Introduction to Galaxy

National Institute of
Environmental Health Sciences
Research Triangle Park, NC
July 18, 2013

Dave Clements, Emory University
http://galaxyproject.org/

Galaxy

amazon web services™

Research Triangle 2013 Galaxy Workshop Tour

NIH National Institute of Environmental Health Sciences

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

Introductions

In 46 seconds or less tell us

- your name
- your affiliation(s)
- something about your research
- something about your goals for today

Goals

1. Introduce Galaxy
2. Introduce bioinformatics concepts and formats
3. Hands-on experience
    - Load and integrate data
    - Perform bioinformatic analysis with Galaxy
    - Save, share describe and publish your analyses
    - Visualize your results
    - Learn how to set up a Galaxy server in the cloud

This workshop will not cover details of how tools are implemented, or new algorithm designs, or which assembler or mapper or ... is best for you.

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

Basic Analysis: We have

an assembly of an archaeal organism
gene annotation
TF binding sites

Which genes have most overlapping TFBs?

http://cloud1.galaxyproject.org/
http://cloud2.galaxyproject.org/
http://cloud3.galaxyproject.org/

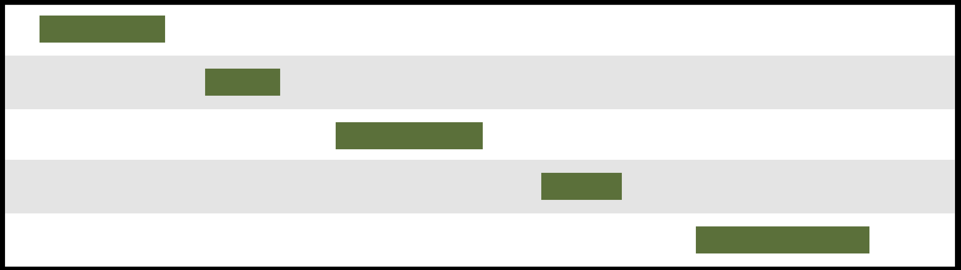
(~ http://usegalaxy.org/galaxy101)

Genes & TFBs: A General Plan

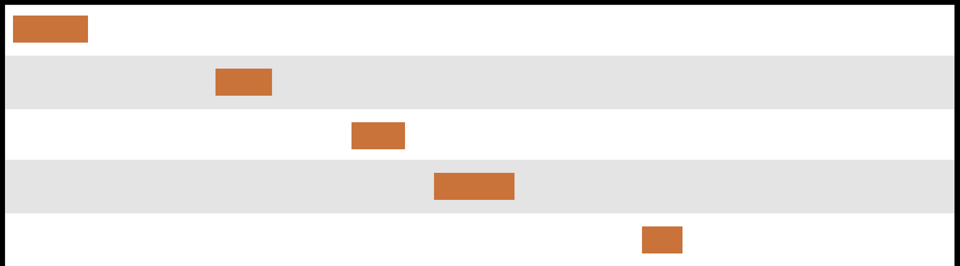
- Get some data
  - Sequence, genes/exons, TFBs
- Mess with it
  - Identify which genes/exons have TFBs
  - Count TFBs per exon
  - Visualize, save, download, ... exons with most TFBs

http://cloud1.galaxyproject.org/
http://cloud2.galaxyproject.org/
http://cloud3.galaxyproject.org/

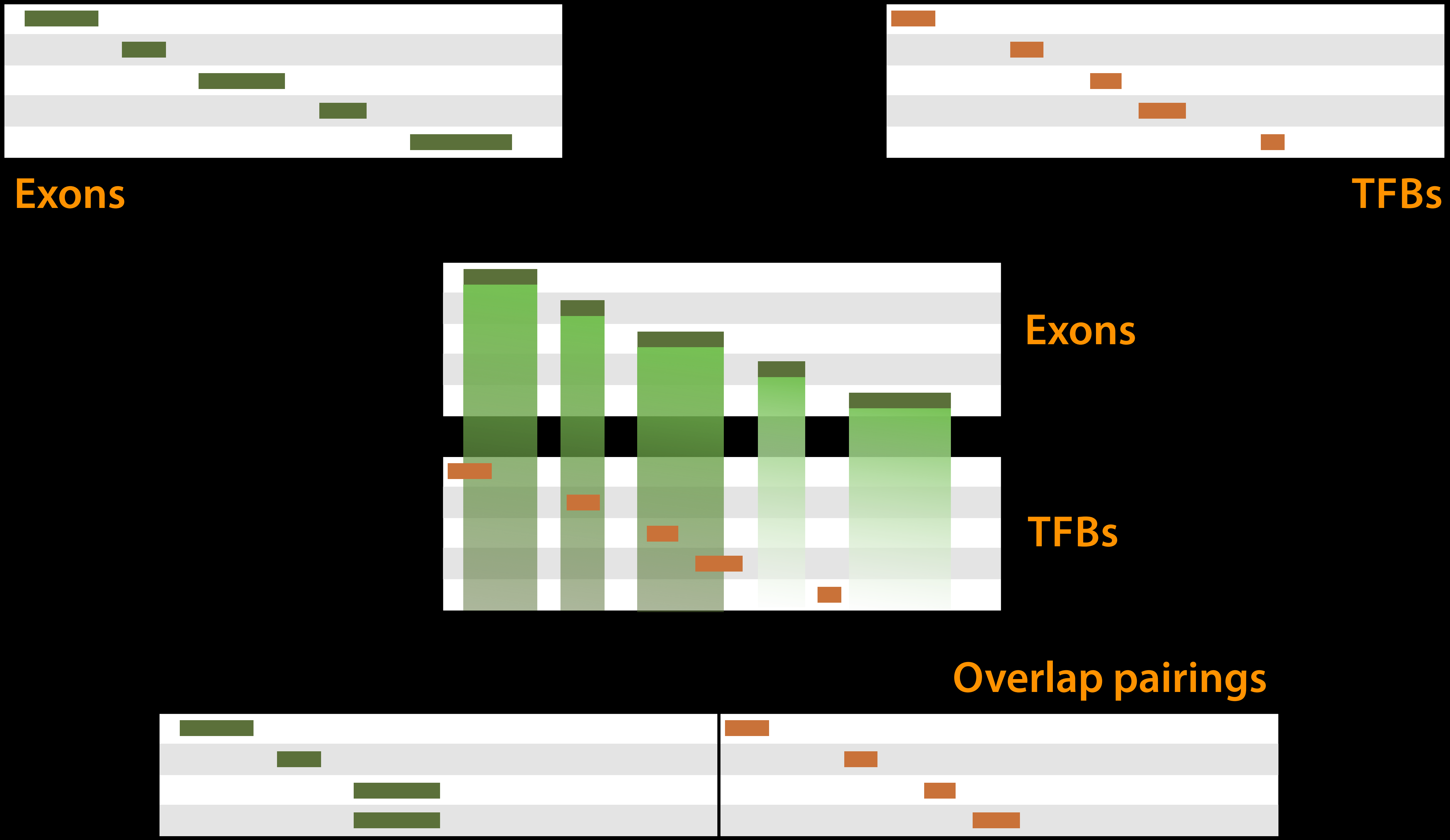
(~ http://usegalaxy.org/galaxy101)



