Galaxy for NGS Data Analysis

A Hands-on Workshop

Plant & Animal Genome XXII
San Diego, January 14, 2014

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

Anushka Brownley
The BioTeam
http://bioteam.net

The Agenda

4:00 Introduction to Galaxy

Hands-on Analysis

Running a Local Galaxy

Community Resources

6:10 Done

This workshop complements tomorrow's talks:

http://bit.ly/gxypag2014

Galaxy @

Plant and Animal Genome XXII (PAG 2014)

UCSC Genome Browser

Sat 4:00-6:10, California Room
Robert Kuhn

The Banana Genome Hub

Tues 11:50-12:10, Pacific Salon 6-7
Gaëtan Droc, *et al.*

Galaxy for NGS Analysis
A Hands-on Workshop

Tues 4:00-6:10, California Room
Dave Clements, Anushka Brownley

This workshop will introduce the Galaxy platform and walk participants through a multi-step next generation sequencing data analysis, starting with quality control. We will review common choices in NGS data analysis, and demonstrate them within the context of Galaxy, taking advantage of Galaxy's tool set and visualization capabilities.

We will also provide a brief overview of what is needed to set up your own local Galaxy instance. This complements the *Galaxy CloudMan* talk on Wednesday

URGI Plant and Fungi Platform
Distributed Resources Through GMOD Tools

Wed 11:10-11:50, GMOD Workshop Golden West
Joelle Amselem, *et al.*

Galaxy CloudMan
A Gentle Introduction to Data Analysis on the Cloud

Wed 11:50-12:30, GMOD Workshop, Golden West
Dave Clements

Galaxy is open-source and web-based, with over 50 publicly accessible Galaxy servers and hundreds of private installations around the world. Galaxy can also be run on compute clouds using *Galaxy CloudMan*.

This talk will briefly introduce Galaxy, Galaxy CloudMan, and some basic cloud concepts. We'll then show a live demonstration of how to setup a Galaxy server on Amazon Web Services (one of several supported cloud infrastructures) using CloudMan, add a dynamically scalable compute cluster to perform analysis, customize the server by adding new tools, and then shut the server down. All steps can be done through a web browser, without ever using a command line interface.

Poster Sessions

Mon 10:00-11:30

P988: The South Green Bioinformatics Platform, Mathieu Rouard, *et al.*

P1050: Integrative System for Gene Family Gathering and Analysis in a Context of Crops' Stress Response Study, Delphine Lavivière, *et al.*

Mon 3:00-4:30

P135: SNP Genotyping to Accelerate Rice Breeding, Michael Thomson, *et al.*

P1041: RepeatExplorer: Collection of Tools for Mining of Repetitive Elements from NGS Data, Petr Novak, *et al.*

The Galaxy Project

Galaxy is an open source web-based platform for data integration and analysis in life sciences research.

The Galaxy Project is supported by a large and active community.

http://galaxyproject.org

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.

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

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 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

Anushka will cover this more later

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.

Galaxy CloudMan: A Gentle Introduction to Data Analysis on the Cloud

Wed, 11:50, Golden West

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*

A ready-to-use appliance (BioTeam)

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

Consulting & Customization (Arctix, 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

Agenda

4:00 Introduction to Galaxy

Hands-on Analysis

Quality Control

Running a Local Galaxy

Community Resources

6:10 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://cloud1.galaxyproject.org/
http://cloud2.galaxyproject.org/
http://cloud3.galaxyproject.org/

NGS Data Analysis Experience?

Novice — Middling — Expert

I'm getting a terabyte of data! That's a good thing, isn't it?

I've done some, but there has to be an easier way.

Just here to see if Galaxy can help others less fortunate than myself.

NGS Data Quality Control

- Introduce 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!

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./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

