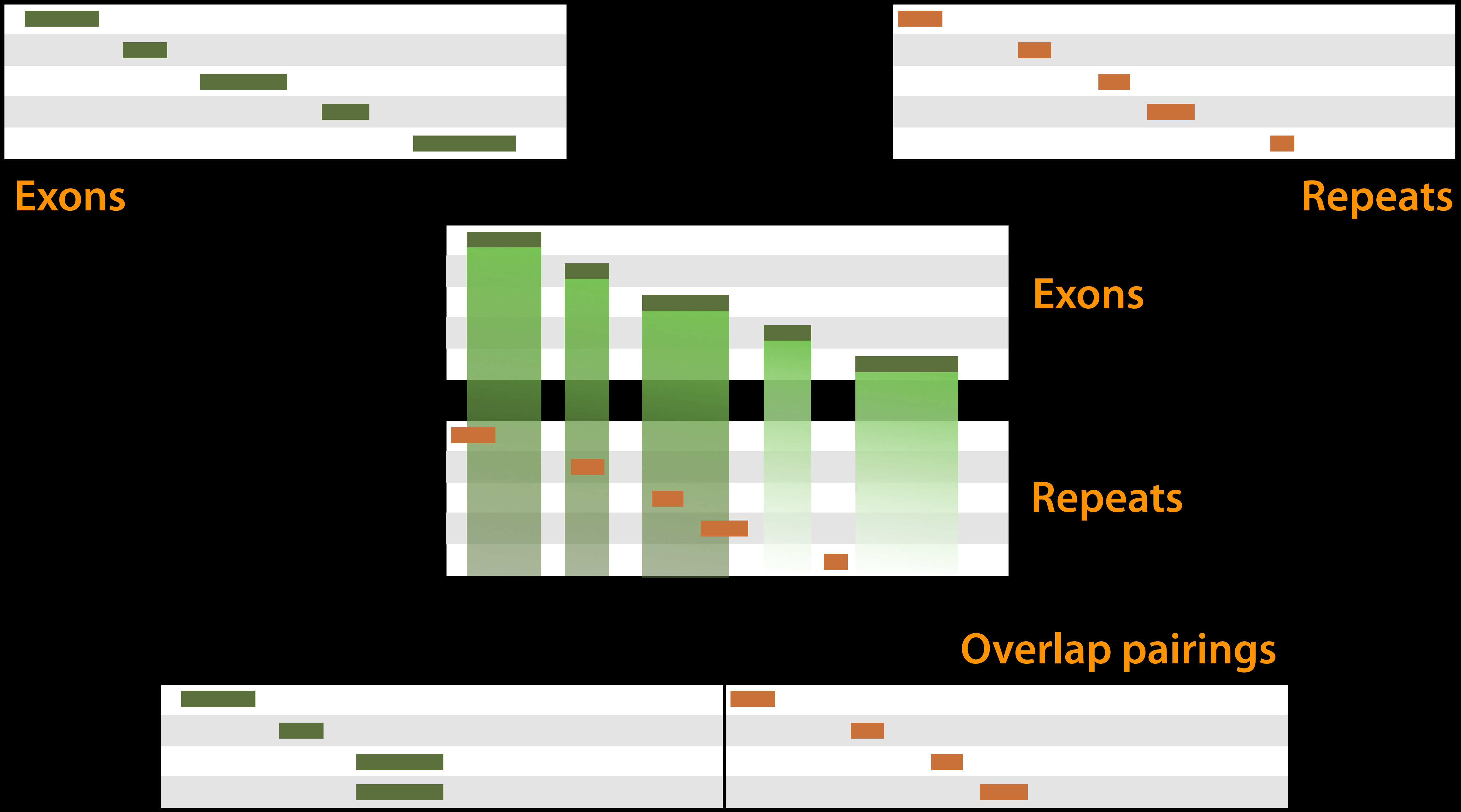


Exons



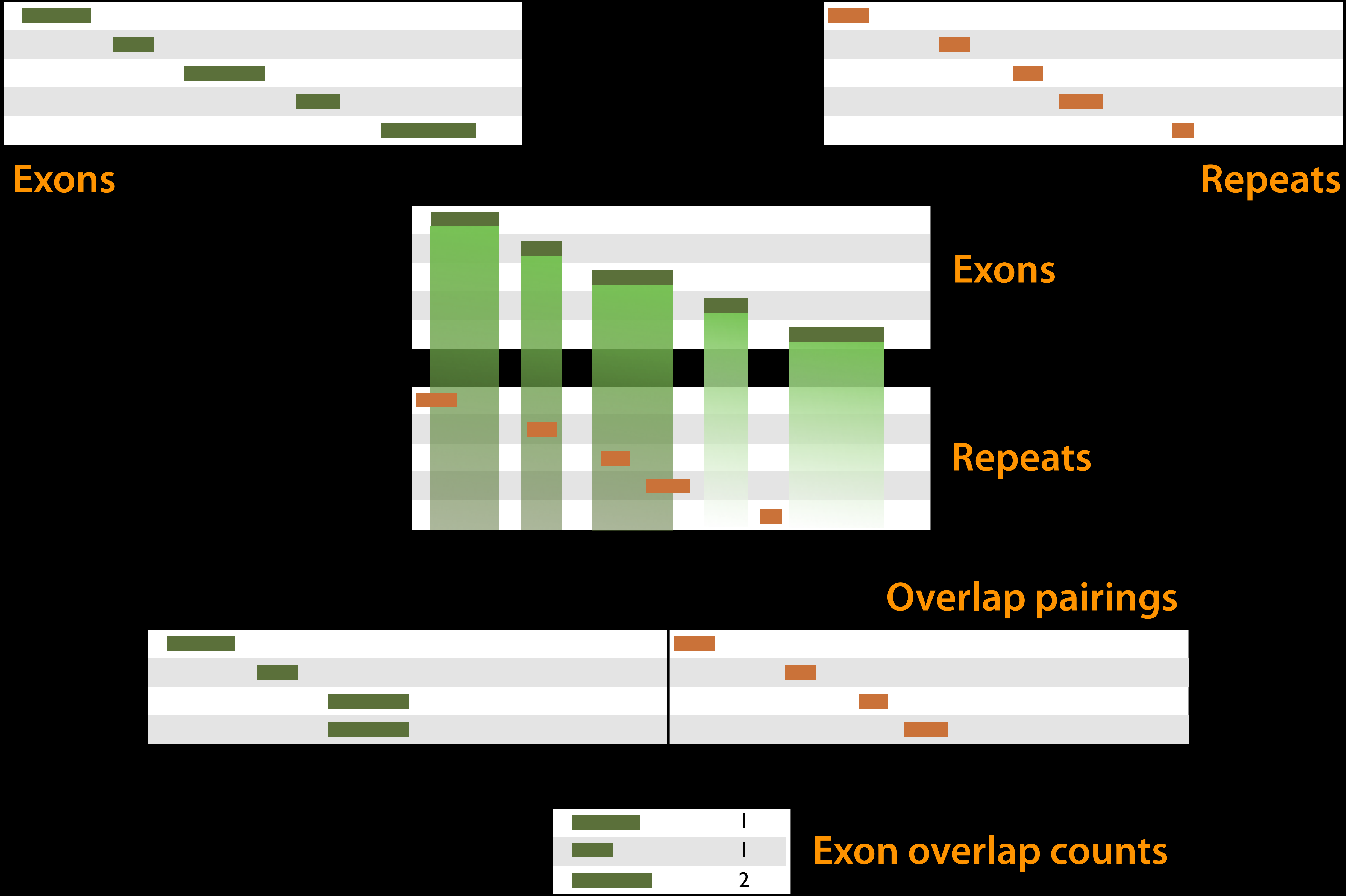
Repeats

(Identify which genes/exons have Repeats)



Operate on Genomic Intervals → **Join**

(Identify which genes/exons have Repeats)



Exons

Repeats

Exons

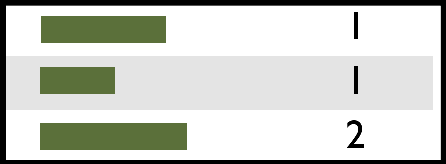
Repeats