Exons



TFBs



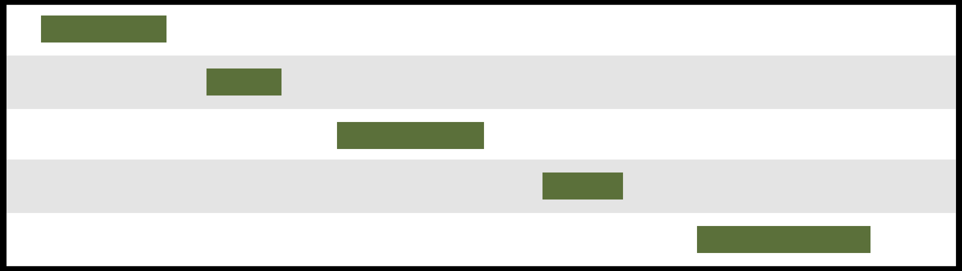
Exons

TFBs

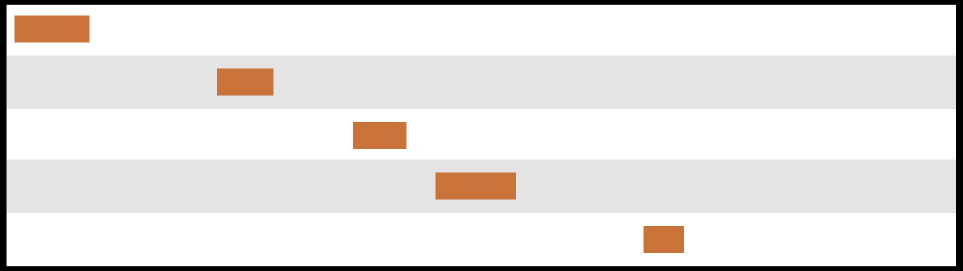
Exons

TFBs

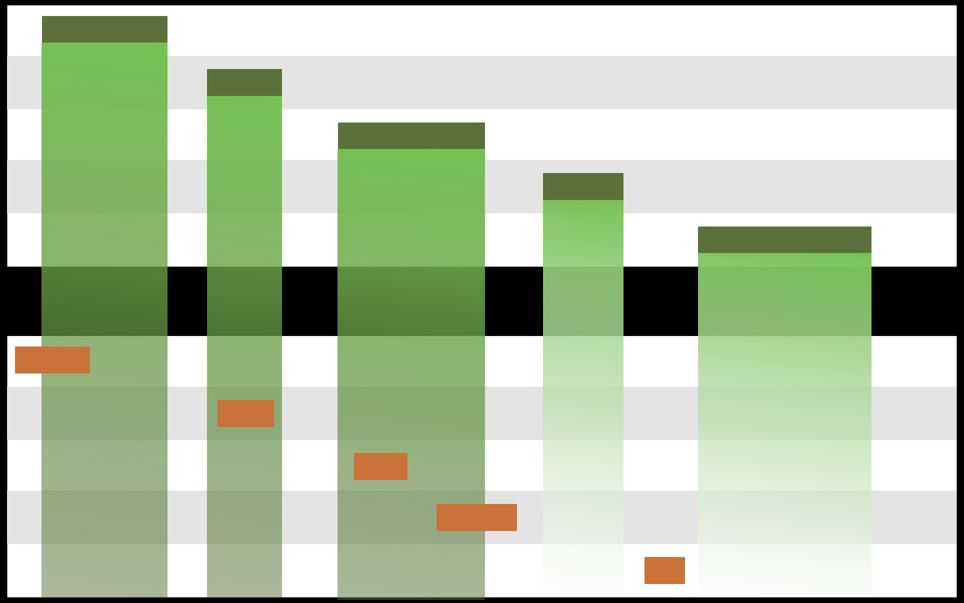
Overlap pairings



Exons



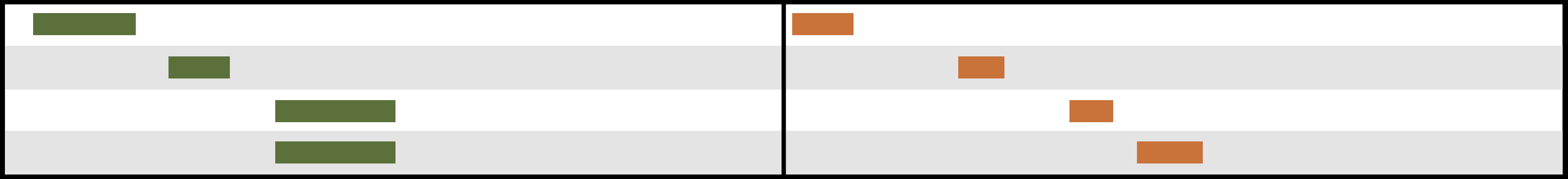
TFBS

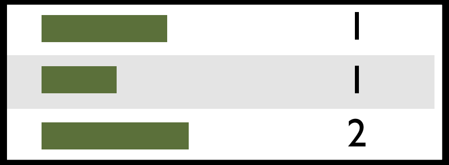


Exons

TFBS

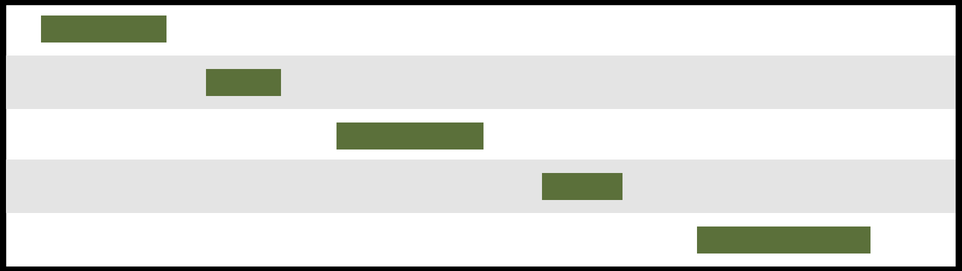
Overlap pairings



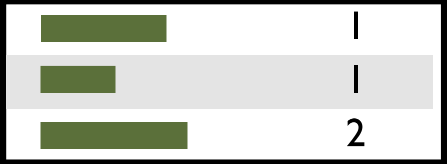


Exon overlap counts

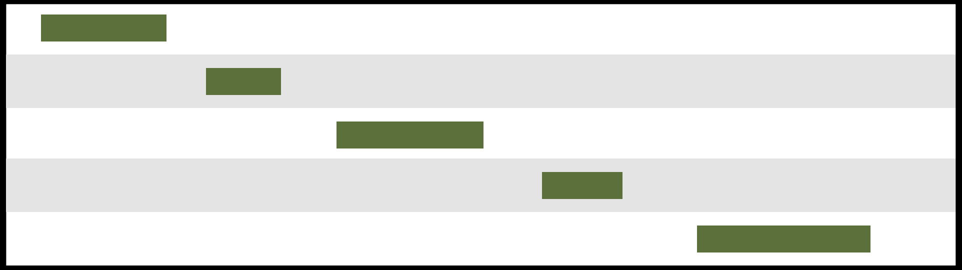
Exon overlap counts



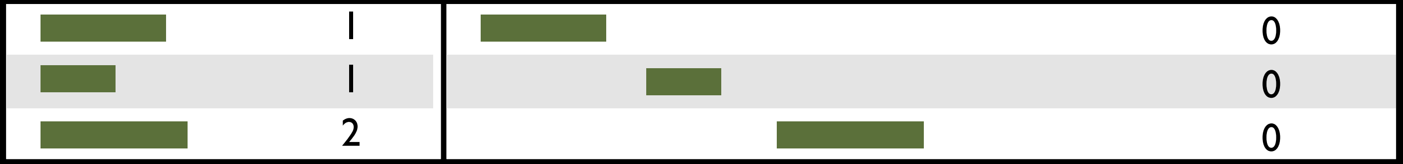
Exons



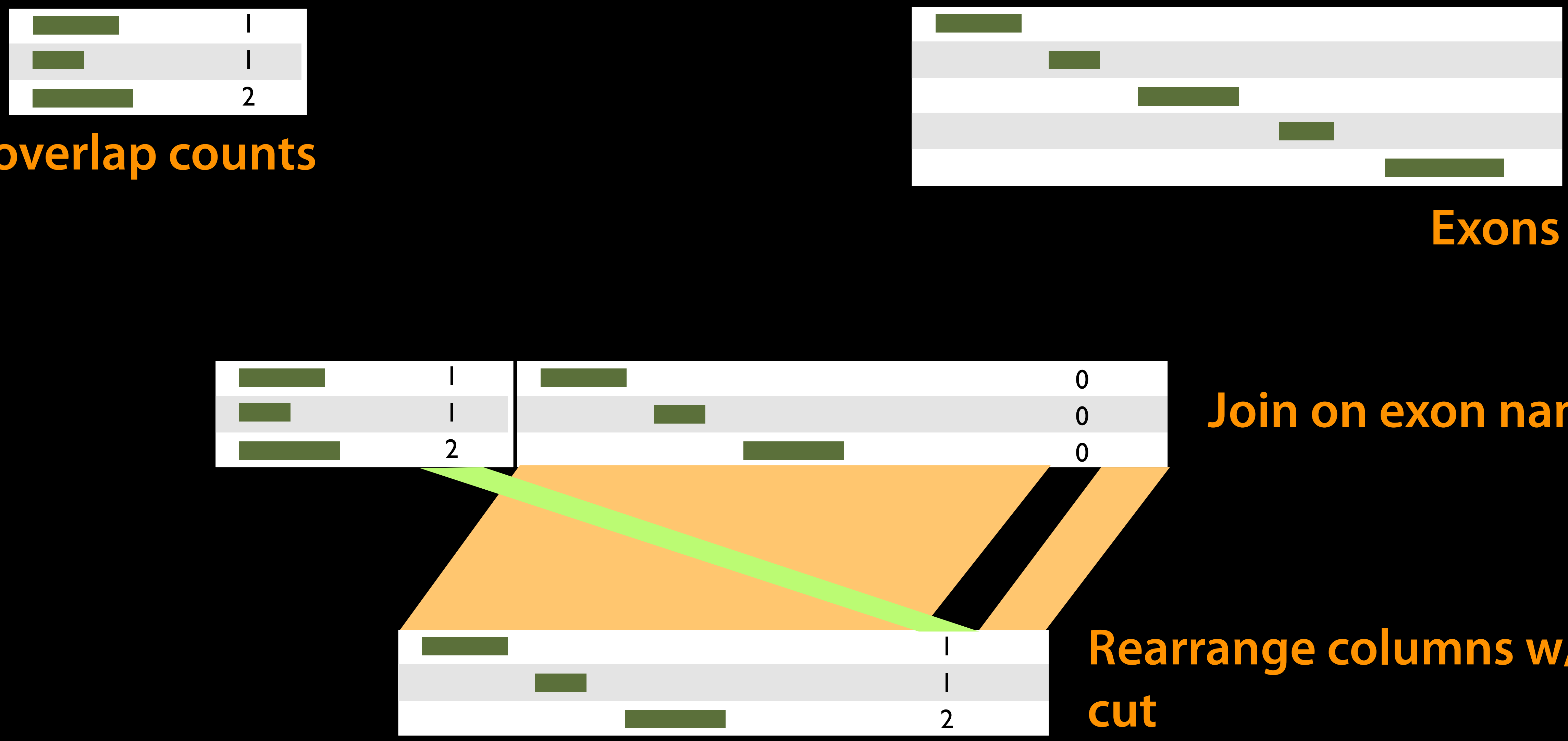
Exon overlap counts



Exons



Join on exon name



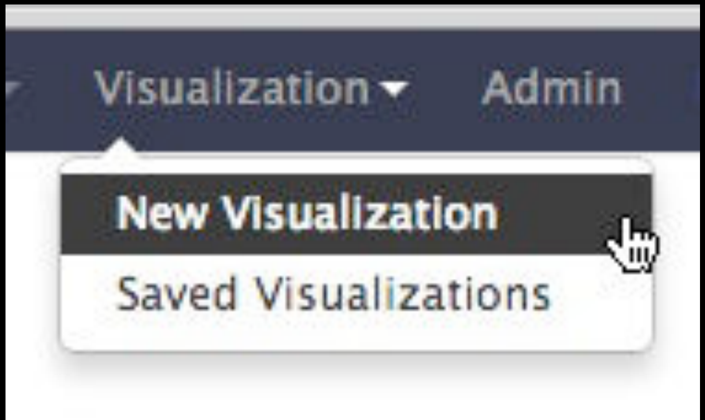
Exon overlap counts

Exons

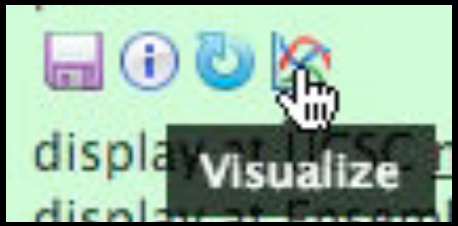
Join on exon name

Rearrange columns w/ cut

Visualize results



or



Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

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

Exons and TFBs *History* → Reusable *Workflow?*

- The analysis we just finished was about
  - An archaea
  - Overlap between exons and TFBs
- But, ...
  - there is nothing inherently in the analysis about archaea, exons or TFBs
  - It is a series of steps that sets the score of one set of features to the number of overlaps from another set of features.

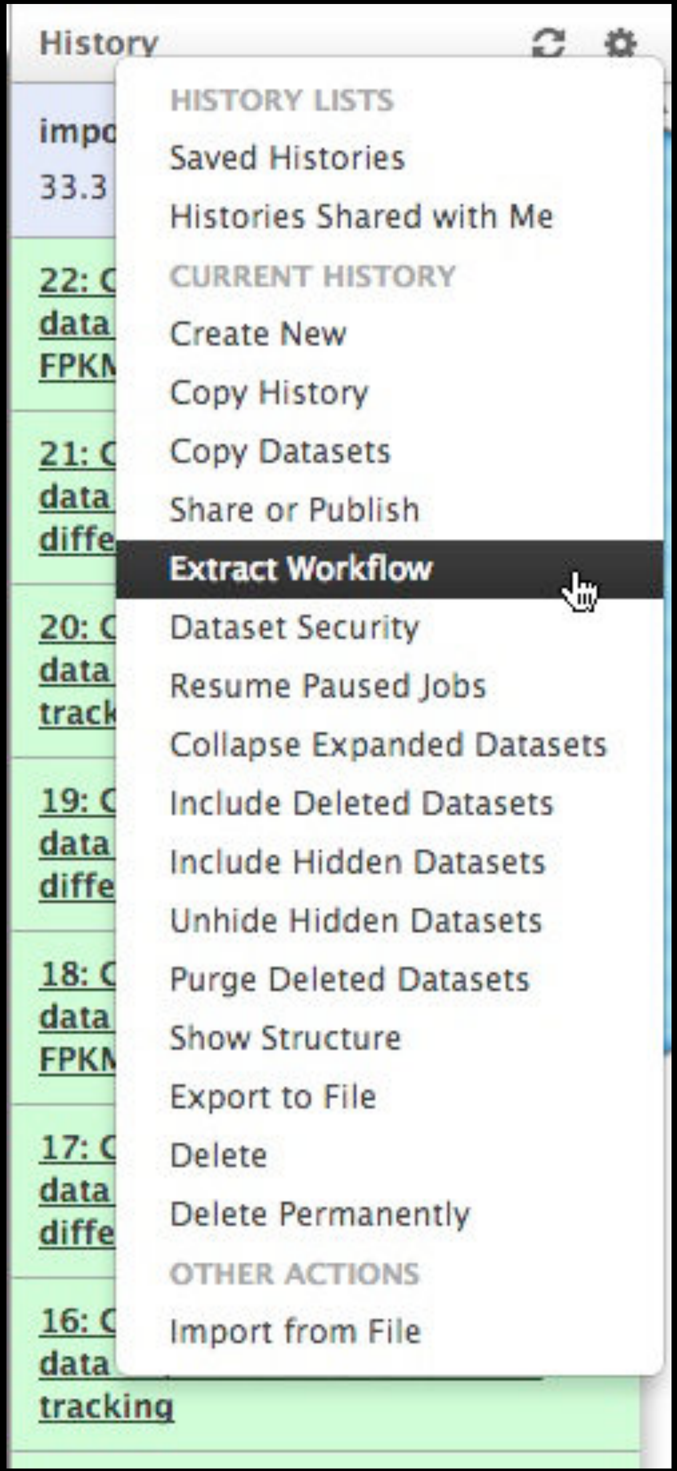
Create a generic *Overlap* Workflow

Extract Workflow from history

- Create a workflow from this history.
- Edit it to make some things clearer.

Run / test it

- Guided: rerun with same inputs
  - Did that work?
- On your own:
  - Count # of exons in each TFBS
  - Did that work? *Why not?*
  - Edit workflow: doc assumptions



Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

RNA-seq Exercise

Shared Data → Published Pages

→ RNA-Seq Analysis Exercise

RNA-seq Exercise: A Plan

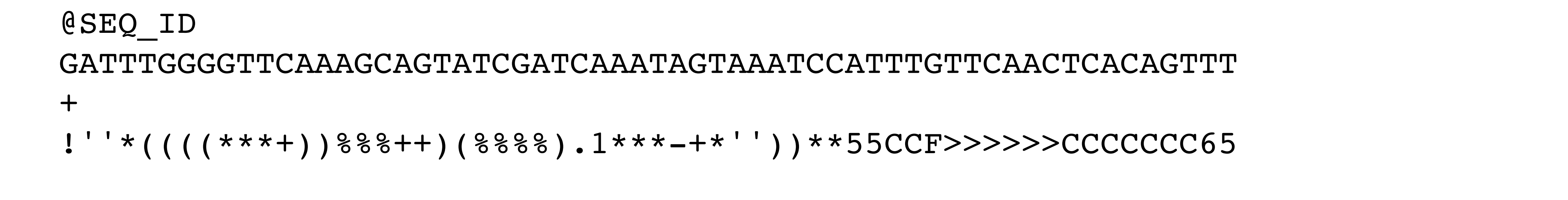
- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit.
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts
- Visualize it

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
  - All datasets are FASTQ and from the Body Map 2.0 project
    - Shared Data → Published Pages → RNA-Seq Analysis Exercise
    - or Shared Data → Data Libraries → RNA-Seq Example

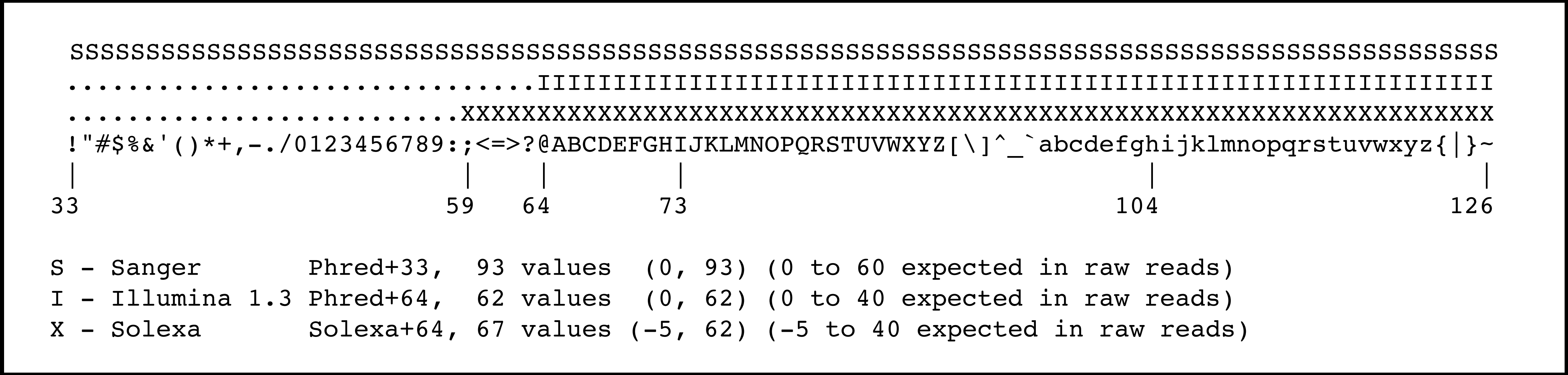
What is FASTQ?

- 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

RNA-seq Exercise: A Plan

Look at quality 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.