→ UC Davis RNA-Seq Human*

→ Select MeOH_REP1_R1, MeOH_REP1_R2
and then Import to current history

UCDAVIS Genome Center Bioinformatics Core

* 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

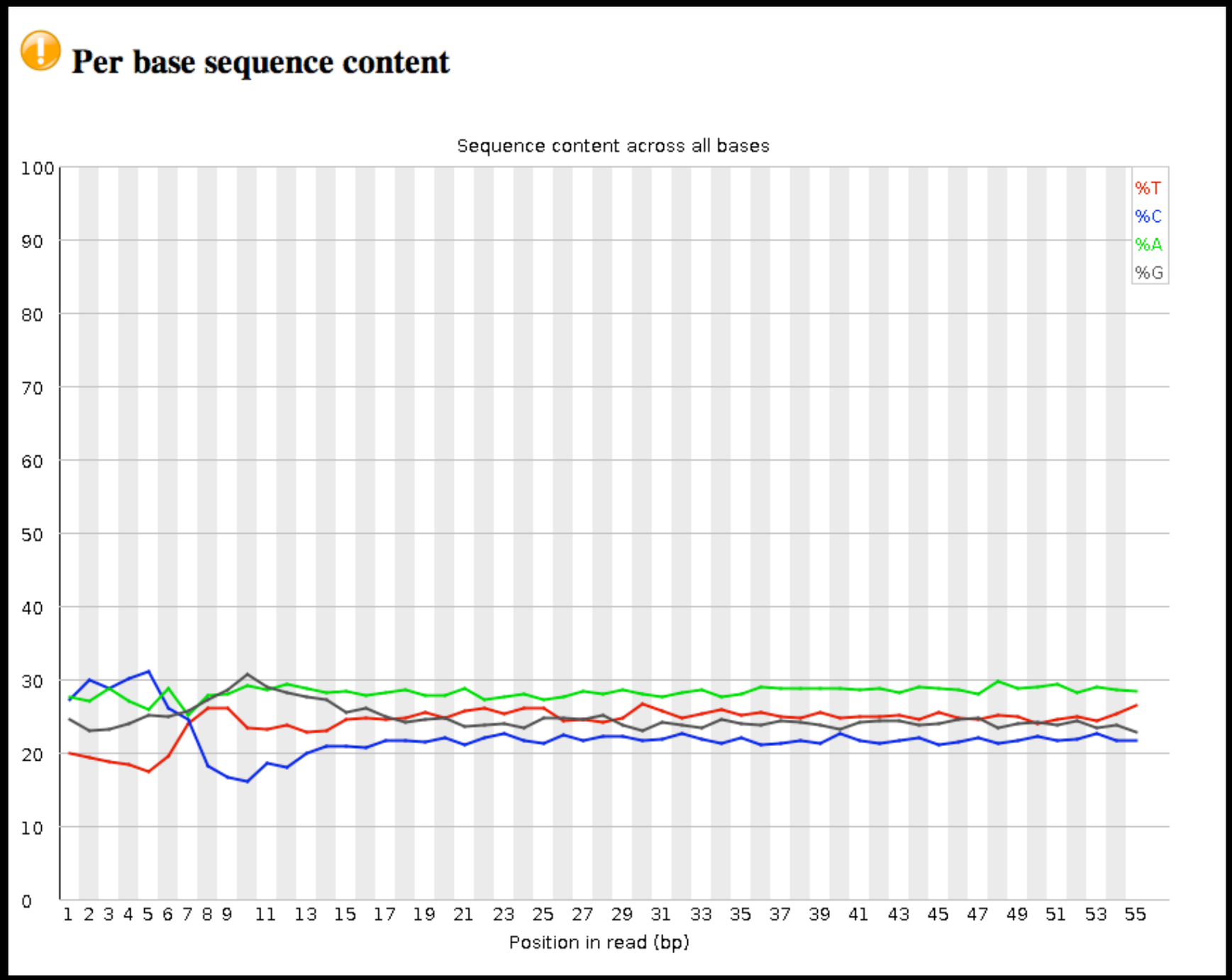
NGS Data Quality: Assessment tools

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?