Overlap pairings

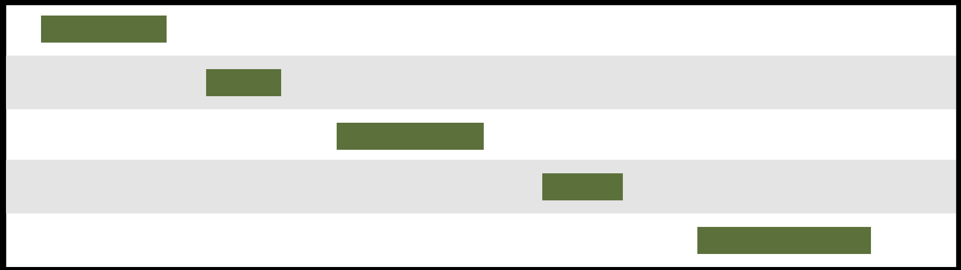
Exon overlap counts

Join, Subtract, and Group → Group

(Count Repeats per exon)



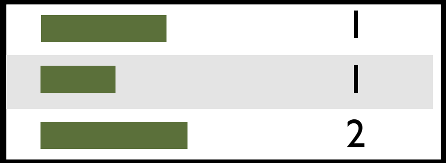
Exon overlap counts



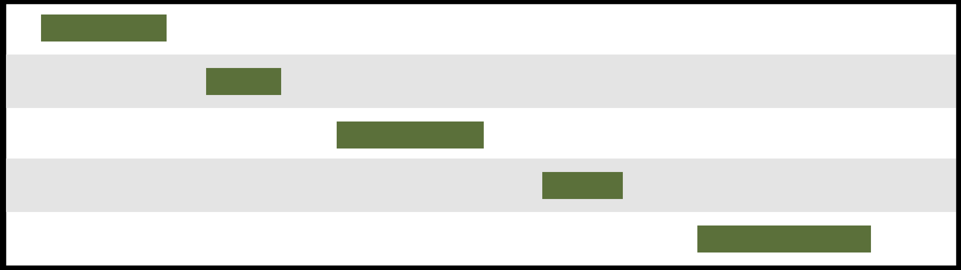
Exons

We’ve answered our question, but we can do better.