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

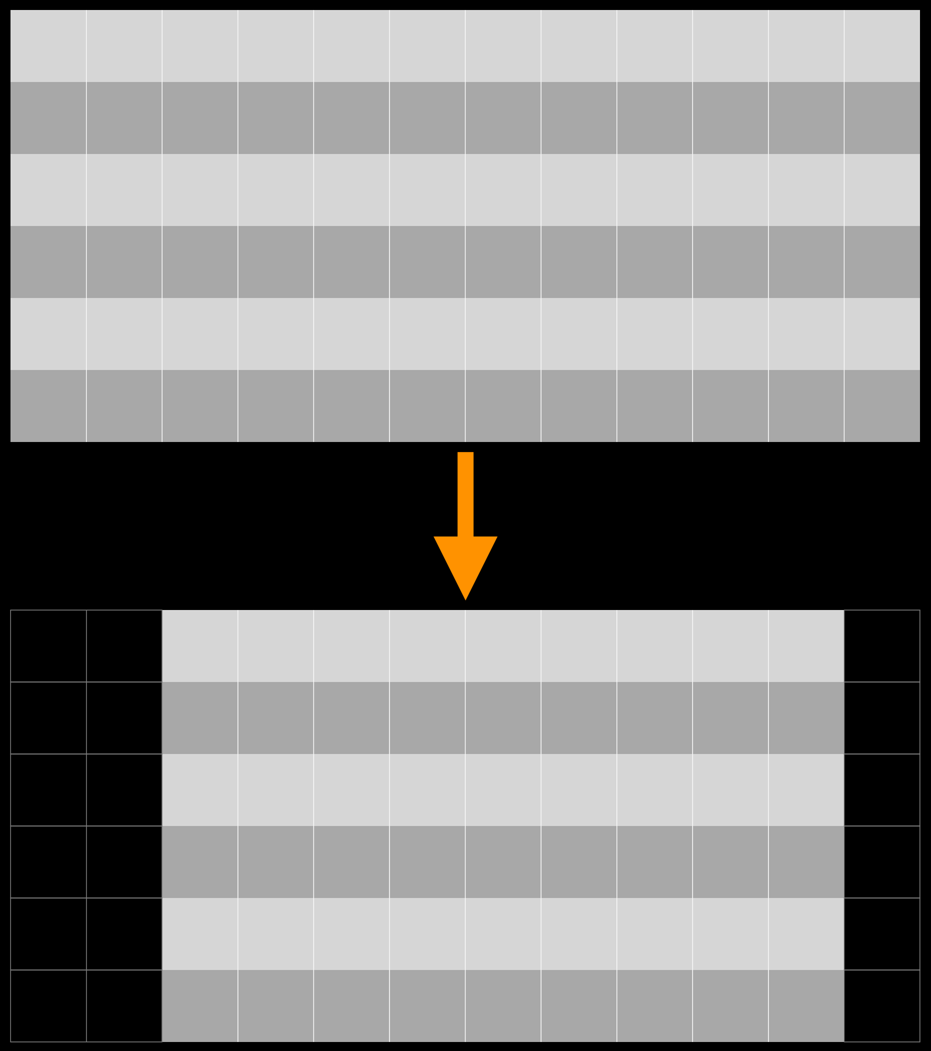
RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality: Option 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

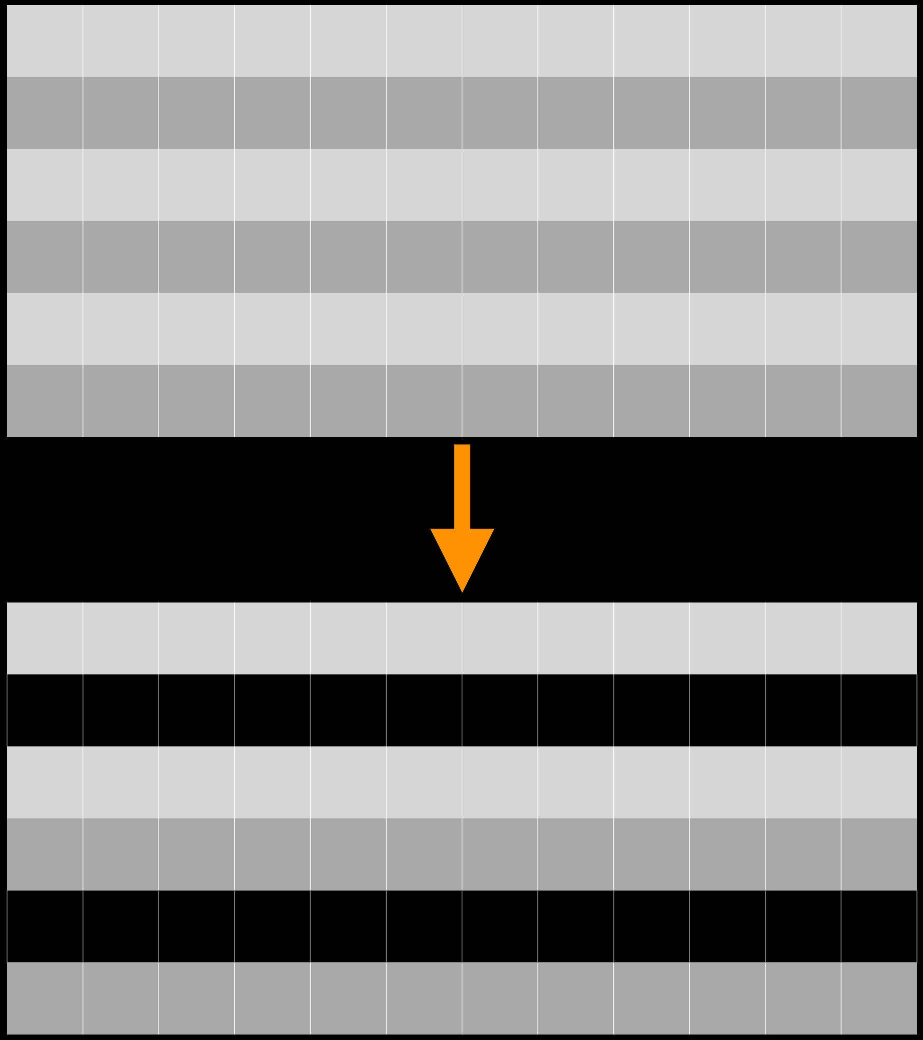
RNA-seq Exercise: A Plan

- Look at quality
- 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



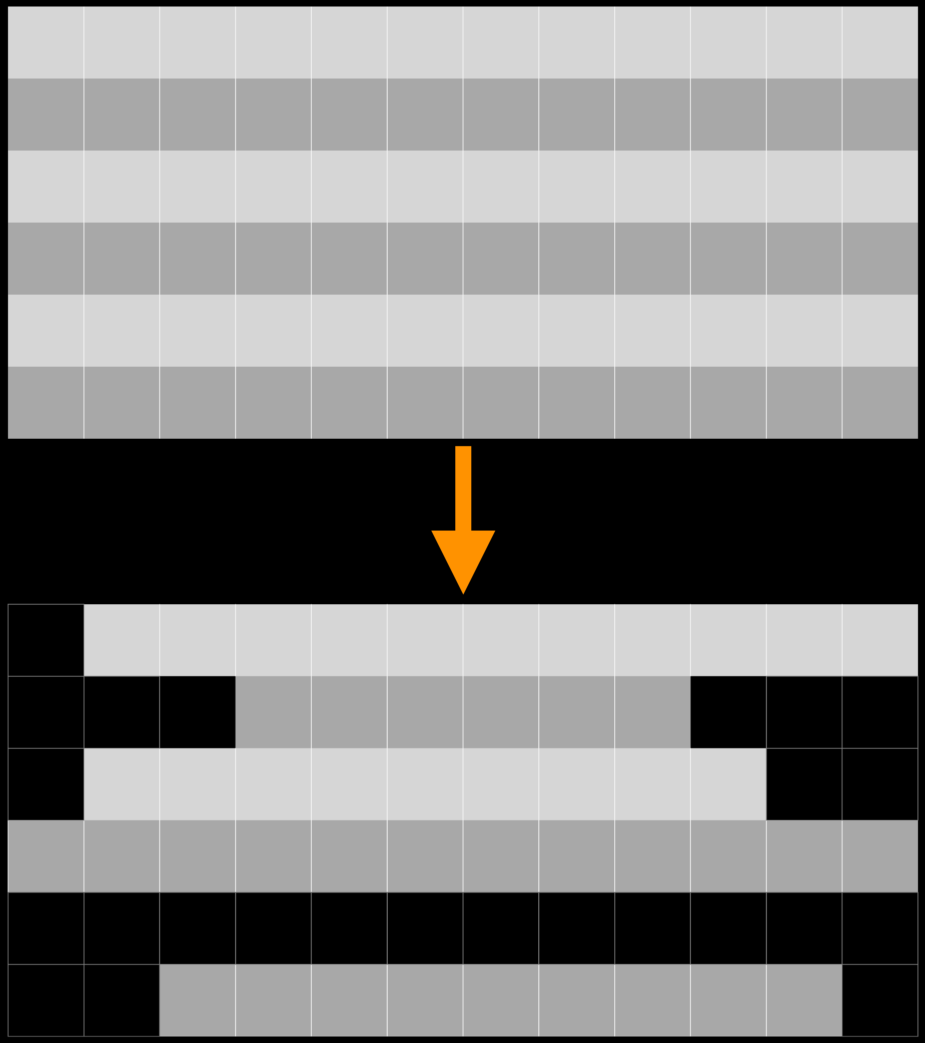
RNA-seq Exercise: A Plan

- Look at quality
- ~~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.

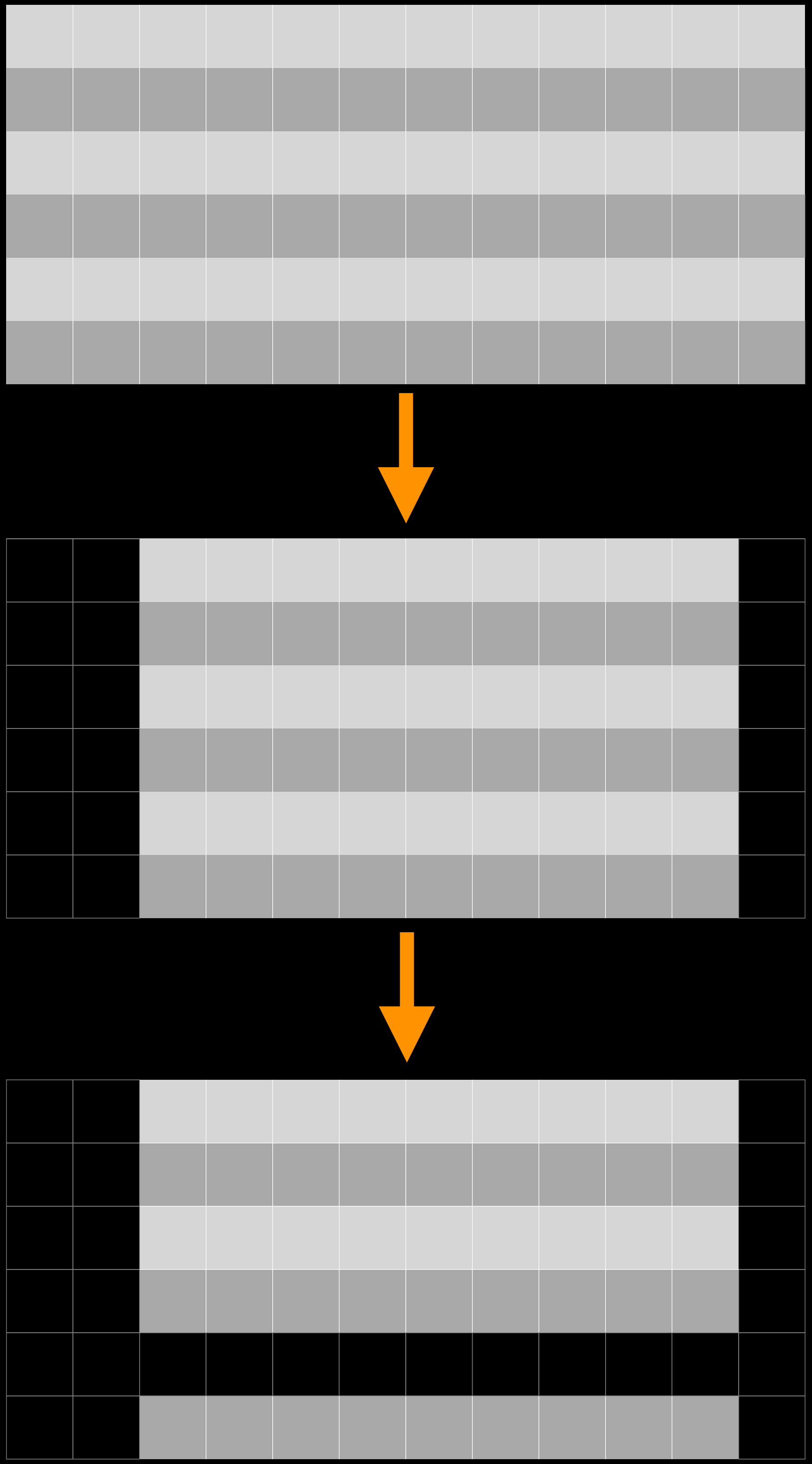


RNA-seq Exercise: A Plan

- Look at quality
- 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

+

Option 2

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?
  - http://biostars.org/
  - http://seqanswers.com/
  - http://galaxyproject.org/search

SEQanswers
the next generation sequencing community

Biostar
BIOINFORMATICS EXPLAINED

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit.
- Map the reads to the human reference using 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.*

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

What is Galaxy?

- **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
- These options result in several **ways to use Galaxy**

http://galaxyproject.org

Galaxy is available ...

As a free (for everyone) web service

http://usegalaxy.org

However, ***a centralized solution cannot scale to meet the analysis needs of the entire world.***

Galaxy is available ...