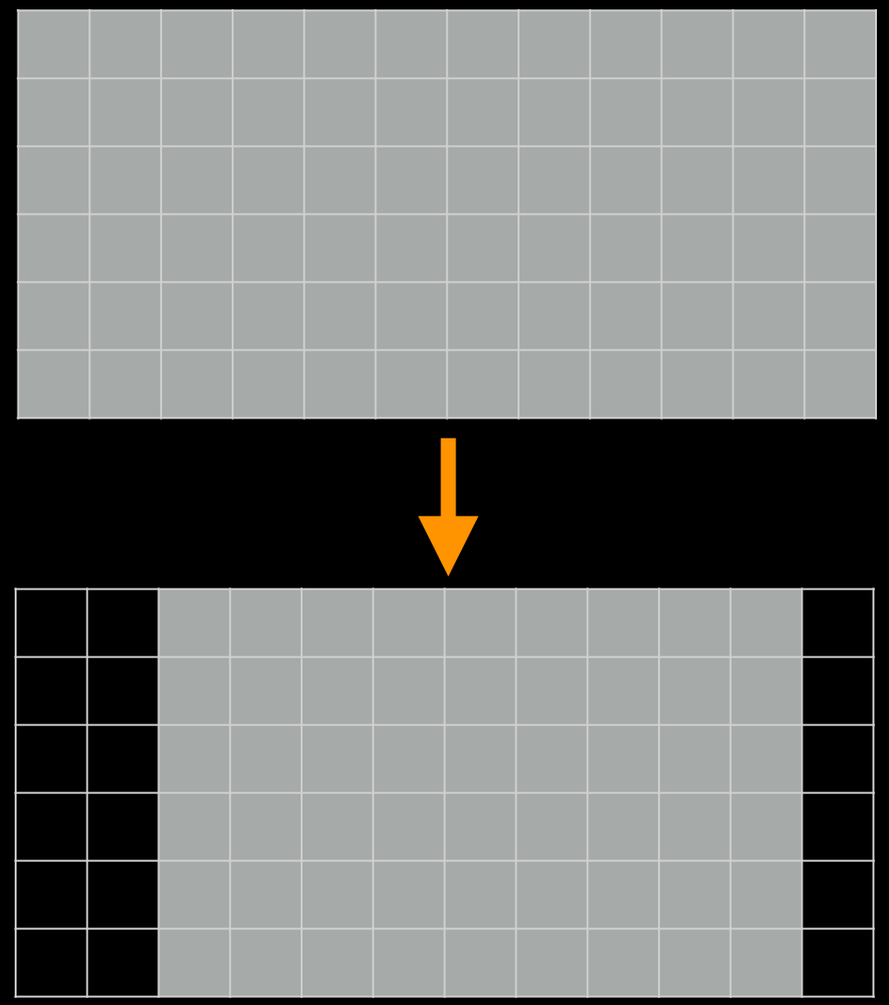


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

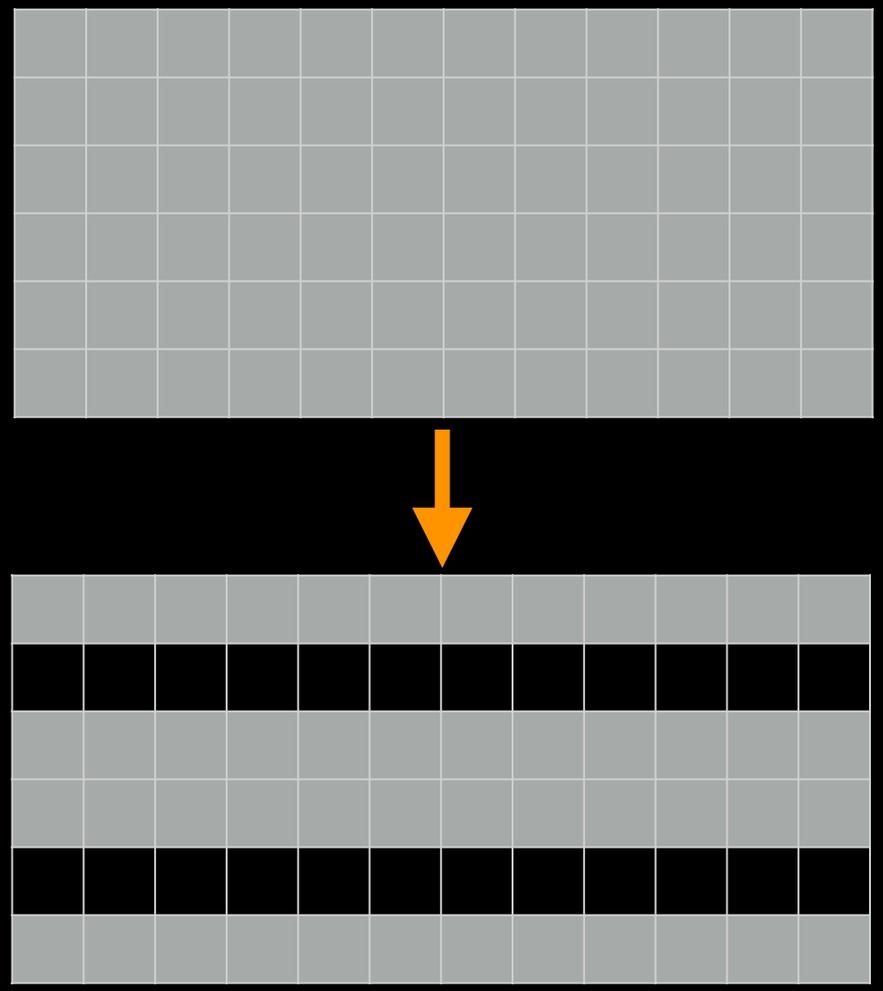
NGS Data Quality: Trim as we see fit

- 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



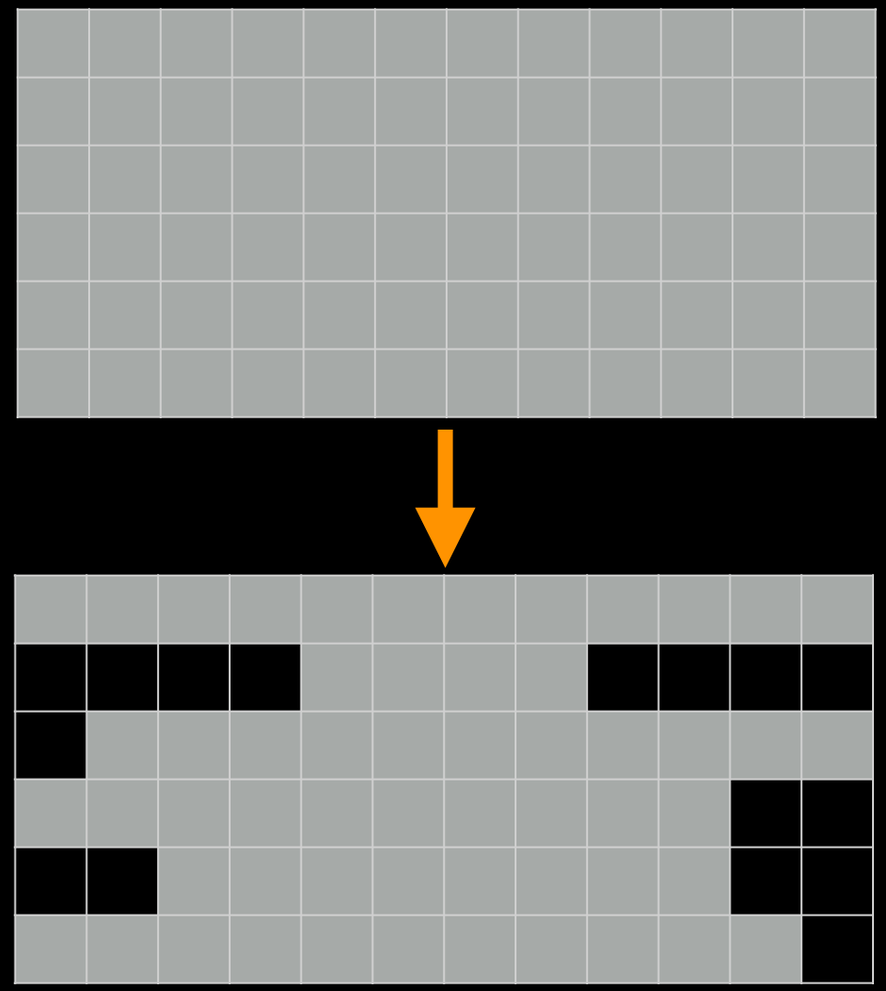
NGS Data Quality: Base Quality Trimming

- ~~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.