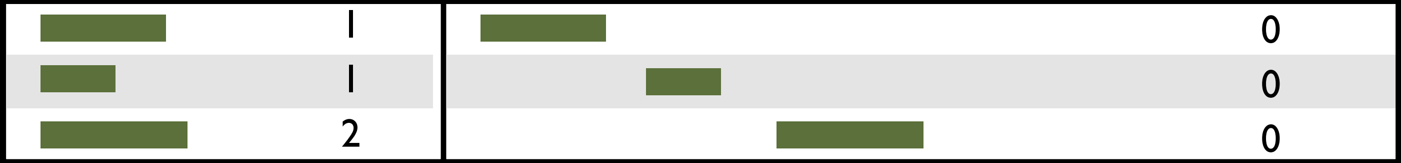
Incorporate the overlap count with rest of Exon information



Exon overlap counts



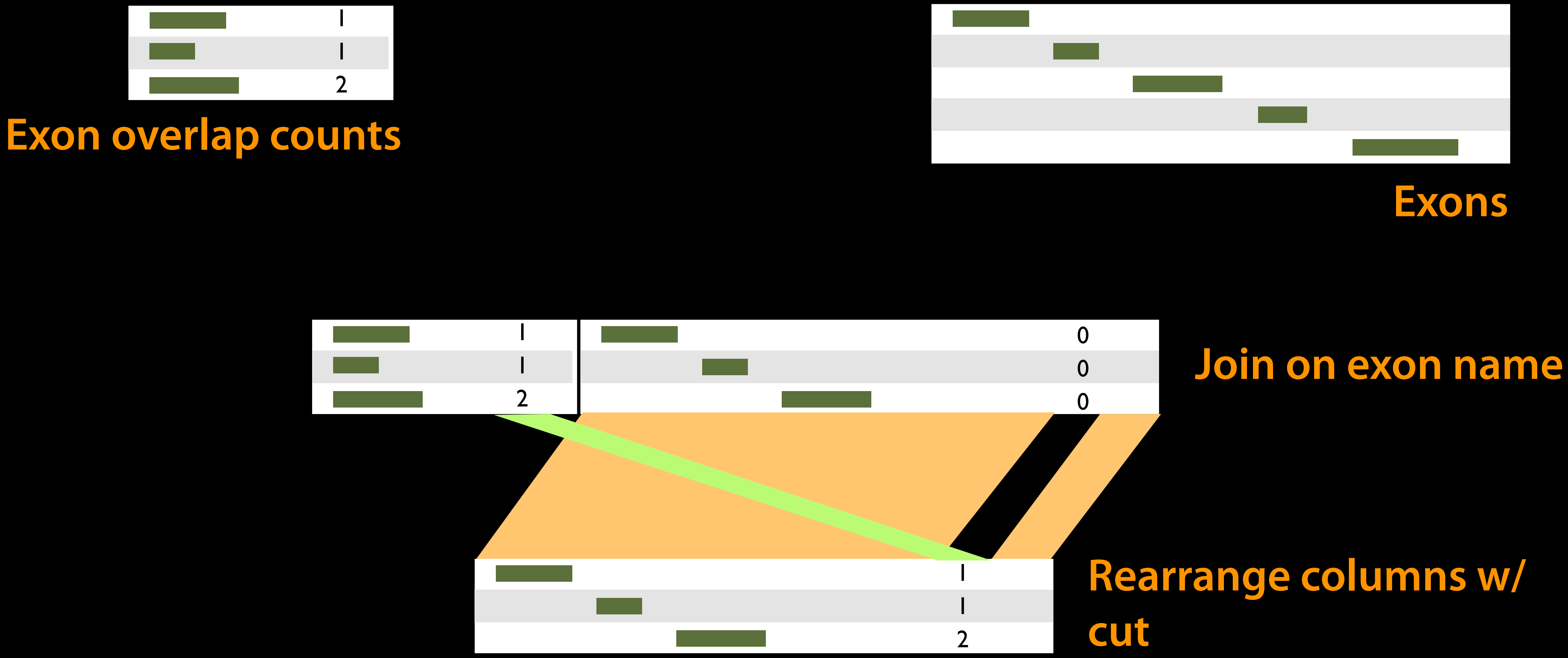
Exons



Join on exon name

Join, Subtract, and Group → Join

(Incorporate the overlap count with rest of Exon information)



Exon overlap counts

Exons

Join on exon name

Rearrange columns w/ cut

Text Manipulation → Cut

(Incorporate the overlap count with rest of Exon information)