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

As Open Source Software: Local Galaxy Instances

- Galaxy is designed for local installation and customization
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters
- Requires a computational resource on which to be deployed

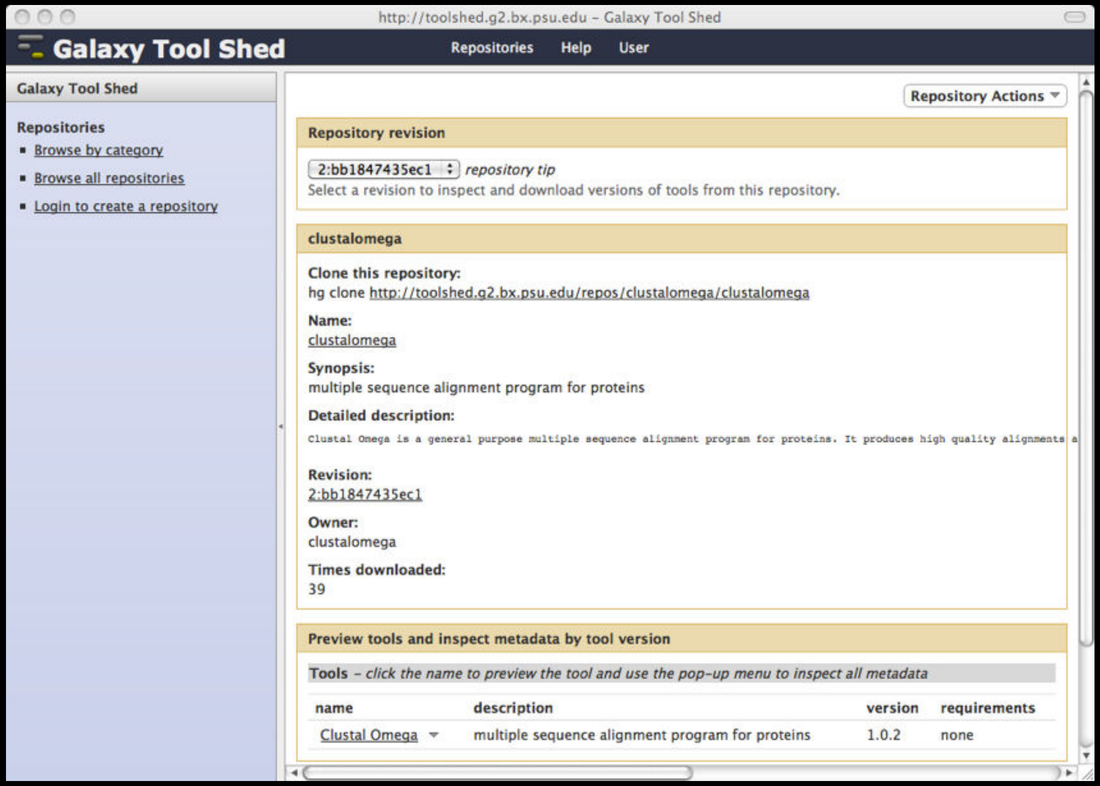
http://getgalaxy.org

Encourage Local Galaxy Instances

- Encourage and support Local Galaxy Instances
  - Support increasingly decentralized model and improve access to existing resources
  - Focus on building infrastructure to enable the community to integrate and share tools, workflows, and best practices

Galaxy Tool Shed

http://toolshed.g2.bx.psu.edu



Repository revision

2:bb1847435ec1 *repository tip*

Select a revision to inspect and download versions of tools from this repository.

clustalomega

Clone this repository:
hg clone http://toolshed.g2.bx.psu.edu/repos/clustalomega/clustalomega

Name:
clustalomega

Synopsis:
multiple sequence alignment program for proteins

Detailed description:
```
Clustal Omega is a general purpose multiple sequence alignment program for proteins. It produces high quality alignments a
```

Revision:
2:bb1847435ec1

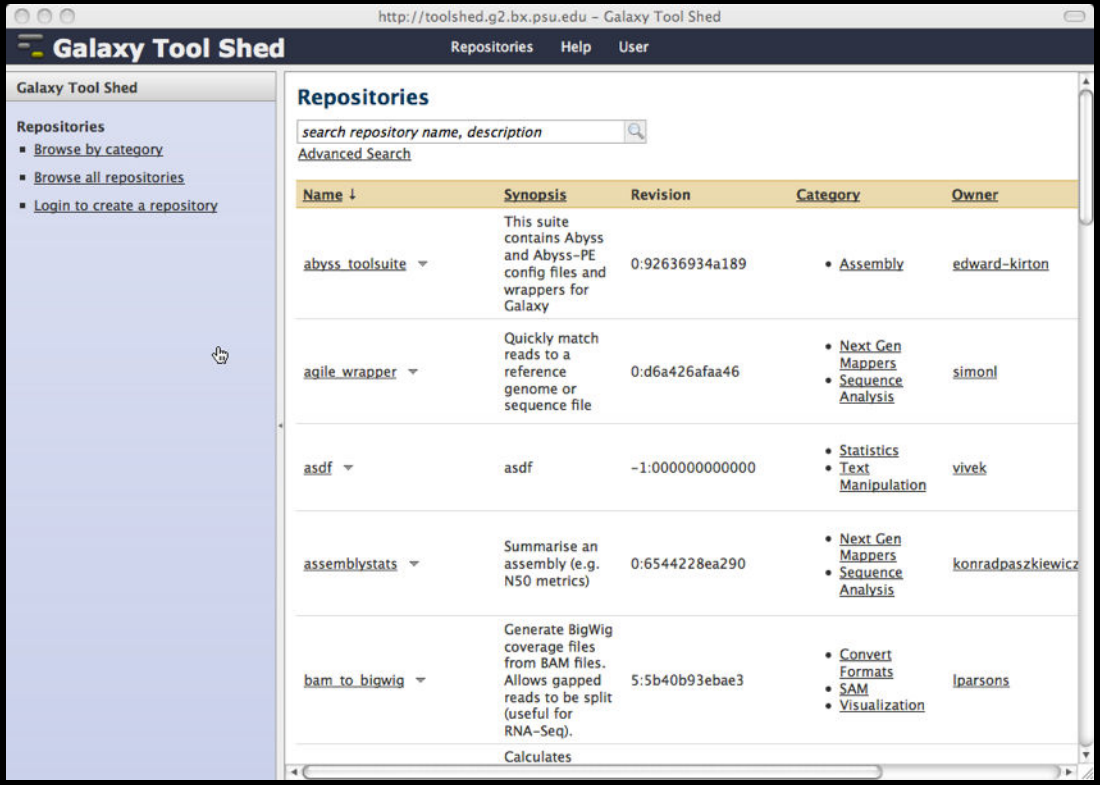
Owner:
clustalomega

Times downloaded:
39

Preview tools and inspect metadata by tool version

Tools – *click the name to preview the tool and use the pop-up menu to inspect all metadata*

name	description	version	requirements
Clustal Omega	multiple sequence alignment program for proteins	1.0.2	none



Repositories

search repository name, description

Advanced Search

Name ↓	Synopsis	Revision	Category	Owner
abyss_toolsuite	This suite contains Abyss and Abyss-PE config files and wrappers for Galaxy	0:92636934a189	• Assembly	edward-kirton
agile_wrapper	Quickly match reads to a reference genome or sequence file	0:d6a426afaa46	• Next Gen Mappers • Sequence Analysis	simonl
asdf	asdf	-1:000000000000	• Statistics • Text Manipulation	vivek
assemblystats	Summarise an assembly (e.g. N50 metrics)	0:6544228ea290	• Next Gen Mappers • Sequence Analysis	konradpaszkiewicz
bam_to_bigwig	Generate BigWig coverage files from BAM files. Allows gapped reads to be split (useful for RNA-Seq).	5:5b40b93ebae3	• Convert Formats • SAM • Visualization	lparsons
	Calculates			

Encourage **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

Plus many more

As Open Source Software: Local Galaxy Instances

- Galaxy is designed for local installation and customization
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters
- Requires a **computational resource** on which to be deployed

http://getgalaxy.org

Got your own cluster?

- Galaxy works with any DRMAA compliant cluster job scheduler (which is most of them).
- Galaxy is just another client to your scheduler.

CLUSTER RESOURCES
TORQUE

GRID ENGINE

Platform Computing

DRMAA
Distributed Resource Management Application API — www.drmaa.org

Galaxy is available ...

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

- *On the Cloud*

  http://usegalaxy.org/cloud

  We are using this right now, and we will demonstrate how to do this later today

http://aws.amazon.com/education

Galaxy is available ...

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

A ready-to-use appliance (BioTeam)

Cloud-based solutions (Appistry, ABgenomica, AIS)

Consulting & Customization (Arctix, Deena Bioinformatics)

Galaxy Resources and Community

Mailing Lists (very active)

Unified Search

Issues Board

Events Calendar, News Feed

Community Wiki

GalaxyAdmins

Screencasts

Tool Shed

Public Installs

CiteULike group, Mendeley mirror

Annual Community Meting

http://wiki.galaxyproject.org

Galaxy Resources and Community: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (42 posts in 2012, 2100+ members)

Galaxy-User

Questions about using Galaxy and usegalaxy.org

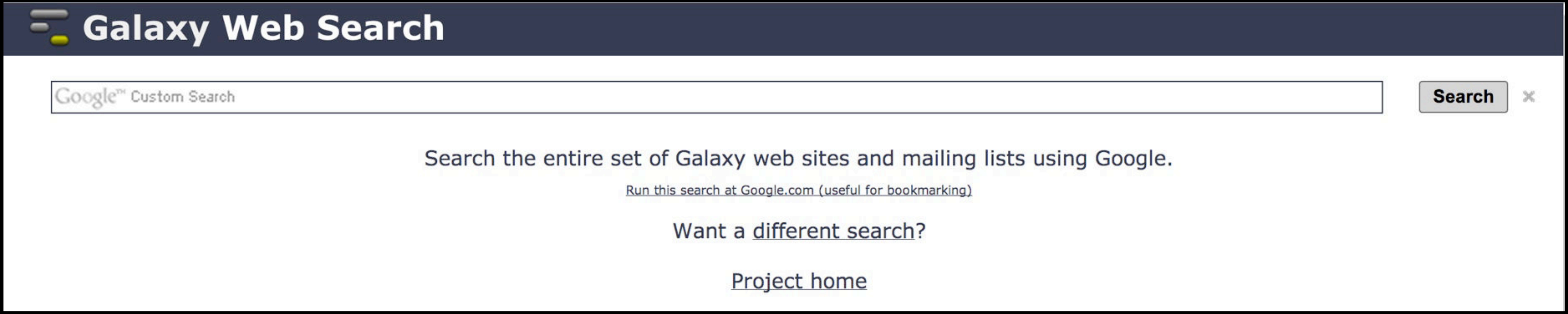
High volume (2900 posts in 2012, 2700+ members)

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (4500 posts in 2012,   900+ members)

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

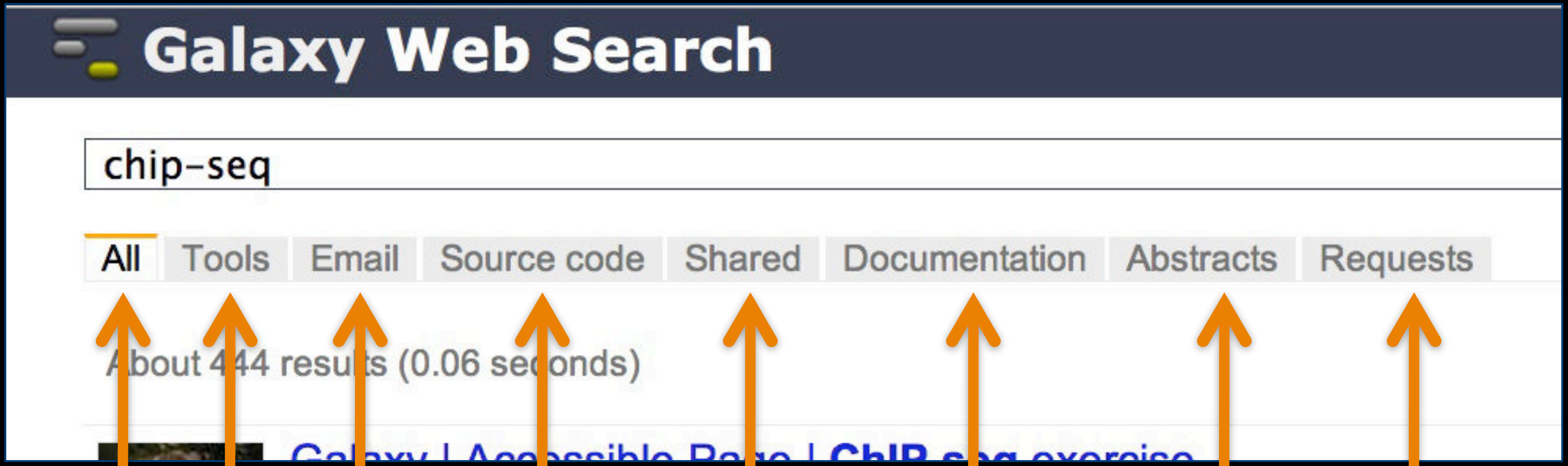


Search the entire set of Galaxy web sites and mailing lists using Google.

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

Want a different search?