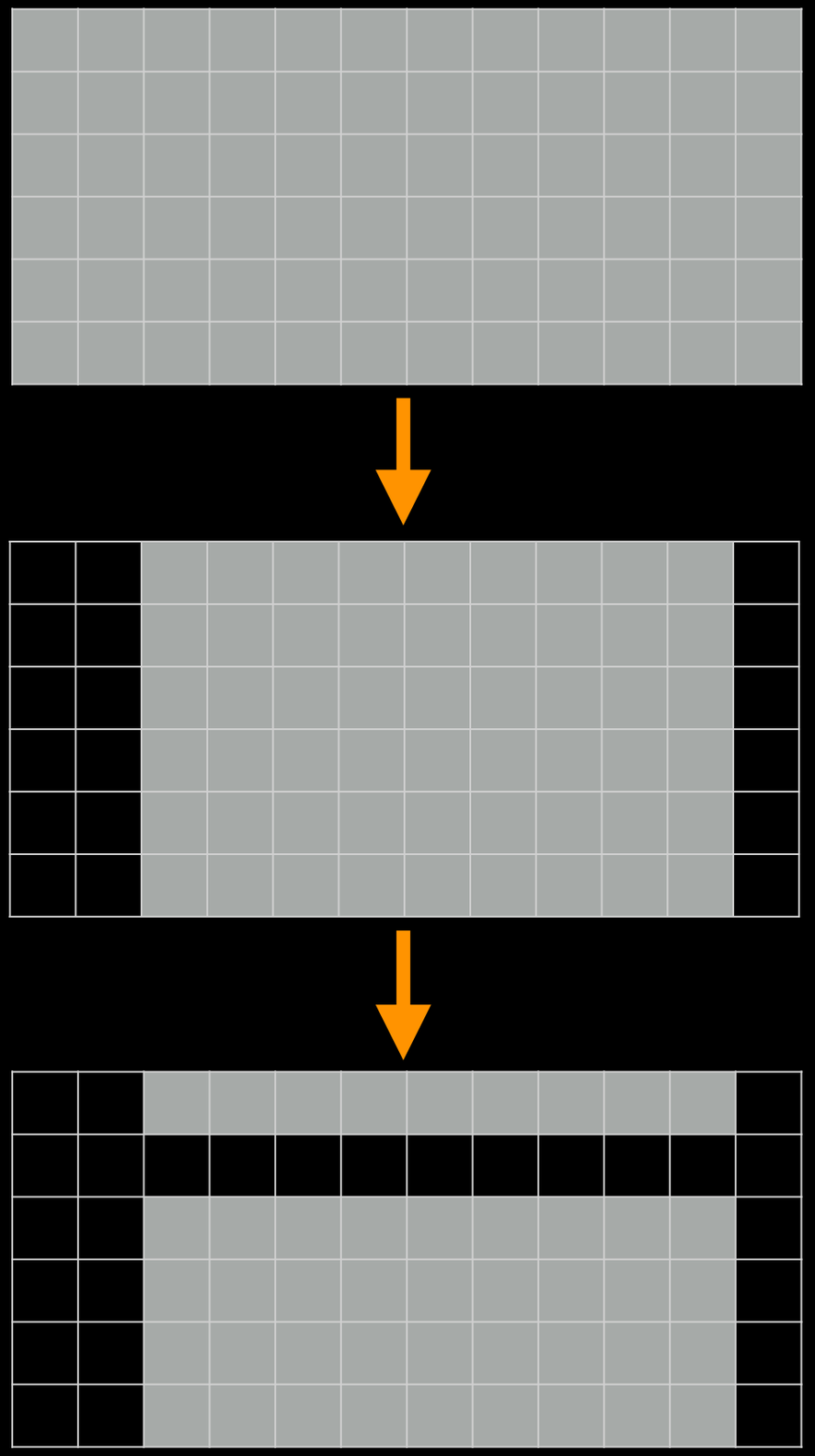


NGS Data Quality: Base Quality Trimming

- 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)