Basic Analysis: Further reading & Resources

http://usegalaxy.org/galaxy101

https://vimeo.com/76343659

Agenda

Introduction to Galaxy

Hands-on Analysis

Community Resources

Galaxy on the cloud

Done

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Dorine Francheteau

Jeremy Goecks

Sam Guerler

Jen Jackson

Greg von Kuster

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

http://wiki.galaxyproject.org/GalaxyTeam

Galaxy is hiring post-docs and software engineers

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Thanks

Dave Clements
Galaxy Project
Johns Hopkins University
clements@galaxyproject.org

Agenda

- Hands-on Analysis
    - Differential Expression Analysis with CuffDiff

Cuffdiff

- Identifies differential expression between multiple datasets
- Uses RPKM/FPKM as its guiding statistic
- RKPM/FKPM attempts to track expression levels of each feature relative to total expression in the dataset

NGS: RNA Analysis → Cuffdiff

Cuffdiff

- Running with 2 Groups: MeOH and R3G
- Each group has 3 replicates each

Cuffdiff

- Which Transcript definitions to use?
  - Official
  - MeOH or R3G Cufflinks transcripts
  - Results of Cuffmerge on MeOH & R3G Cufflinks transcripts
- Depends on what you care about

Cuffdiff

- Produces 15 output files, all explained in doc
- We’ll focus on gene differential expression testing files (also care about gene FPKM files)
- Column 7 (“status”) can be FAIL, NOTEST, LOWDATA or OK
  - Filter and Sort → Filter
    - c7 == ‘OK’
    - Column 14 (“significant”) can be yes or no
  - c14 == ‘yes’

Agenda

Hands-on Analysis

- CuffDiff Alternatives

Alternatives

- We used Tophat (calling Bowtie) to map RNA-Seq reads to the genome
- We used Cuffdiff to identify differentially expressed genes across two experimental conditions
- Tophat, Bowtie and Cuffdiff are widely installed on many Galaxy instances, including CloudMan based instances
- but ...

Alternatives

Lindner R, Friedel CC (2012) "A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq."
PLoS ONE 7(12): e52403. doi:10.1371/journal.pone.0052403

reviews 14 packages (for slightly different problem of transciptome alignment)

Rapaport, *et al.*, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data."
Genome Biology 2013, 14:R95 doi:10.1186/gb-2013-14-9-r95

reviews 7 packages

Each tool has it's own strengths and weaknesses.

What's a biologist to do?

Alternatives: What’s a biologist to do?

Learn the strengths and weaknesses of the tools you have ready access to. Are they a good match for the questions you are asking?

If not, then research alternatives, identify good options and then work with your bioinformatics/systems people to get access to those tools.*

* You can also install alternatives in Galaxy.

Cuffdiff Alternatives: DESeq

Cuffdiff

Uses FPKM/RPKM as a central statistic.
Total # mapped reads heavily influences FPKM/RPKM.
Can lead to challenges when you have very highly expressed genes in the mix.

DESeq (and edgeR)

DESeq is an R based differential expression analysis package where expression analysis is much more effectively isolated between features.

Cuffdiff Alternatives: DESeq

Takes a simple, tab delimited list of features and read counts across different samples.
First, have to create that list.

htseq-count
Is a tool that walks BAM files producing these lists

Where are DESeq and htseq-count?

Tool Shed.

Cuffdiff Alternatives: Further Reading & Resources

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data
by Rapaport, *et al*.