Project home



About 444 results (0.06 seconds)

Galaxy | Accessible Page | ChIP-seq exercise

Find

Everything on …

Tools for …

Email about …

Source code for …

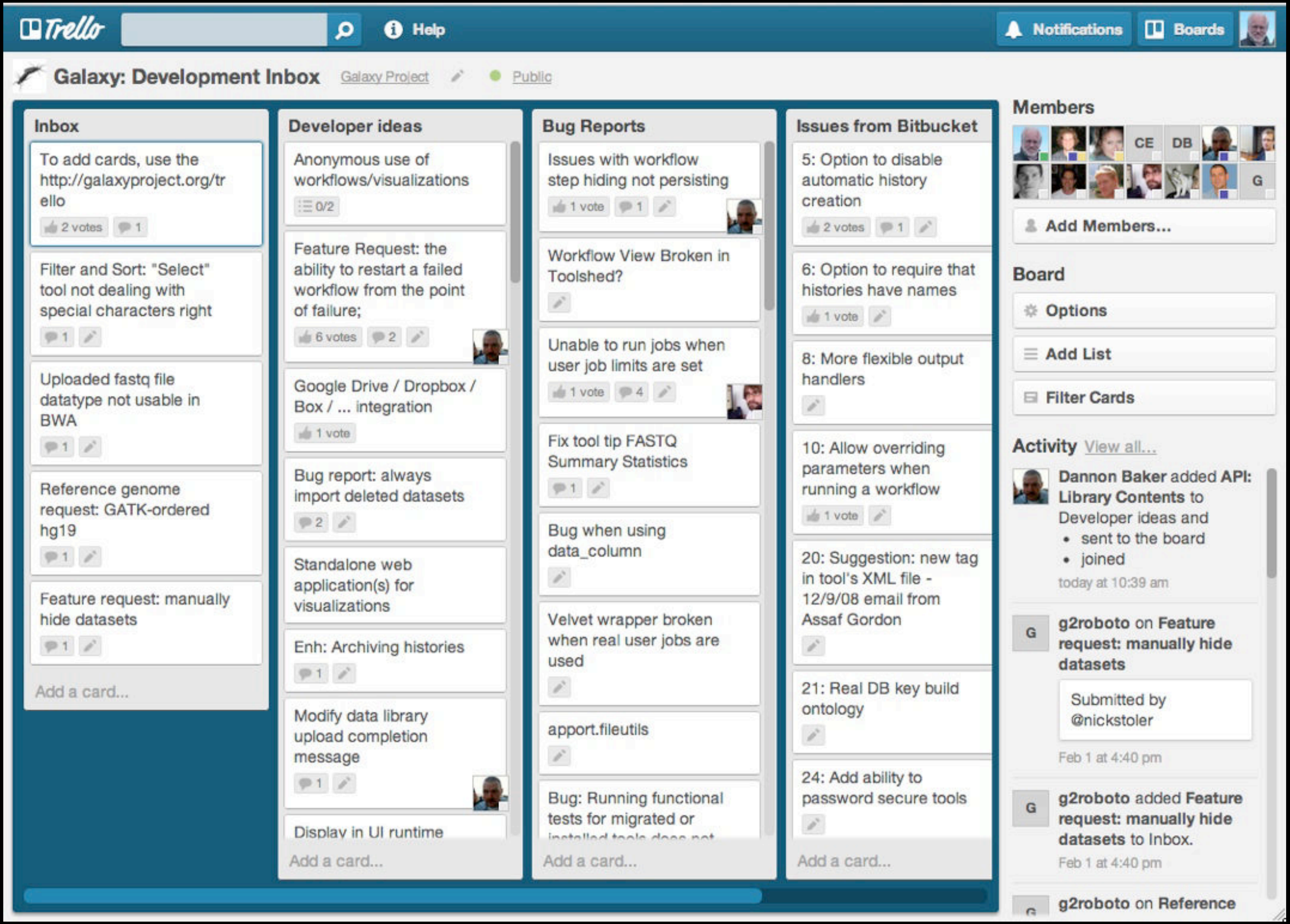
Published Histories, Pages, Workflows, about …

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

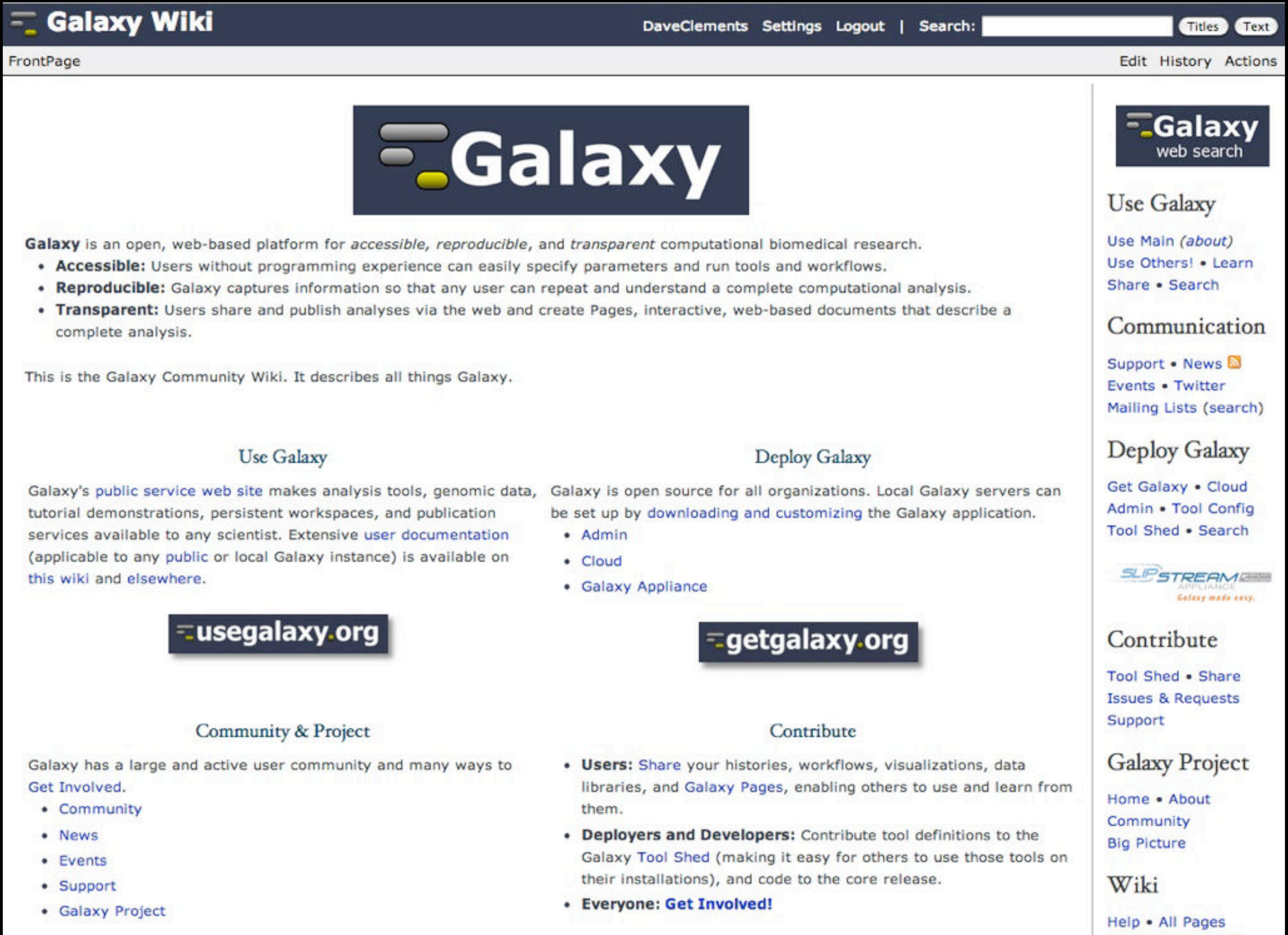
g2roboto added **Feature request: manually hide datasets** to Inbox.

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.

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

Community & Project

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

- Community
- News
- Events
- Support
- Galaxy Project

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.
- **Everyone: Get Involved!**

Use Galaxy

Use Main *(about)*
Use Others! • Learn
Share • Search

Communication

Support • News
Events • Twitter
Mailing Lists (search)

Deploy Galaxy

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

Contribute

Tool Shed • Share
Issues & Requests
Support

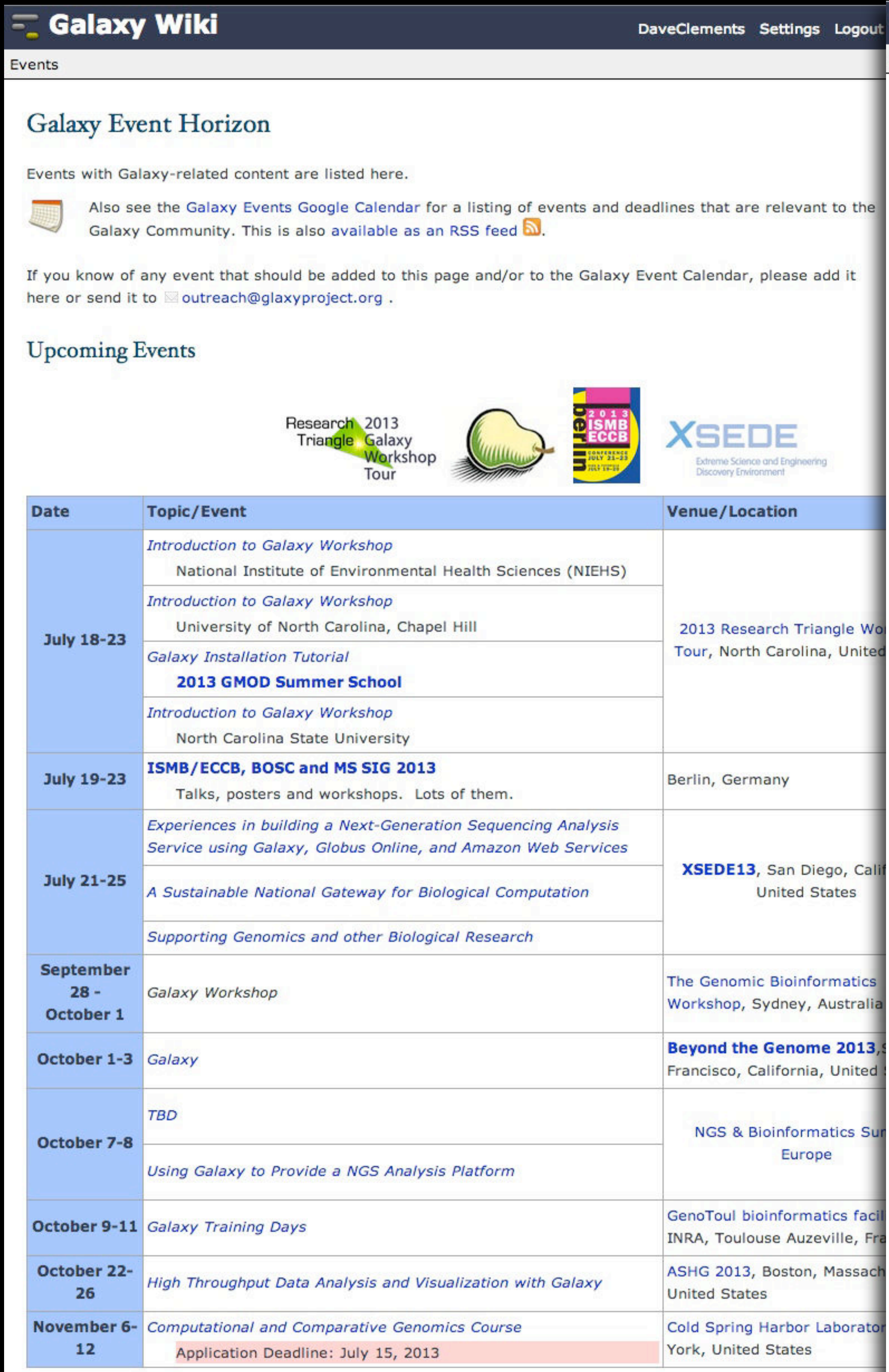
Galaxy Project

Home • About
Community
Big Picture

Wiki

Help • All Pages

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.

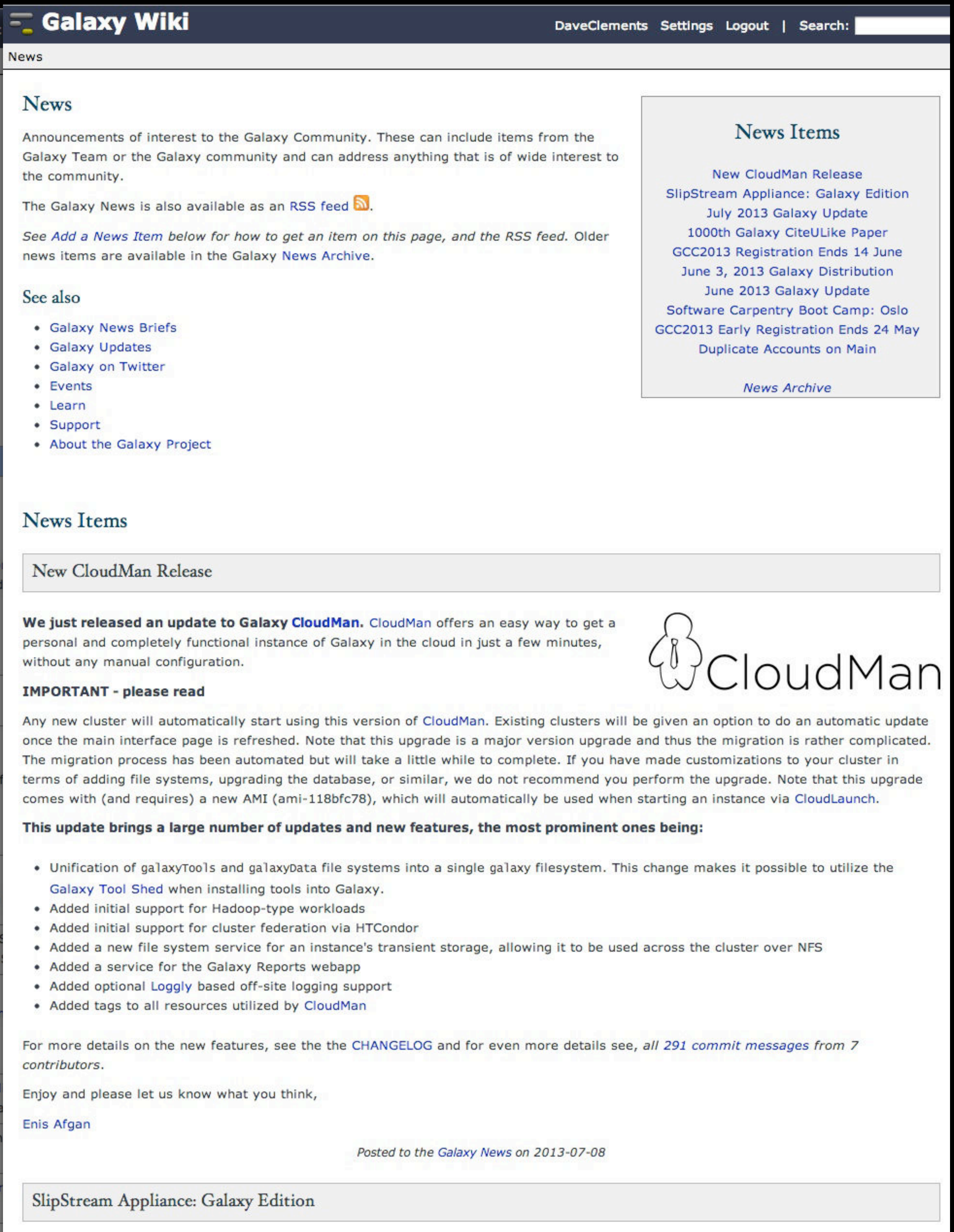
If you know of any event that should be added to this page and/or to the Galaxy Event Calendar, please add it here or send it to outreach@glaxyproject.org .