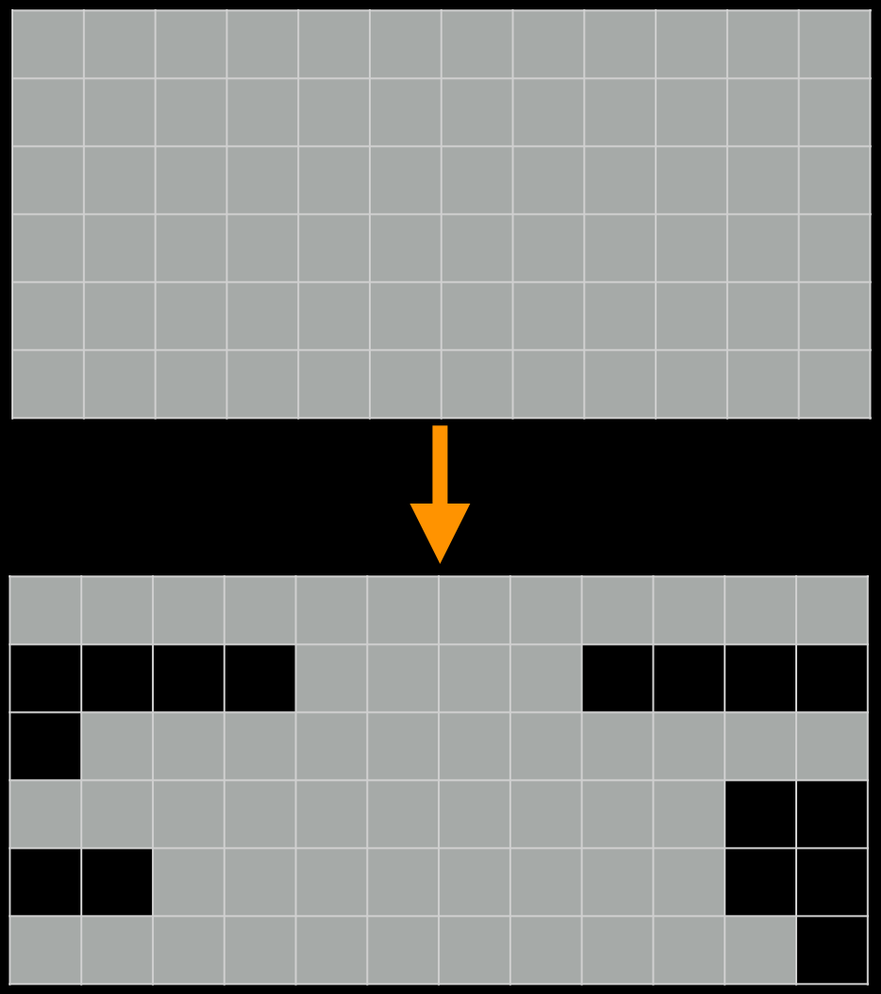
Trim? *As we see fit?*

- 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

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."

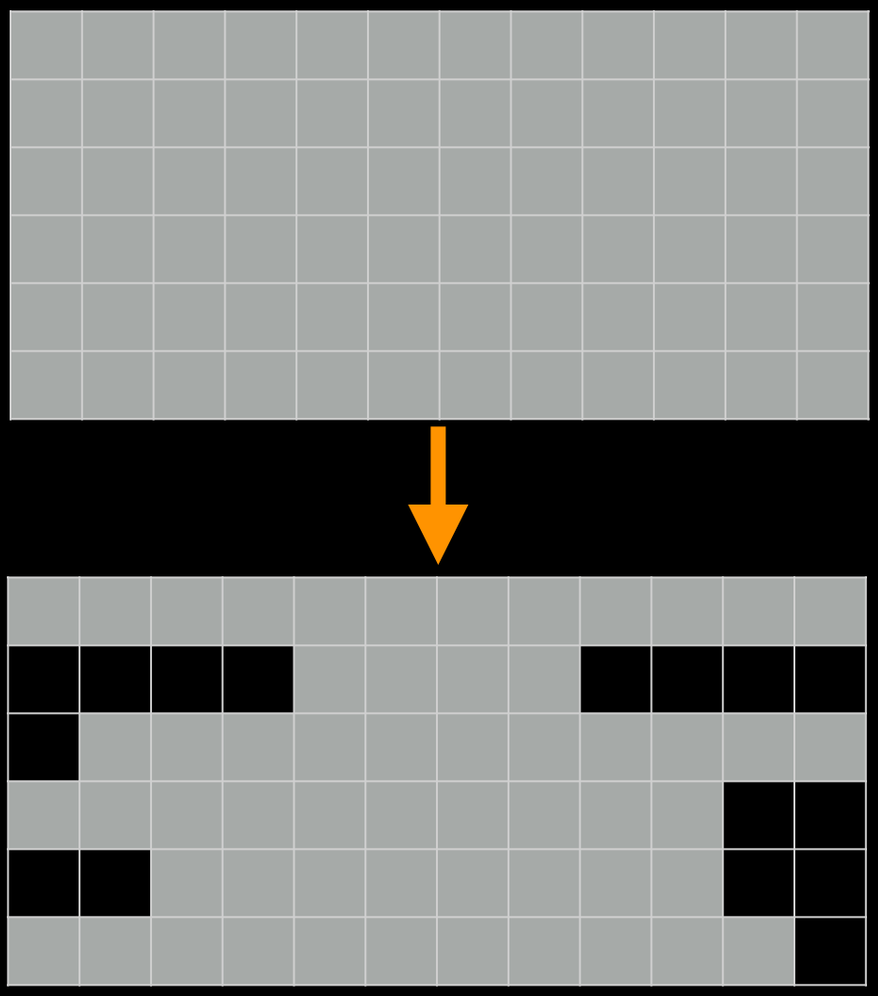
Dave C, mortal

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
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	342	0.2052981325073385	No Hit
CGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	325	0.1950932545756871	No Hit
CTGCATTATAAAAAGGACAGCCAGATATCAACTGTTACAGAAATGAAATAAGACG	230	0.1380659955458709	No Hit
CGGCCGCAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	199	0.1194571004940361	No Hit
GTCAGCTCAACTTGTAGGCCCCAAAAGAAAACAGCGTCTTACTGGGGAGGGATAT	197	0.1182565266197242	No Hit

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

But this will break pairings.