DESeq Reference Manual

DESeq Galaxy Wrapper
by Nikhil Joshi

htseq-count Galaxy Wrapper
by Lance Parsons

Genes & Repeats: Exercise

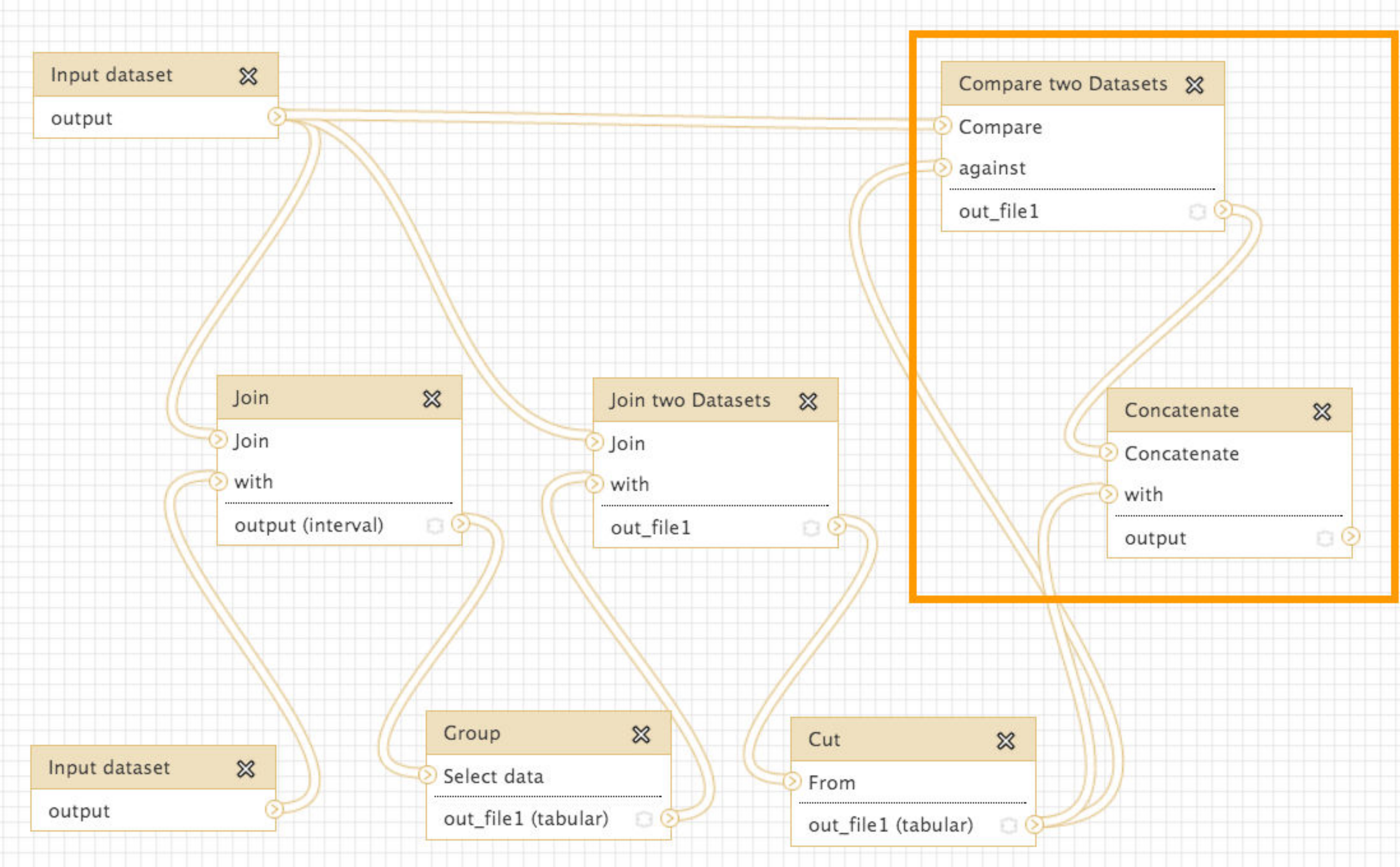
Include genes/exons with no overlaps in final output. Set the score for these to 0.

Everything you need will be in the toolboxes we used in the first Gene/Exon-Repeats exercise.

http://cloud2.galaxyproject.org/

http://cloud3.galaxyproject.org/

One Possible Solution



Solution from Stanford Kwenda and Caron Griffiths in Pretoria.

Takes advantage of the fact that Exons already have 0 scores.

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

A series of inputs, analysis steps, intermediate datasets, and outputs

Workflow:

A series of analysis steps

Can be repeated with different data

Transcript Prediction: Cufflinks

- Run Cufflinks on Tophat output to assemble reads into transcripts
  - Tophat does not make any predictions about how the reads it mapped assemble together into transcripts.
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here*

NGS: RNA Analysis → Cufflinks

Cufflinks: Min Isoform Fraction

Cufflinks can predict many different transcripts for a gene.

One transcript is likely to dominate.

Min Isoform Fraction tells Cufflinks to ignore any isoforms that fall below this level of expression, *relative to the dominant isoform.*

Higher values: less noise; less likely to report/discover low-expression transcripts.

Cufflinks: Pre mRNA Fraction

From the Cufflinks Manual

> “Some RNA-Seq protocols produce a significant amount of reads that originate from incompletely spliced transcripts, and these reads can confound the assembly of fully spliced mRNAs. Cufflinks uses this parameter to filter out alignments that lie within the intronic intervals implied by the spliced alignments. The minimum depth of coverage in the intronic region covered by the alignment is divided by the number of spliced reads, and if the result is lower than this parameter value, the intronic alignments are ignored. The default is 15%.”

Basically, sets your tolerance for noise / novel constructs in intronic regions.

Cufflinks: Normalization and Correction

How hard should Cufflinks work to do the right thing?

Quartile Optimization: Attempt to compensate for skew caused by highly expressed genes

Bias Correction: Attempt to compensate for known issues with use of random hexamers in library preparation.*

Multi-Read Correct: Try to make reads that mapped to multiple locations more useful**

* see Kasper D. Hansen, Steven E. Brenner, Sandrine Dudoit, Biases in Illumina transcriptome sequencing caused by random hexamer priming Nucleic Acids Research, Volume 38, Issue 12 (2010)

** see http://cufflinks.cbcb.umd.edu/howitworks.html#hmul

Cufflinks: Reference Annotation

How biased should we be, based on what we already know?

Reference Annotation: Use the reference annotation as dogma. Only doing quantification of known transcripts

Reference Annotation as Guide: Take advantage of what we already know, but be open to novel transcripts, if there is sufficient evidence

No: Transcript prediction will be based entirely on mapped reads in this dataset.