Upcoming Events

Research Triangle 2013 Galaxy Workshop Tour
2013 ISMB ECCB berlin
XSEDE Extreme Science and Engineering Discovery Environment

Date	Topic/Event	Venue/Location
July 18-23	*Introduction to Galaxy Workshop* National Institute of Environmental Health Sciences (NIEHS)	2013 Research Triangle Wo Tour, North Carolina, United
	Introduction to Galaxy Workshop University of North Carolina, Chapel Hill	
	Galaxy Installation Tutorial **2013 GMOD Summer School**	
	Introduction to Galaxy Workshop North Carolina State University	
July 19-23	**ISMB/ECCB, BOSC and MS SIG 2013** Talks, posters and workshops. Lots of them.	Berlin, Germany
July 21-25	*Experiences in building a Next-Generation Sequencing Analysis Service using Galaxy, Globus Online, and Amazon Web Services*	**XSEDE13**, San Diego, Calif United States
	A Sustainable National Gateway for Biological Computation	
	Supporting Genomics and other Biological Research	
September 28 - October 1	*Galaxy Workshop*	The Genomic Bioinformatics Workshop, Sydney, Australia
October 1-3	*Galaxy*	**Beyond the Genome 2013**, Francisco, California, United
October 7-8	*TBD*	NGS & Bioinformatics Su Europe
	Using Galaxy to Provide a NGS Analysis Platform	
October 9-11	*Galaxy Training Days*	GenoToul bioinformatics facil INRA, Toulouse Auzeville, Fra
October 22-26	*High Throughput Data Analysis and Visualization with Galaxy*	ASHG 2013, Boston, Massach United States
November 6-12	*Computational and Comparative Genomics Course* Application Deadline: July 15, 2013	Cold Spring Harbor Laborator York, United States

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.

News Items

New CloudMan Release
SlipStream Appliance: Galaxy Edition
July 2013 Galaxy Update
1000th Galaxy CiteULike Paper
GCC2013 Registration Ends 14 June
June 3, 2013 Galaxy Distribution
June 2013 Galaxy Update
Software Carpentry Boot Camp: Oslo
GCC2013 Early Registration Ends 24 May
Duplicate Accounts on Main

News Archive

See also

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

News Items

New 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.

IMPORTANT - please read

Any new cluster will automatically start using this version of CloudMan. Existing clusters will be given an option to do an automatic update once the main interface page is refreshed. Note that this upgrade is a major version upgrade and thus the migration is rather complicated. The migration process has been automated but will take a little while to complete. If you have made customizations to your cluster in terms of adding file systems, upgrading the database, or similar, we do not recommend you perform the upgrade. Note that this upgrade comes with (and requires) a new AMI (ami-118bfc78), which will automatically be used when starting an instance via CloudLaunch.

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

- Unification of `galaxyTools` and `galaxyData` file systems into a single `galaxy` filesystem. This change makes it possible to utilize the Galaxy Tool Shed when installing tools into Galaxy.
- Added initial support for Hadoop-type workloads
- Added initial support for cluster federation via HTCondor
- Added a new file system service for an instance's transient storage, allowing it to be used across the cluster over NFS
- Added a service for the Galaxy Reports webapp
- Added optional Loggly based off-site logging support
- Added tags to all resources utilized by CloudMan

For more details on the new features, see the the CHANGELOG and for even more details see, *all 291 commit messages from 7 contributors*.

Enjoy and please let us know what you think,

Enis Afgan

Posted to the Galaxy News on 2013-07-08

SlipStream Appliance: Galaxy Edition

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | SUMMER 2014

http://bit.ly/gcc2014

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

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Dave Clements

Nate Coraor

Carl Eberhard

Dorine Francheteau

Jeremy Goecks

Sam Guerler

Jen Jackson

Greg von Kuster

Ross Lazarus

Anton Nekrutenko

James Taylor

http://wiki.galaxyproject.org/GalaxyTeam

Galaxy is hiring post-docs and software engineers

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

RNA-seq Exercise: A Plan

- ...
- Map the reads to the human reference using Tophat
- 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.*

RNA-seq Exercise: A Plan

- ...
- Map the reads to the human reference using Tophat
- 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 it

Visualizing Genomics

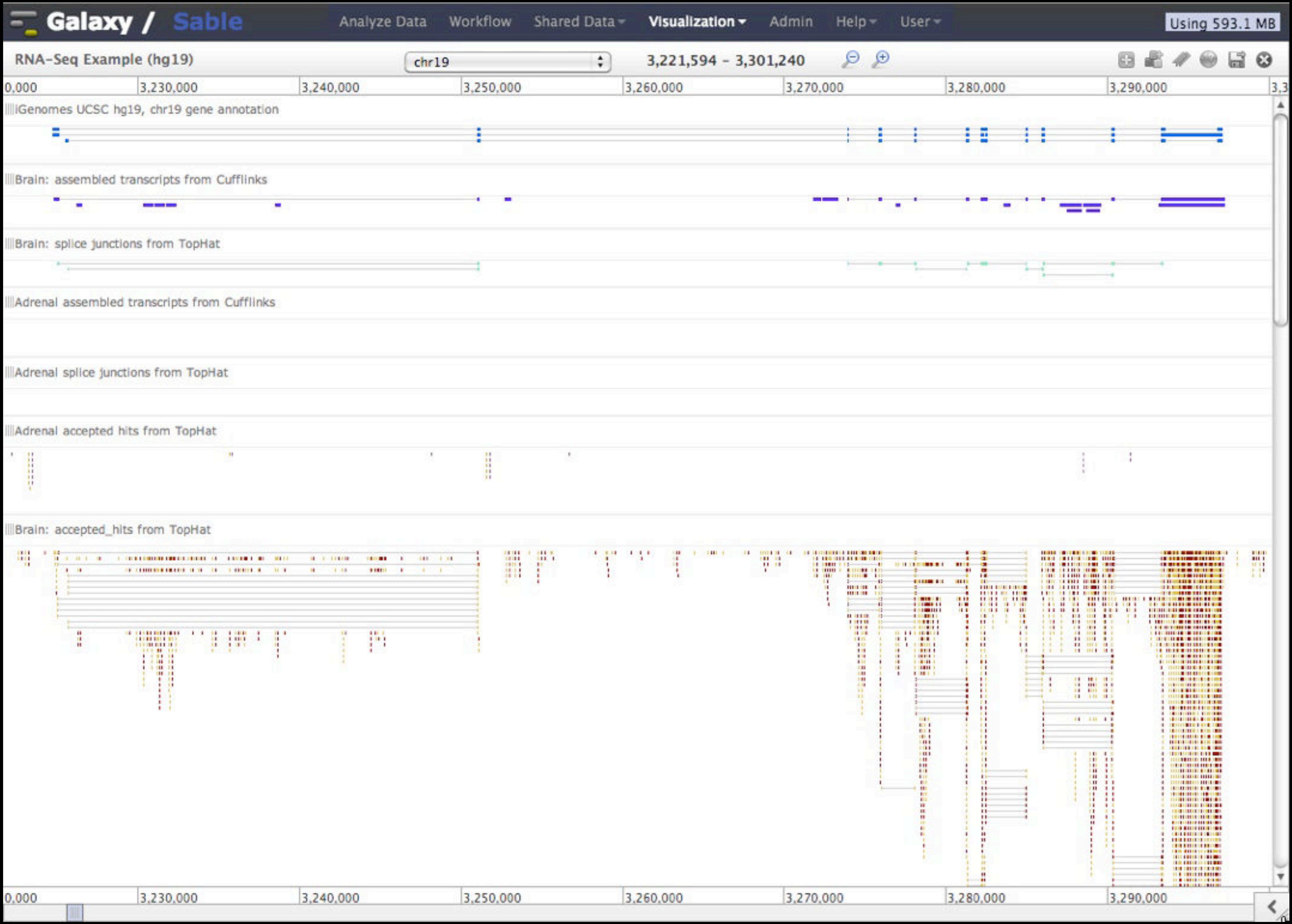
Supported external browsers

- UCSC
- Ensembl
- GBrowse
- IGB
- IGV

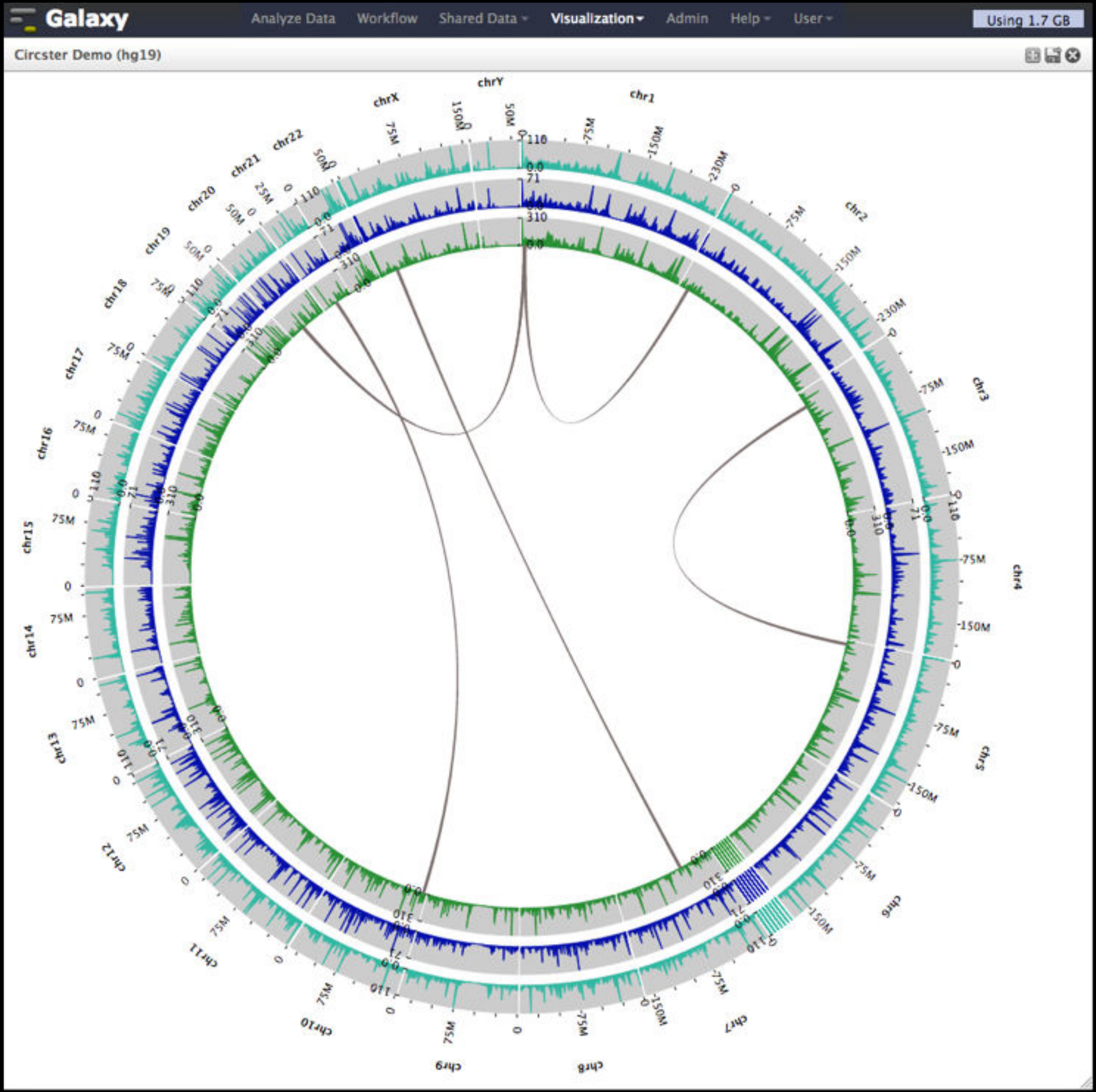
Traditional browser strengths:

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

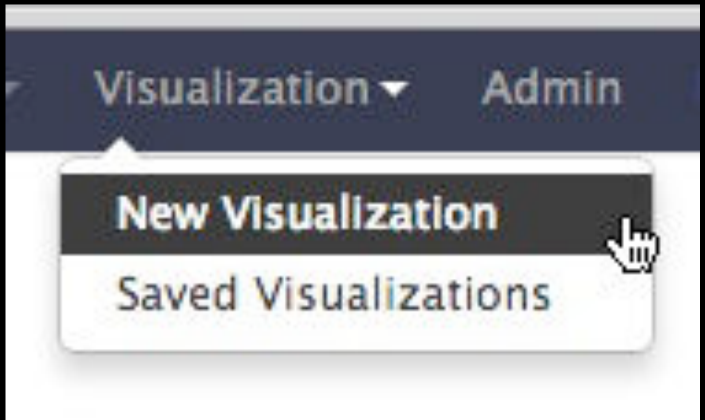
Trackster: Galaxy's embedded track browser