NGS Data Quality: Done with 1st Replicate!

Now, only 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.

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

4:00 Introduction to Galaxy

Hands-on Analysis

Mapping with TopHat

Running a Local Galaxy

Community Resources

6:10 Done

RNA-seq Exercise: Mapping with Tophat

Create a new history

Import all datasets from library:

UC Davis RNA-Seq → RNA-Seq reads filtered

Get all datasets, and

UC Davis RNA-Seq → Chr12

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: **How did we do?**

NGS: SAM Tools → flagstat

Mapping with Tophat: **Lets do it some more!**

NGS: RNA Analysis → TopHat

for the remaining 3 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

4:00 Introduction to Galaxy

Hands-on Analysis

Running a Local Galaxy

Community Resources

6:10 Done

Agenda

4:00 Introduction to Galaxy

Hands-on Analysis

Running a Local Galaxy

Community Resources

6:10 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

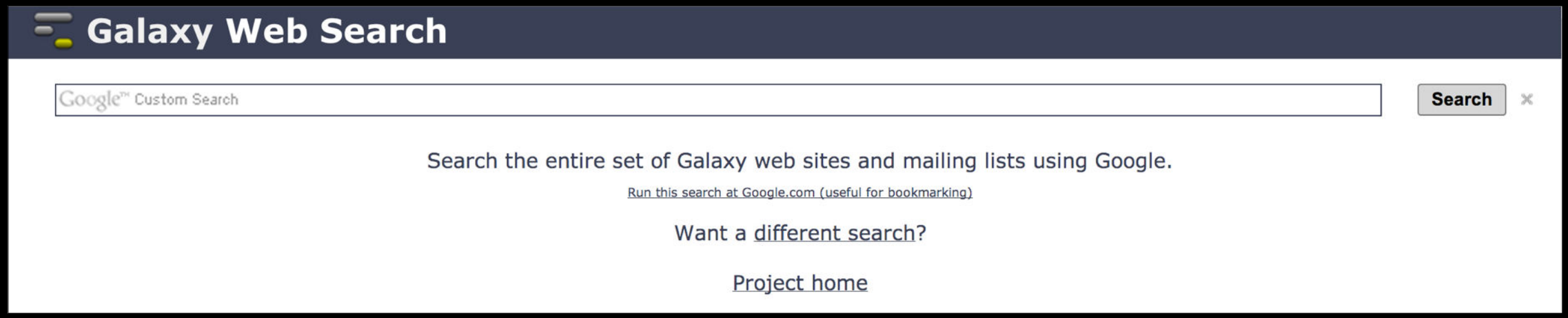
Reasoning with ontologies?
- ✓ OPPL Galaxy