Transcript Prediction: Cuffmerge

- Each Cufflinks run creates a set of transcript predictions.
- Cuffmerge unifies all those predictions into a single set.
- Makes this incredibly tedious task easy.

Transcript Prediction: Cufflinks

- Run Cufflinks on Tophat output to assemble reads into transcripts
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here.*
  - Visualize *and refine* our analysis

Visualizing Genomics

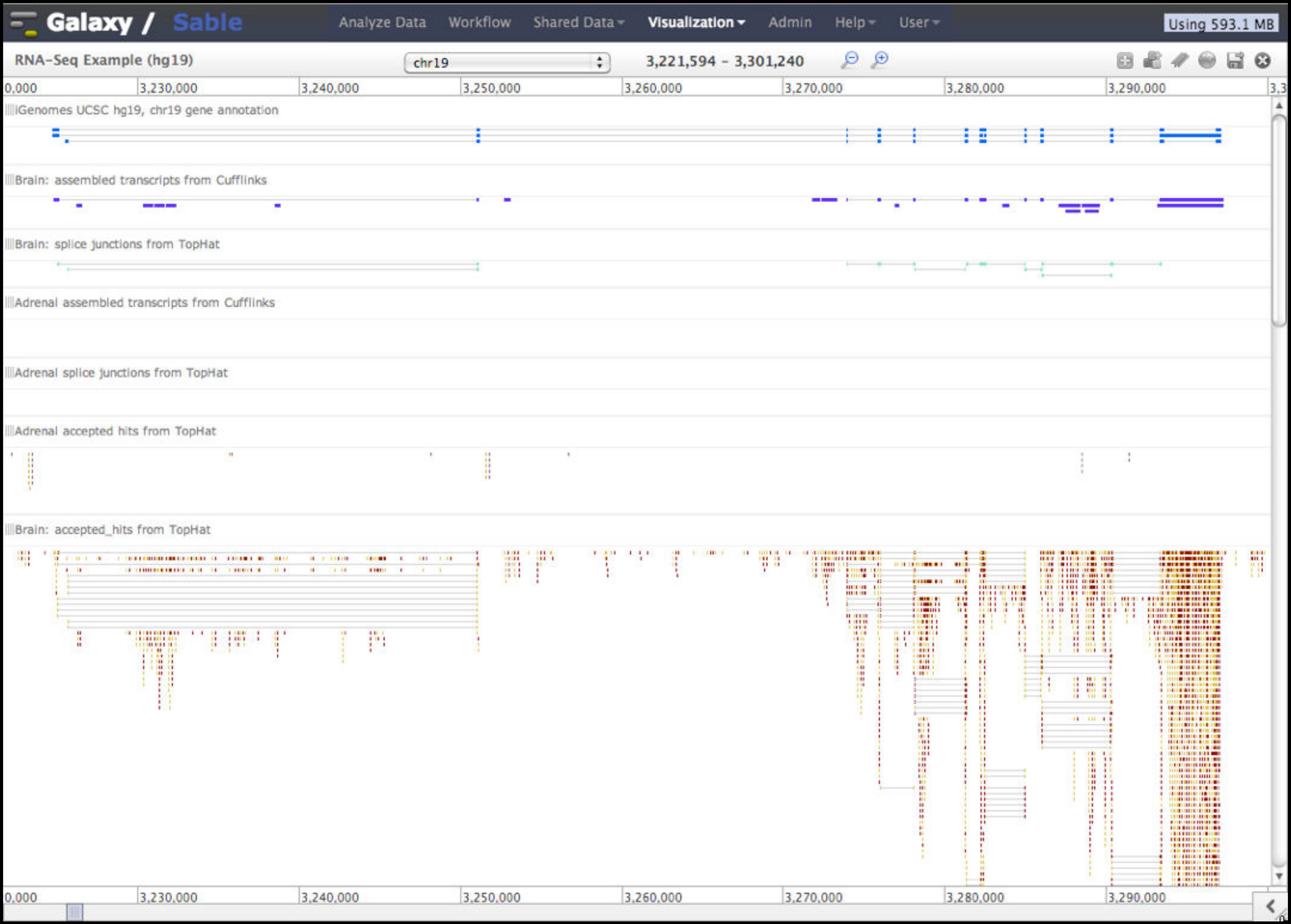
Supported external browsers

- UCSC
- Ensembl
- GBrowse
- IGB
- IGV

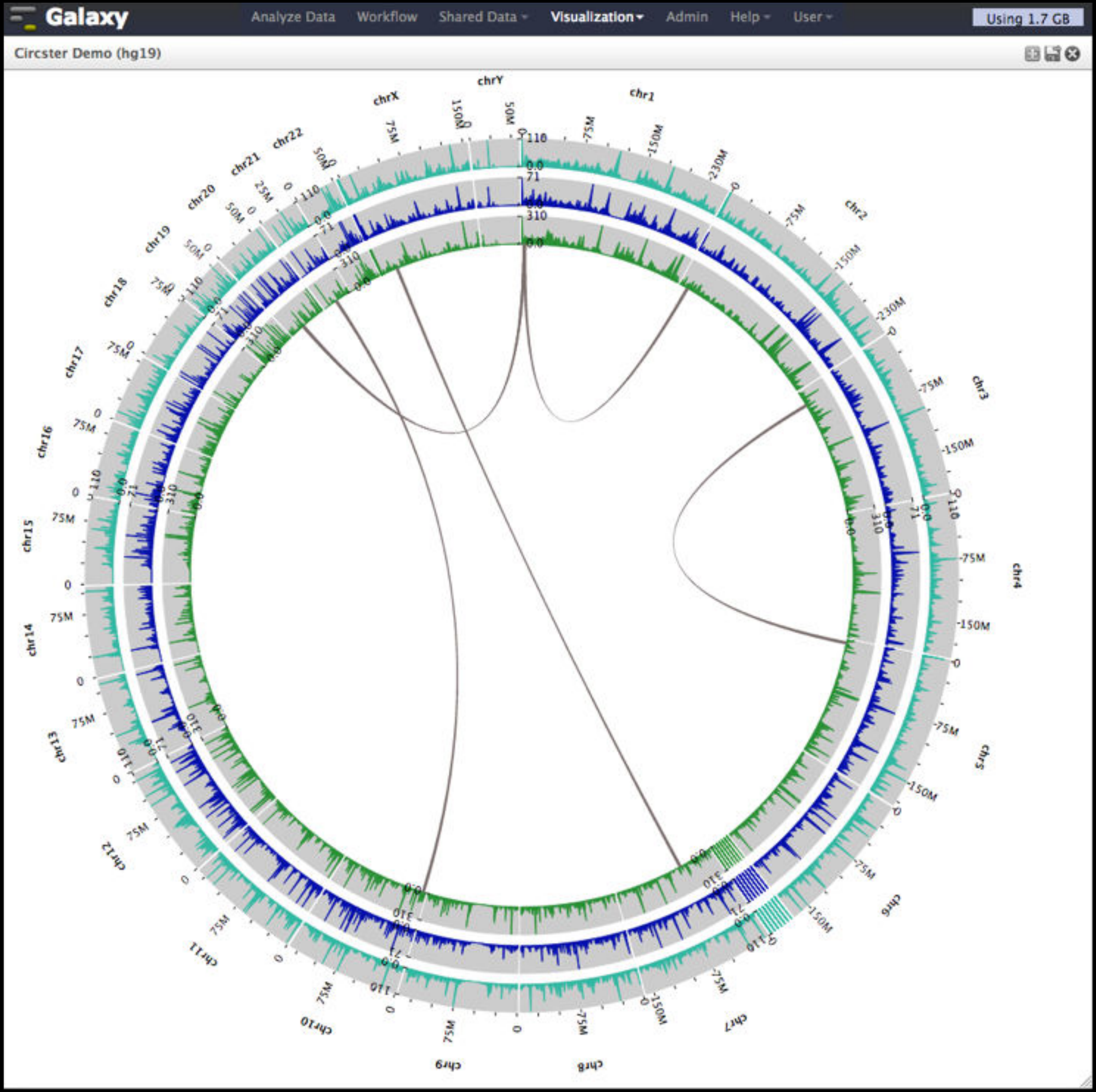
Traditional browser strengths:

- Showing what is nearby
