Introduction to Galaxy

The Genome Analysis Centre (TGAC)
Norwich, United Kingdom
9 May 2014

Dave Clements
Johns Hopkins University

http://galaxyproject.org/

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done

The Agenda

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.

While this workshop does cover RNA-Seq, we are only using that specific example to learn general principles.

What is Galaxy?

A free (for everyone) web server

Open source software

These options result in several ways to use Galaxy

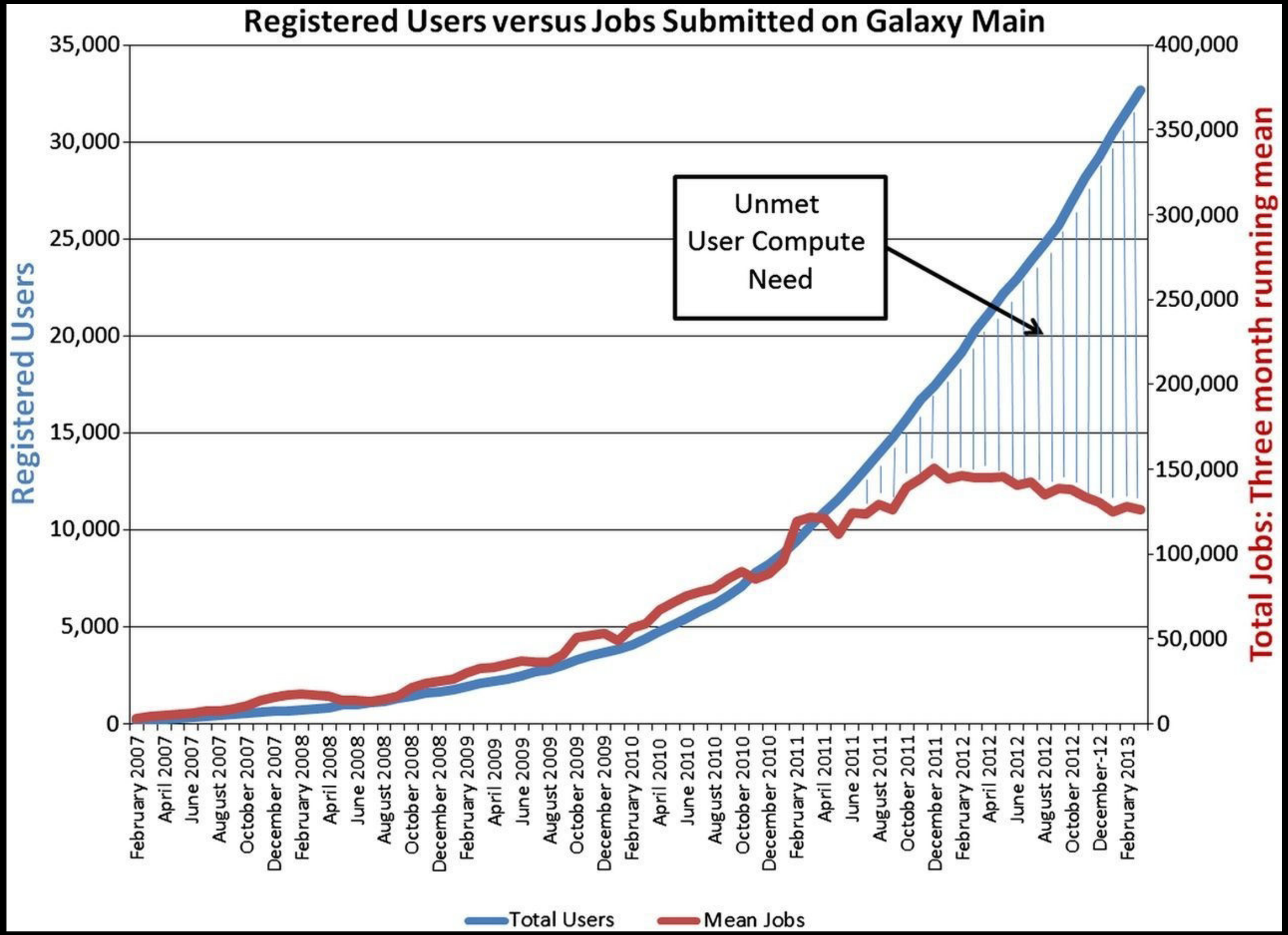
http://galaxyproject.org

Galaxy is available ...