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
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

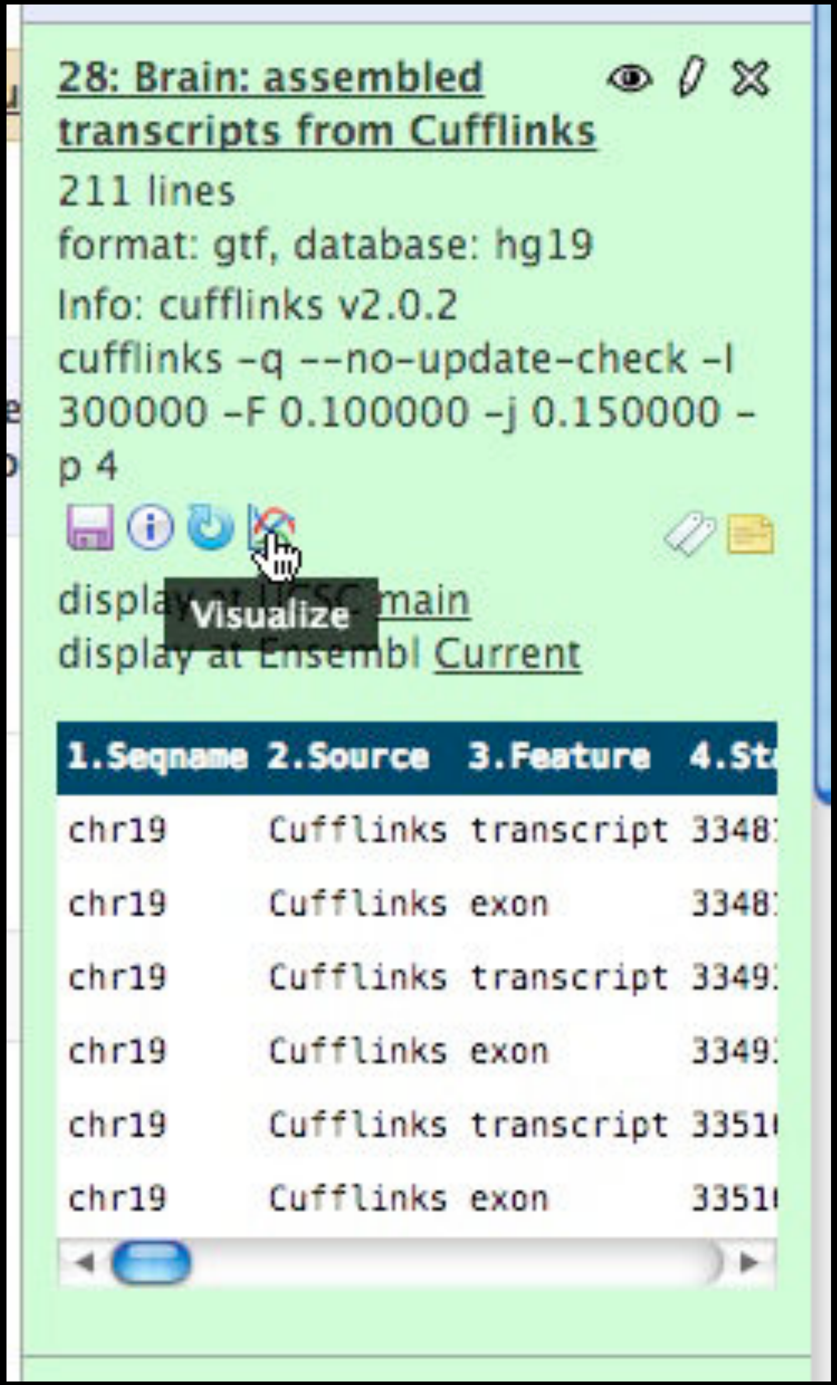
Circster


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 1.7 GB
Circster Demo (hg19)
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
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
- Expose basic analyses in visualization to make it more informative
- Make the *analyze-visualize-refine* loop seamless and fast
- Enable experimenting with tools and their parameter space
- Support custom genome browsers

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Sharing & Publishing enables Reproducibility

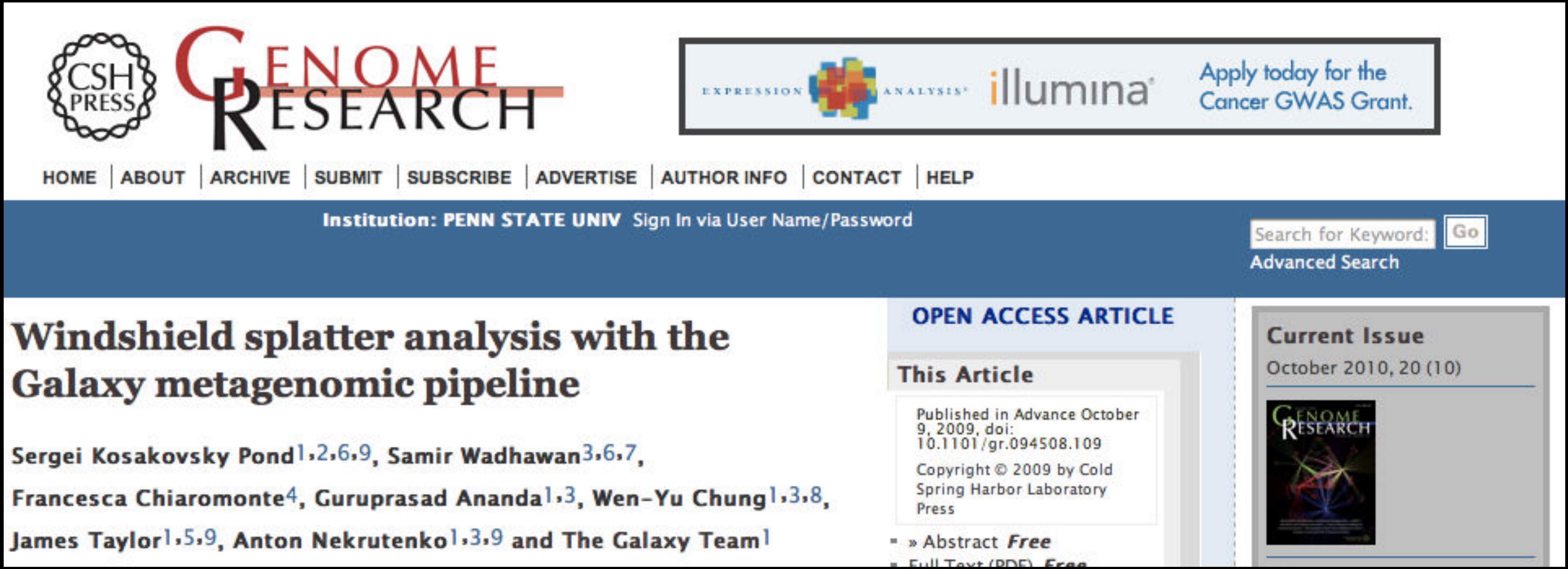
Reproducibility: Everybody talks about it, but ...

Galaxy aims to push the goal of reproducibility from the bench to the bioinformatics realm

All analysis in Galaxy is recorded without any extra effort from the user.

Histories, workflows, visualizations and ***pages*** can be shared with others or published to the world.

Sharing & Publishing enables Reproducibility



CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP

Institution: PENN STATE UNIV Sign In via User Name/Password

Search for Keyword: Go

Advanced Search

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]

OPEN ACCESS ARTICLE

This Article

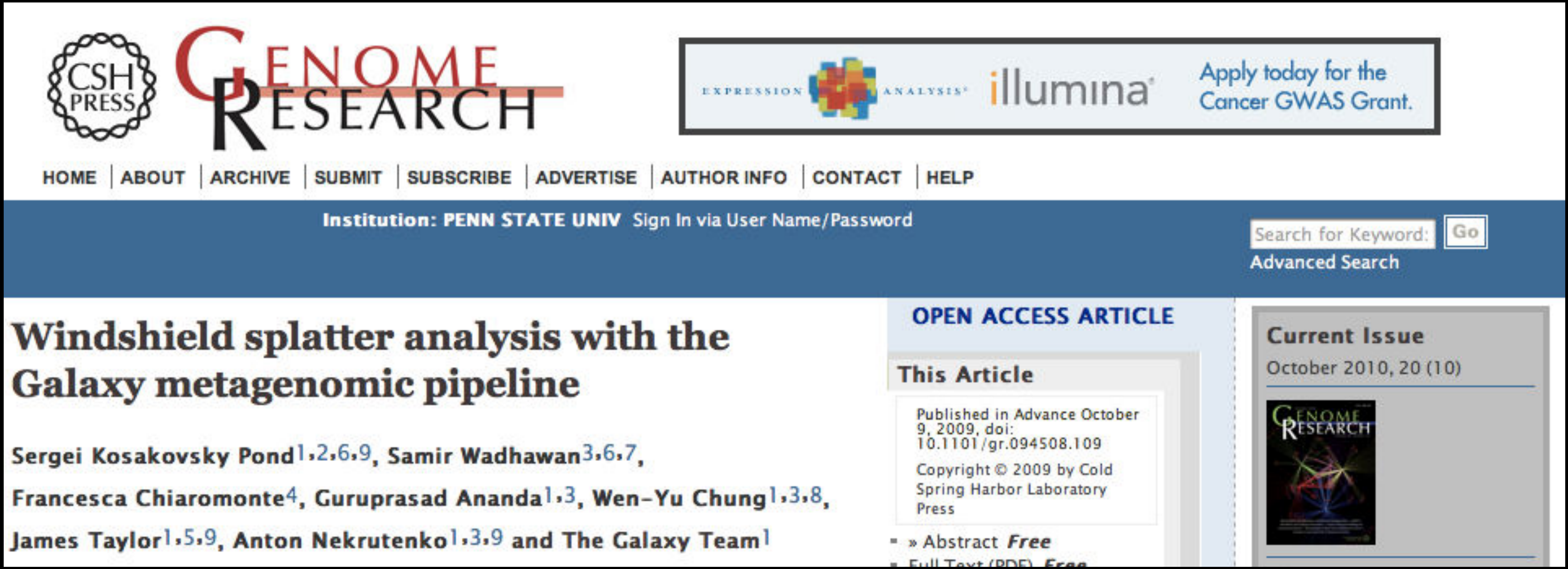
Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

- » Abstract *Free*
- Full Text (PDF) *Free*

Current Issue

October 2010, 20 (10)

Sharing & Publishing enables Reproducibility


CSH PRESS
GENOME RESEARCH
EXPRESSION ANALYSIS
illumina
Apply today for the Cancer GWAS Grant.
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP
Institution: PENN STATE UNIV Sign In via User Name/Password
Search for Keyword:
Go
Advanced Search
OPEN ACCESS ARTICLE
Current Issue
October 2010, 20 (10)
GENOME RESEARCH

Windshield splatter analysis with the Galaxy metagenomic pipeline

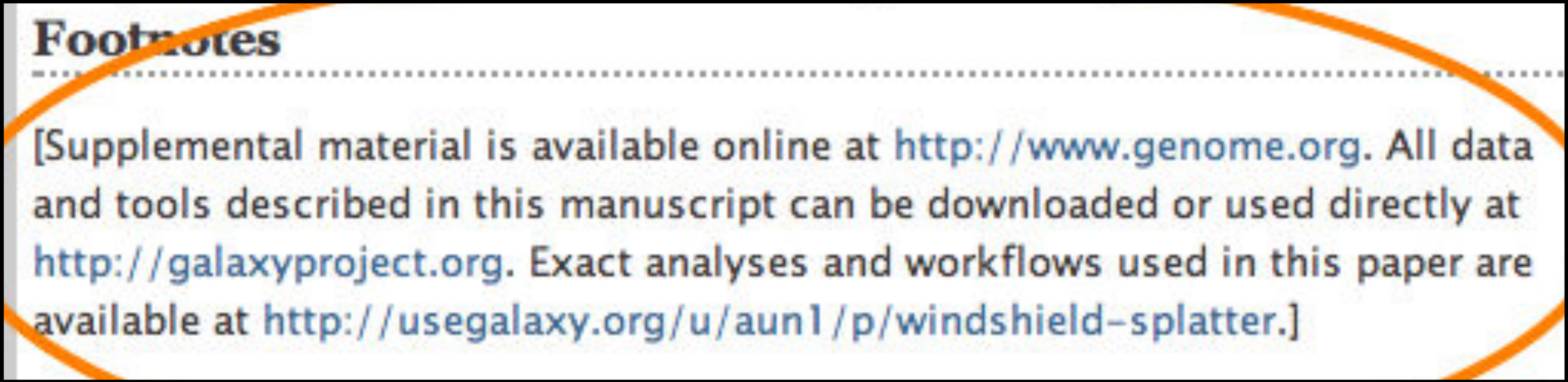
Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