- what else is happening here
- highlighting correlations
- integrating many datasets

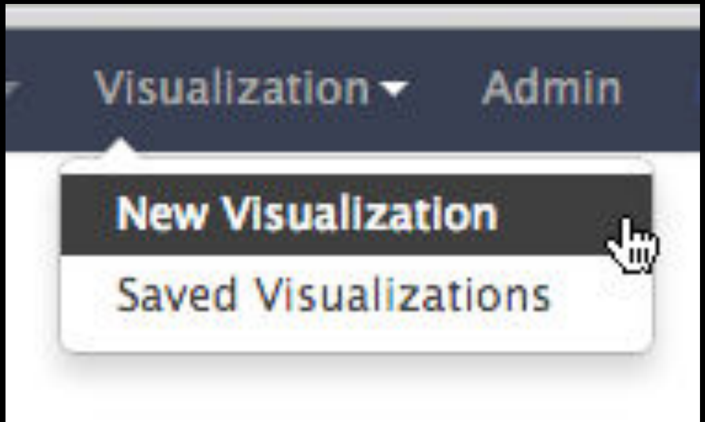
Trackster: Galaxy’s embedded track browser


Galaxy / Sable
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 593.1 MB
RNA-Seq Example (hg19)
chr19
3,221,594 - 3,301,240
0,000
3,230,000
3,240,000
3,250,000
3,260,000
3,270,000
3,280,000
3,290,000
3,3
iGenomes UCSC hg19, chr19 gene annotation
Brain: assembled transcripts from Cufflinks
Brain: splice junctions from TopHat
Adrenal assembled transcripts from Cufflinks
Adrenal splice junctions from TopHat
Adrenal accepted hits from TopHat
Brain: accepted_hits from TopHat