As a free (for everyone) web server 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

It is installed in locations around the world, including:

http://galaxy.qbcg.uga.edu/

Galaxy is available ...

- As a free (for everyone) web ser
  http://usegalaxy.org
- As open source software
  http://getgalaxy.org
- *On the Cloud*

We are using this today.

http://aws.amazon.com/education
http://globus.org/

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, 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

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
11:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done

Basic Analysis

Which genes have most overlapping Repeats?

http://cloud1.galaxyproject.org/

http://cloud2.galaxyproject.org/

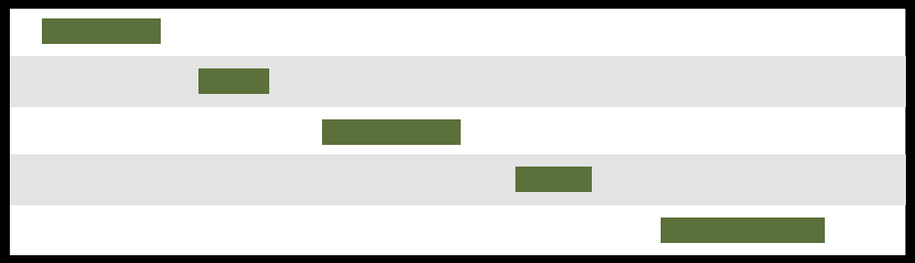
(~ http://usegalaxy.org/galaxy101)

Genes & Repeats: A General Plan

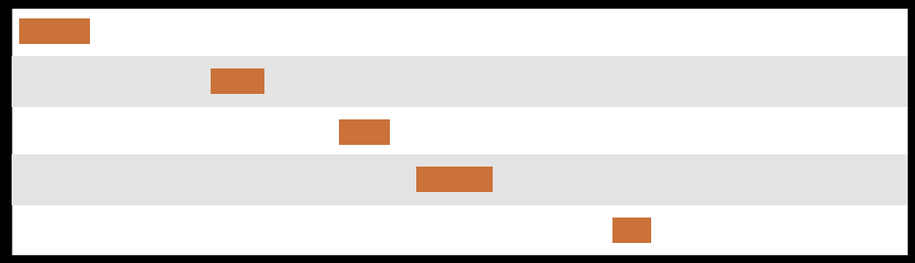
- Get some data
  - Get Data → UCSC Table Browser
- Identify which genes/exons have Repeats
- Count Repeats per exon
- Visualize, save, download, ... exons with most Repeats