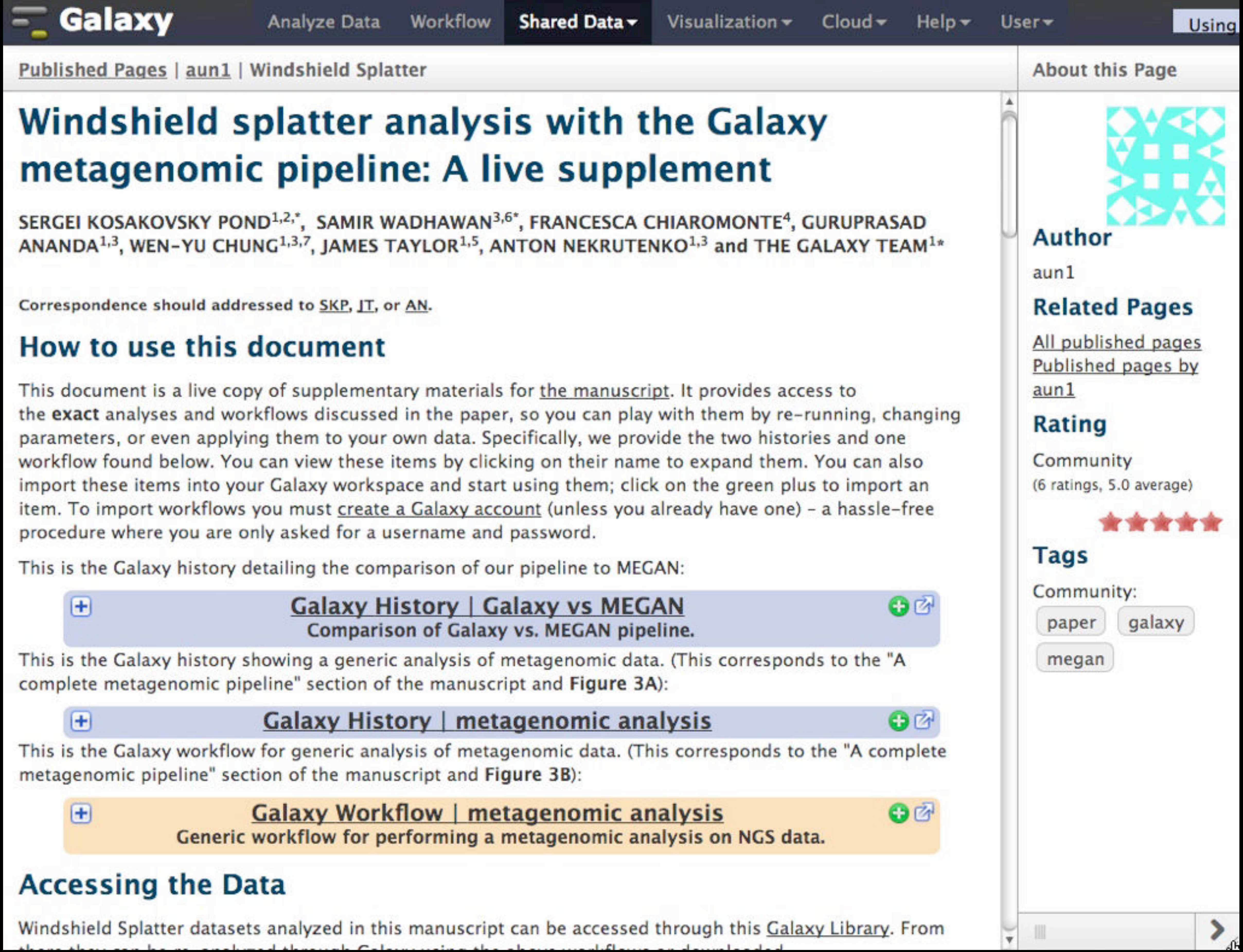
Copyright © 2009 by Cold Spring Harbor Laboratory Press

- » Abstract *Free*
- Full Text (PDF) *Free*



Footnotes

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]



Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

⊞ Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

⊞ Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

⊞ Galaxy Workflow | metagenomic analysis
Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From

Author

aun1

Related Pages

All published pages
Published pages by aun1

Rating

Community
(6 ratings, 5.0 average)
★★★★★

Tags

Community:
paper galaxy megan

http://usegalaxy.org/u/aun1/p/windshield-splatter

Sharing for Galaxy Administrators Too

Data Libraries

Make data easy to find

Genome Builds

Care about a particular subset of life?

Galaxy Tool Shed

Wrapping tools and datatypes

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done

Galaxy CloudMan

http://usegalaxy.org/cloud

- Start with a **fully configured and populated** (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center.
- **We are using this today.**

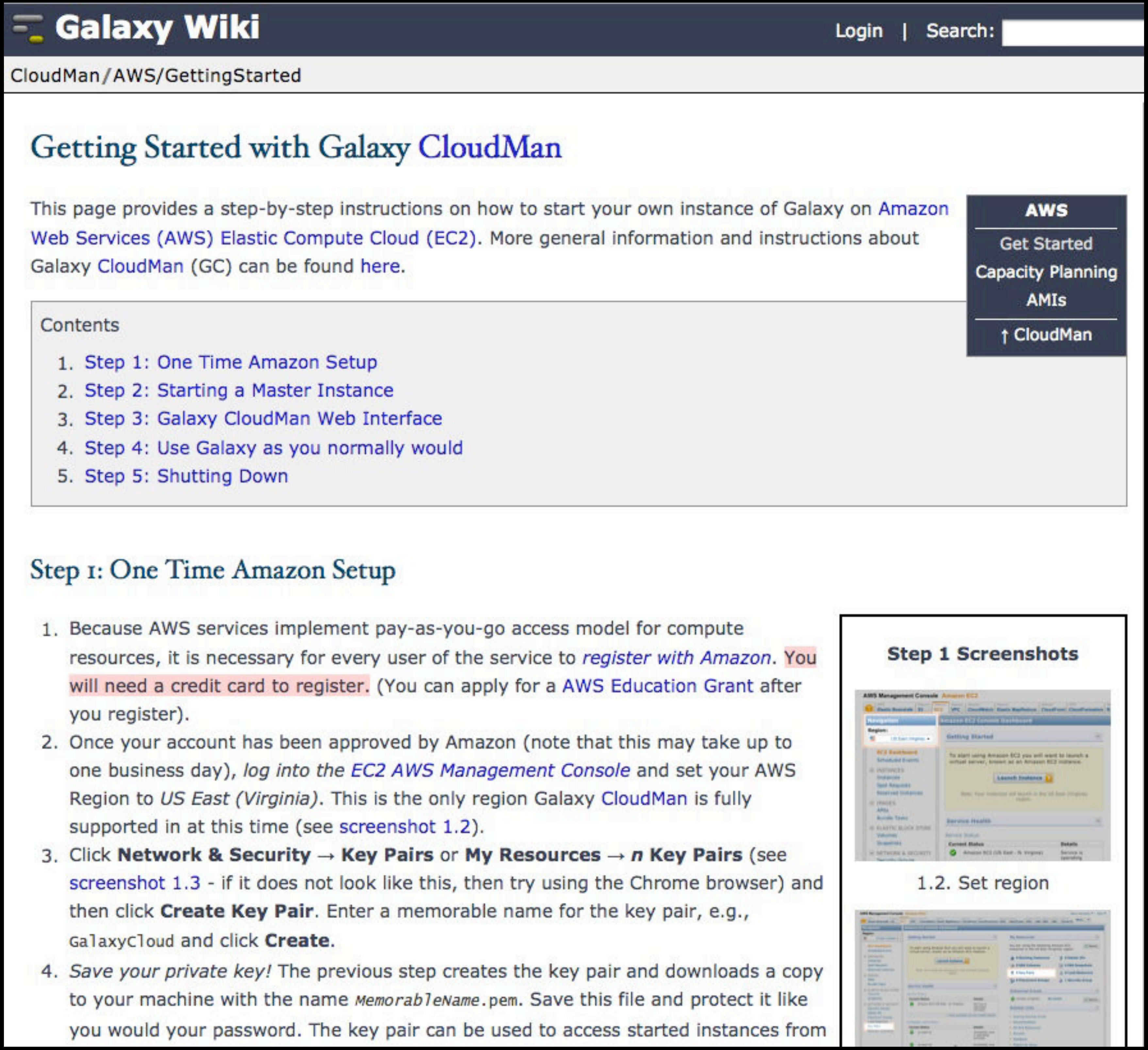
amazon web services™

- **You will set up an instance now**

http://aws.amazon.com/education

Could do this step by step, but ...

http://bit.ly/GXYAWSGetStarted



Galaxy Wiki

Login | Search:

CloudMan/AWS/GettingStarted

Getting Started with Galaxy CloudMan

This page provides a step-by-step instructions on how to start your own instance of Galaxy on Amazon Web Services (AWS) Elastic Compute Cloud (EC2). More general information and instructions about Galaxy CloudMan (GC) can be found here.

AWS

Get Started

Capacity Planning

AMIs

↑ CloudMan

Contents

1. Step 1: One Time Amazon Setup
2. Step 2: Starting a Master Instance
3. Step 3: Galaxy CloudMan Web Interface
4. Step 4: Use Galaxy as you normally would
5. Step 5: Shutting Down

Step 1: One Time Amazon Setup

1. Because AWS services implement pay-as-you-go access model for compute resources, it is necessary for every user of the service to *register with Amazon*. You will need a credit card to register. (You can apply for a AWS Education Grant after you register).
2. Once your account has been approved by Amazon (note that this may take up to one business day), *log into the EC2 AWS Management Console* and set your AWS Region to *US East (Virginia)*. This is the only region Galaxy CloudMan is fully supported in at this time (see screenshot 1.2).
3. Click **Network & Security** → **Key Pairs** or **My Resources** → ***n* Key Pairs** (see screenshot 1.3 - if it does not look like this, then try using the Chrome browser) and then click **Create Key Pair**. Enter a memorable name for the key pair, e.g., `GalaxyCloud` and click **Create**.
4. *Save your private key!* The previous step creates the key pair and downloads a copy to your machine with the name `*MemorableName*.pem`. Save this file and protect it like you would your password. The key pair can be used to access started instances from

Step 1 Screenshots

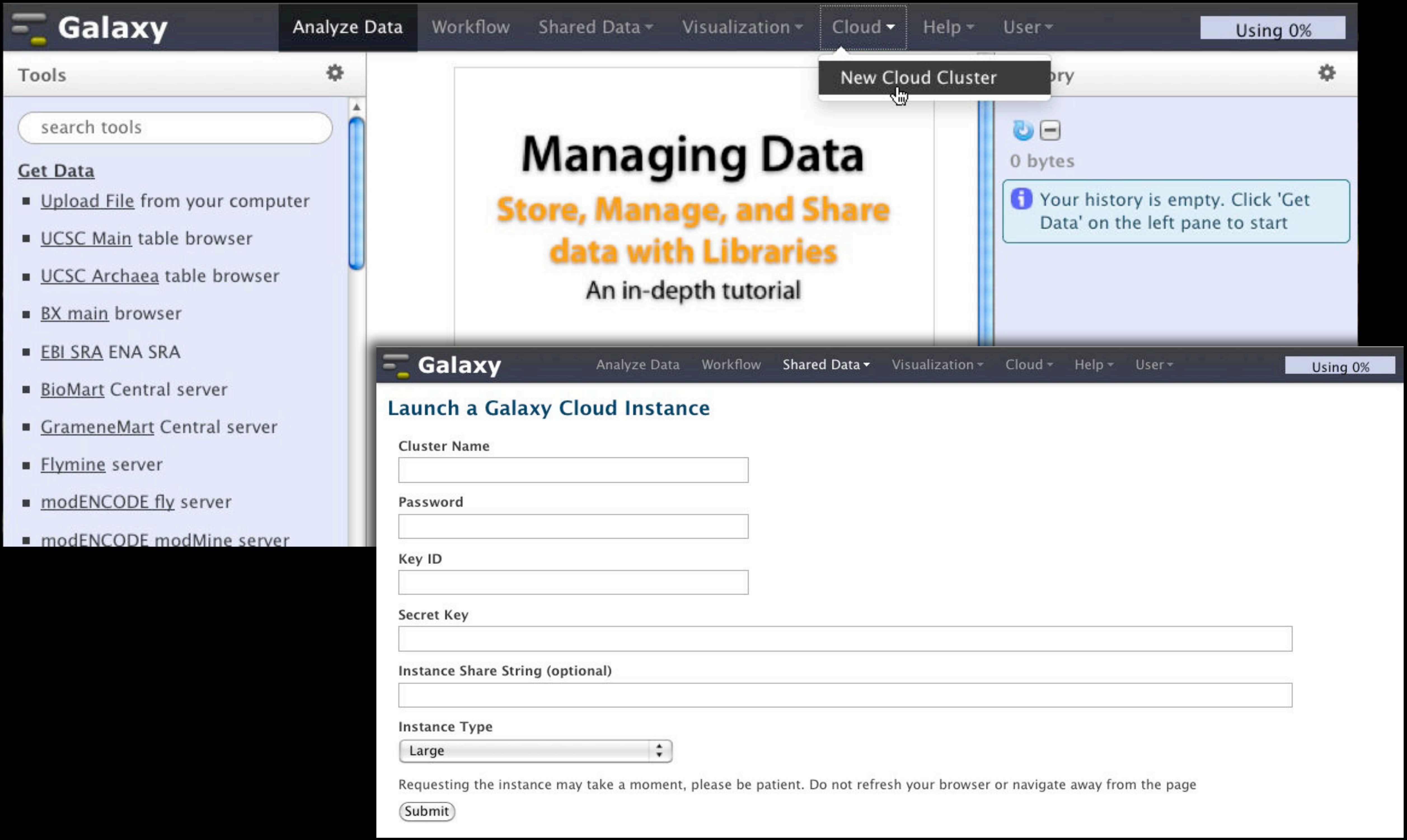
1.2. Set region

AWS Credentials

http://bit.ly/niehsAWS

Instant CloudMan

http://usegalaxy.org/cloudlaunch



Launch a Galaxy Cloud Instance

Cluster Name

Password

Key ID

Secret Key

Instance Share String (optional)

Instance Type

Large

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

Submit

http://bit.ly/niehsAWS

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy
10:40	Break
11:00	Basic Analysis into Reusable Workflows
11:20	RNA-Seq Example Part I
12:20	Lunch
1:35	Galaxy Project Overview
1:55	RNA-Seq Example Part II
2:45	Break
3:05	Sharing, Publishing and Reproducibility
3:25	Setting up Galaxy on the Amazon Cloud
5:00	Done, almost

Instant Feedback

http://bit.ly/20130718Gxy

Acknowledgements

Tom Randall
Debbie Wilson
You

Trudy Mackay
Barrie Hayes
The Galaxy Team

NIEHS

AWS Education Grant

NIH NSF Huck Institute
Penn State University Emory University

http://bit.ly/20130718Gxy

Thanks

http://bit.ly/20130718Gxy

Dave Clements

Galaxy Project
Emory University

clements@galaxyproject.org