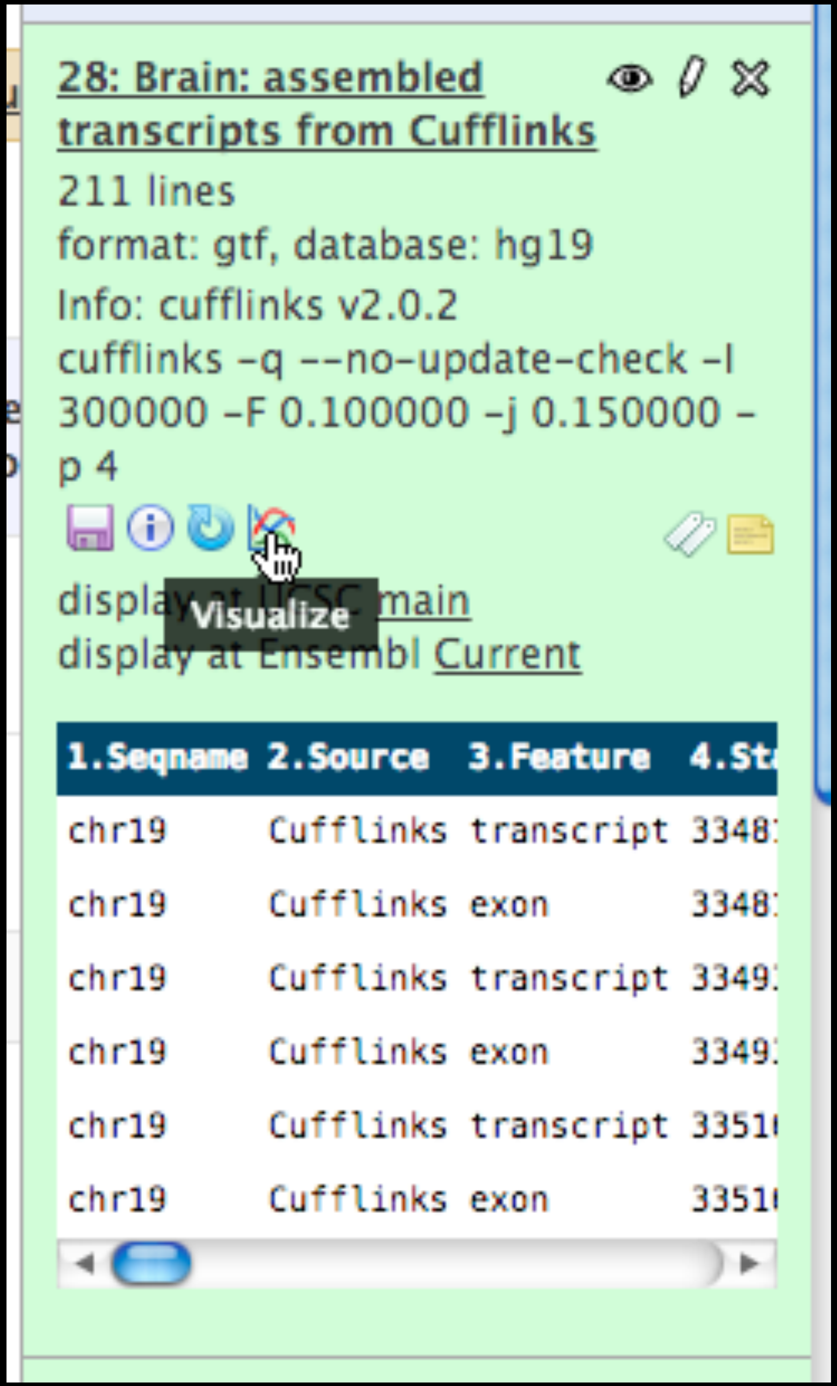
Circster


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 1.7 GB
Circster Demo (hg19)
chrY
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
75M
150M
230M
110
71
310
0.0

Create a visualization in Galaxy



or


28: Brain: assembled transcripts from Cufflinks

211 lines

format: gtf, database: hg19

Info: cufflinks v2.0.2

cufflinks -q --no-update-check -I 300000 -F 0.100000 -j 0.150000 -p 4

display at UCSC main

display at Ensembl Current

1.Seqname	2.Source	3.Feature	4.St
chr19	Cufflinks	transcript	3348
chr19	Cufflinks	exon	3348
chr19	Cufflinks	transcript	3349
chr19	Cufflinks	exon	3349
chr19	Cufflinks	transcript	3351
chr19	Cufflinks	exon	3351

Vizualization inside Galaxy

- Levarge visualization to evaluate and refine analyses
- Make the *analyze-visualize-refine* loop seamless and fast
- Enable experimenting with tools and their parameter space
- Support custom genome browsers

Transcript Prediction: Further Reading & Resources

Princeton HTSEQ Users RNA-Seq Tutorial
by Lance Parsons

Gene Construction
By Monica Britton

Web-based visual analysis for high-throughput genomics
by Goecks, et al.

Cufflinks Manual