Repeats!
- ✓ RepeatExplorer

Everything?
- ✓ Andromeda

Over 50 public Galaxy servers

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



Find

Everything on …

Tools for …

Email about …

Source code for …

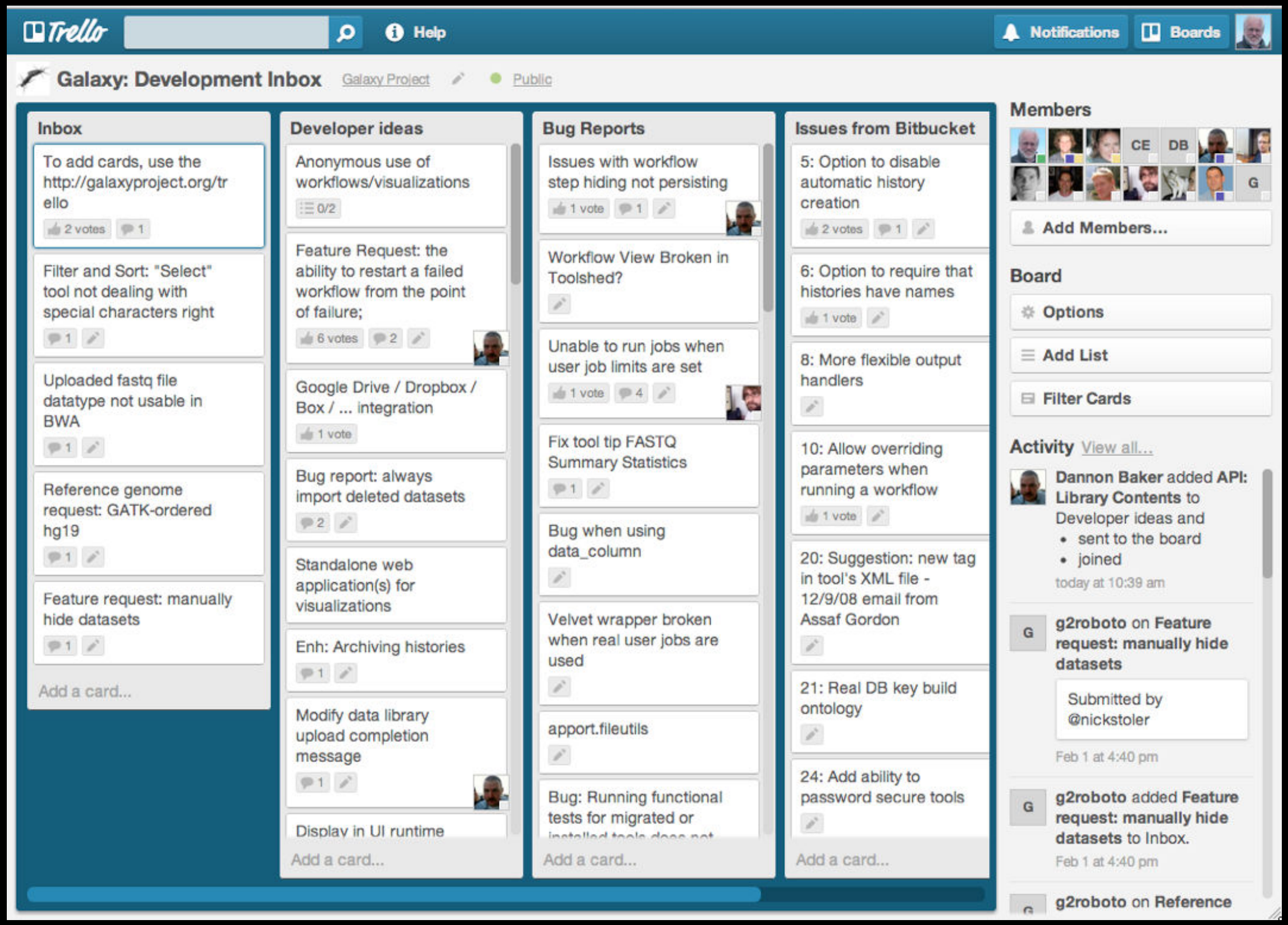
Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

Community can create, vote and comment on issues



http://bit.ly/gxyissues

http://wiki.galaxyproject.org

Galaxy Wiki

Login | Search: Titles Text

FrontPage

Locked History Actions

Galaxy

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 Wiki

Login | Search:

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 .

Contents

1. Upcoming Eve
2. Other Calenda
3. Past Events
    1. 2014
    2. Archive

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

Galaxy Wiki

Login | Search:

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.

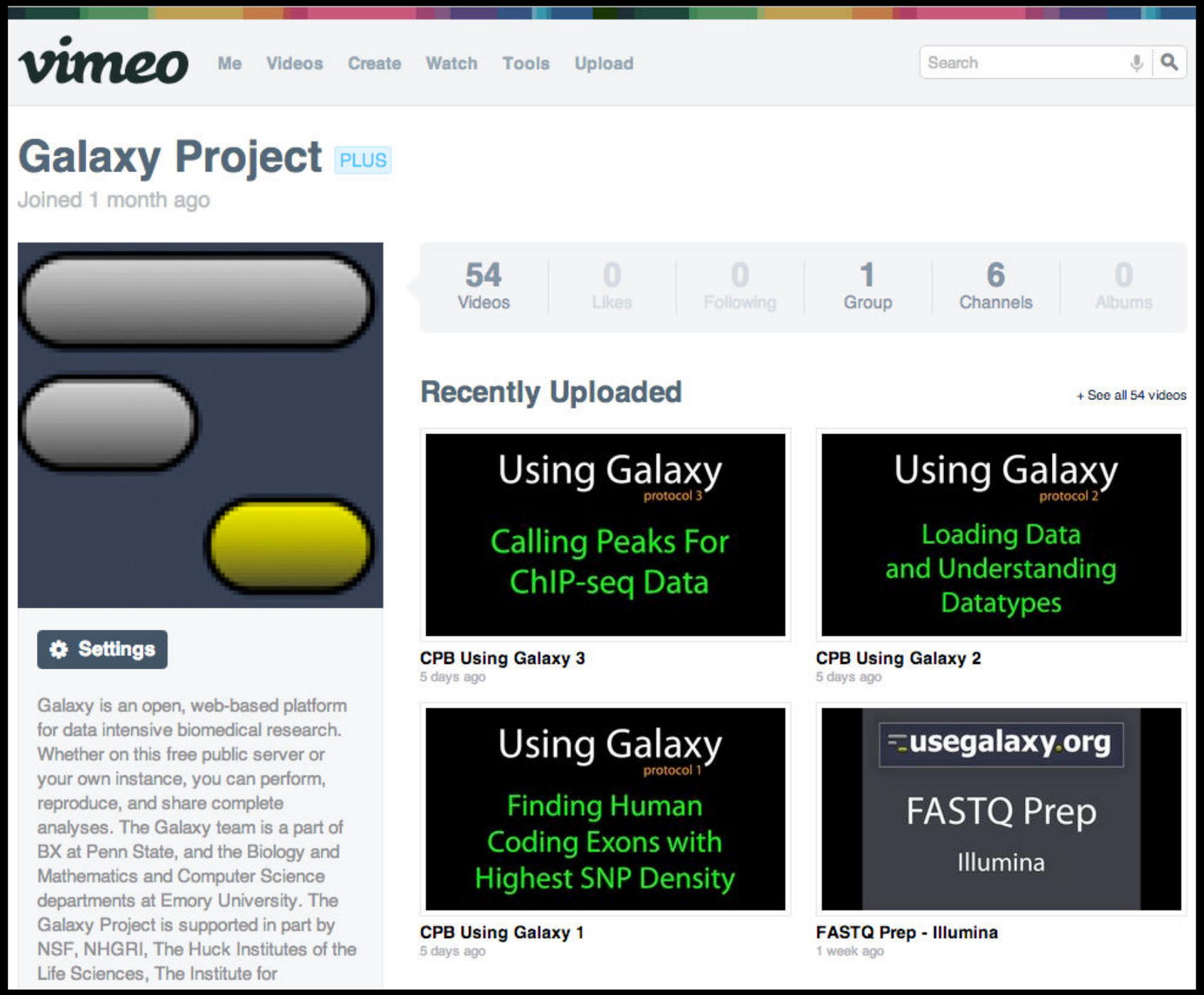
This update brings a large number of updates and new features, the most prominent ones being:

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

http://bit.ly/gcc2014

Galaxy Resources & Community: Videos

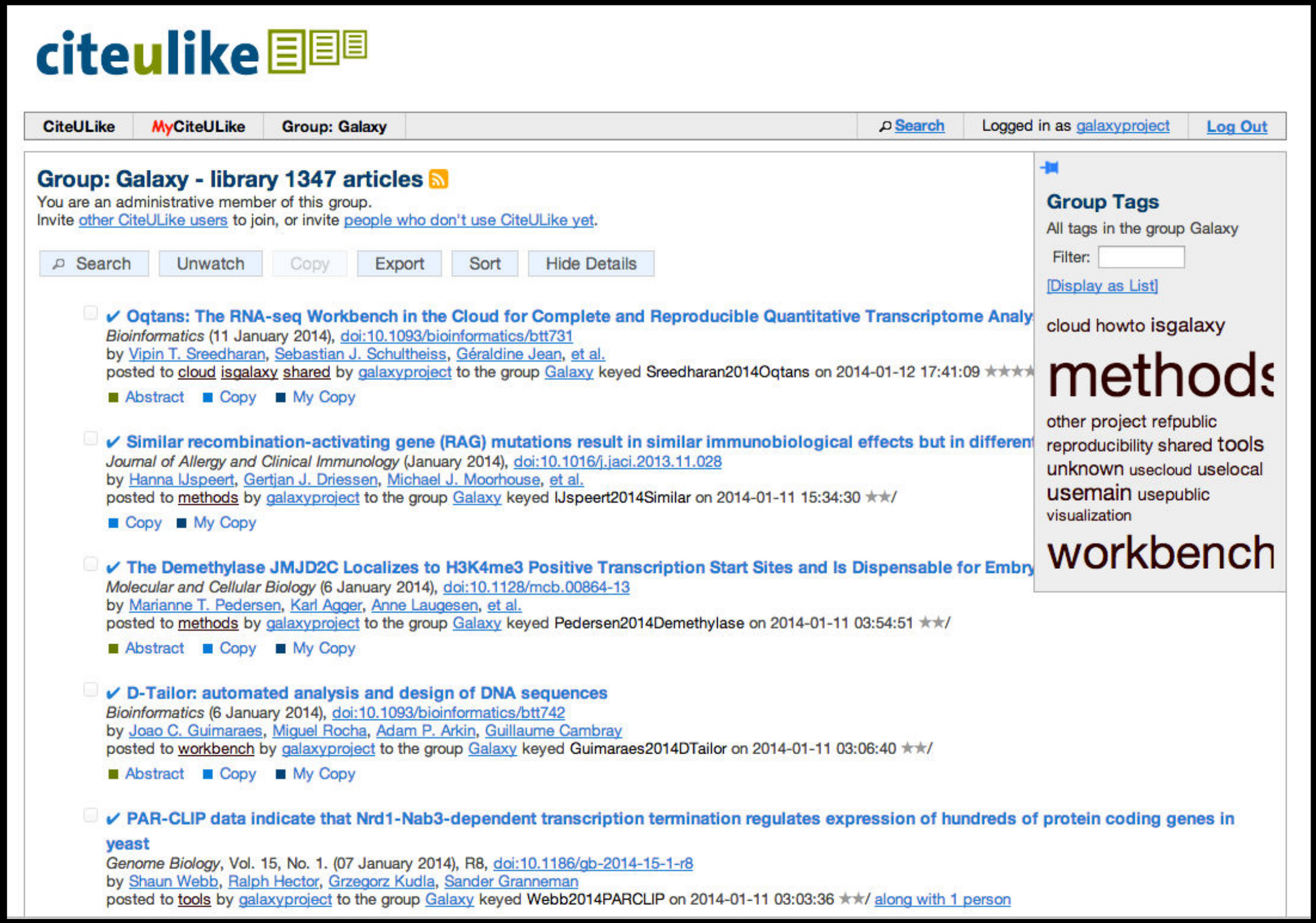


"How to" screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject

Galaxy Resources & Community: CiteULike Group



Over 1300 papers

17 different tags

http://bit.ly/gxycul

Thanks

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

Anushka Brownley
The BioTeam
http://bioteam.net