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

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

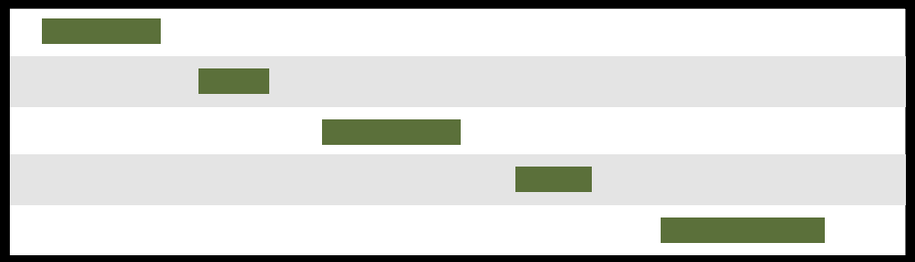


Exons

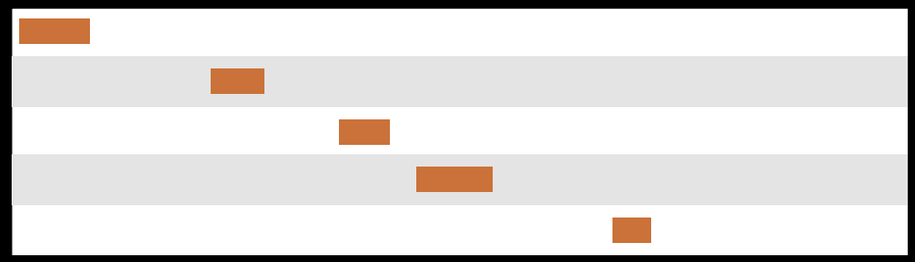


Repeats

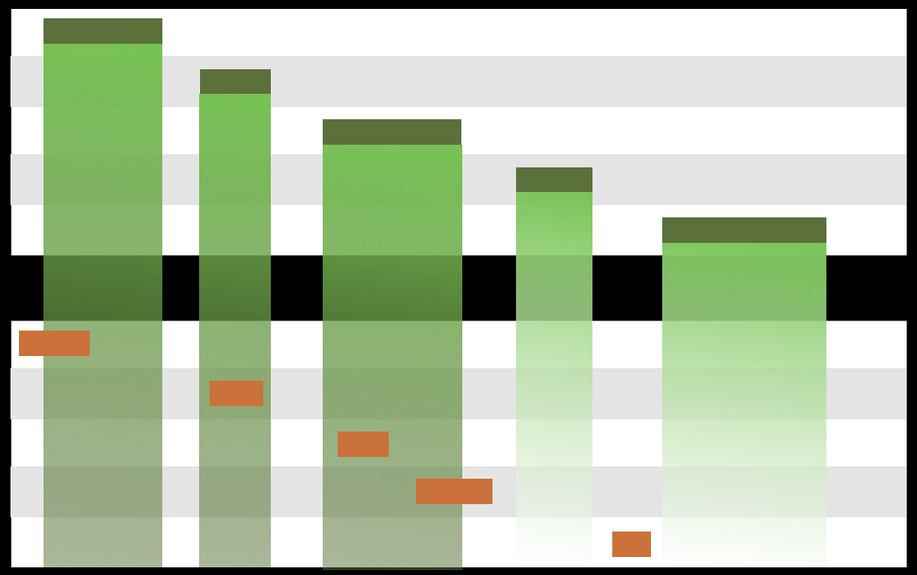
(Identify which genes/exons have Repeats)



Exons



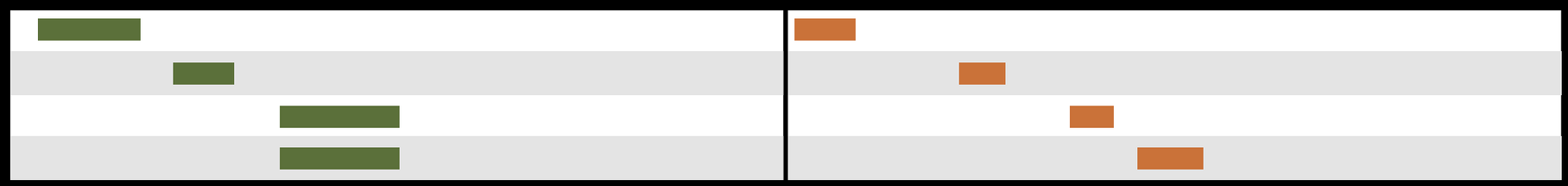
Repeats



Exons

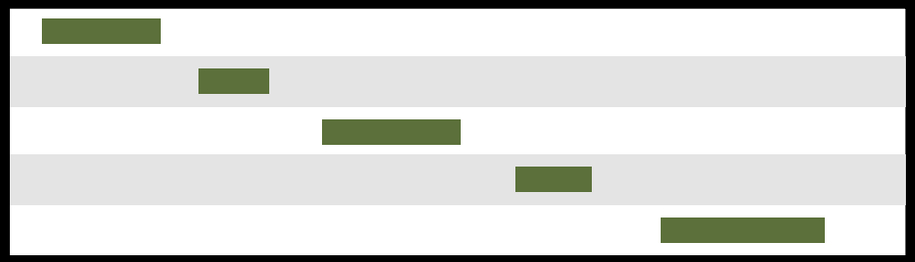
Repeats

Overlap pairings

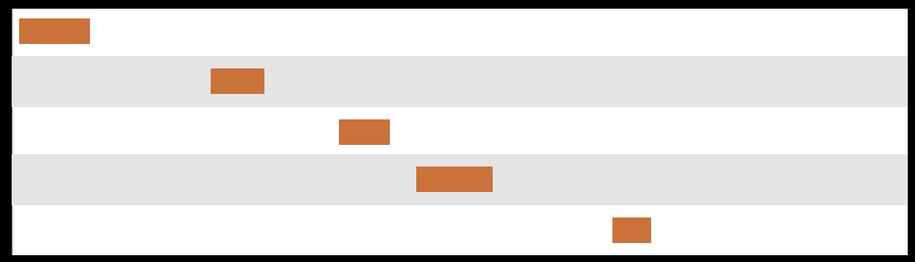


Operate on Genomic Intervals → **Join**

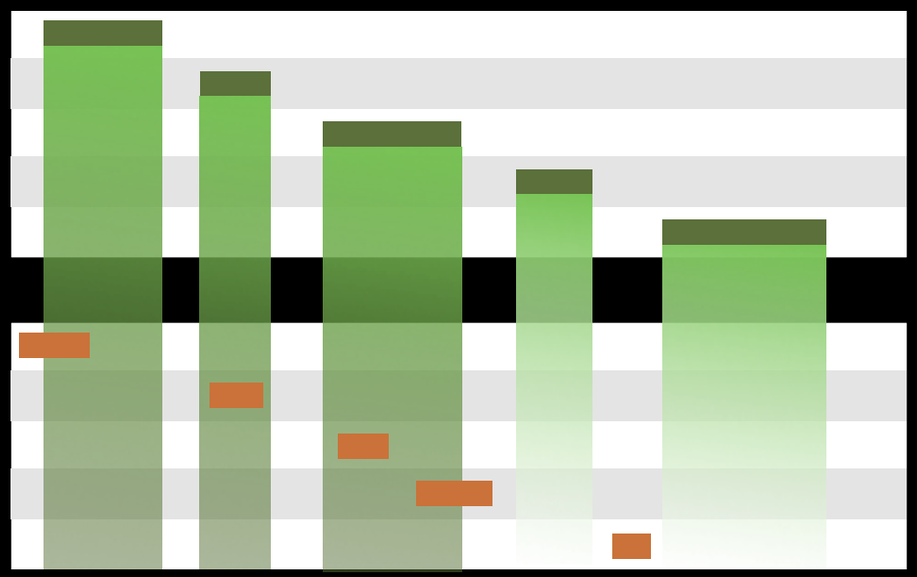
(Identify which genes/exons have Repeats)



Exons



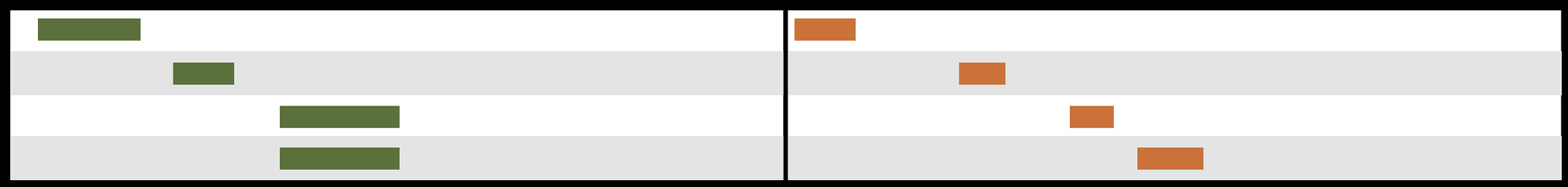
Repeats



Exons

Repeats

Overlap pairings

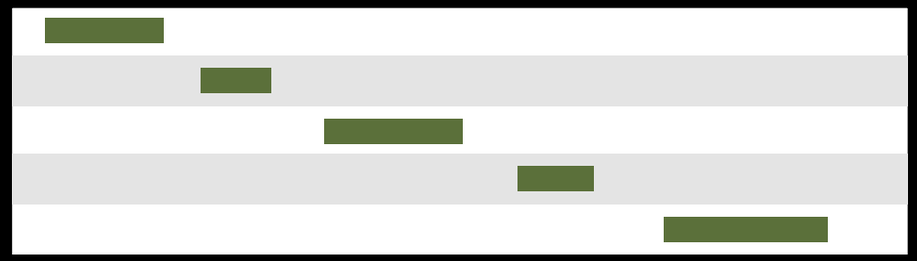


Exon overlap counts

Join, Subtract, and Group → Group

(Count Repeats per exon)

Exon overlap counts

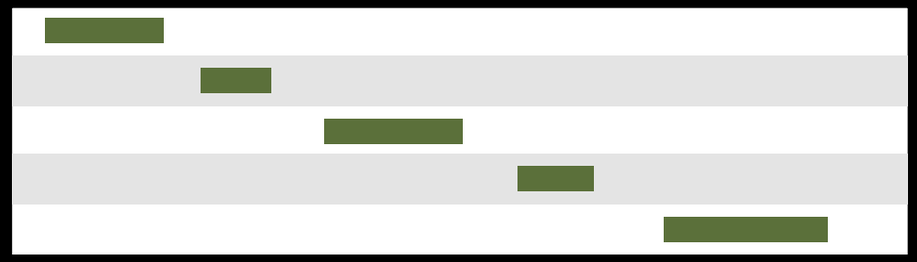


Exons

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

Incorporate the overlap count with rest of Exon information

Exon overlap counts



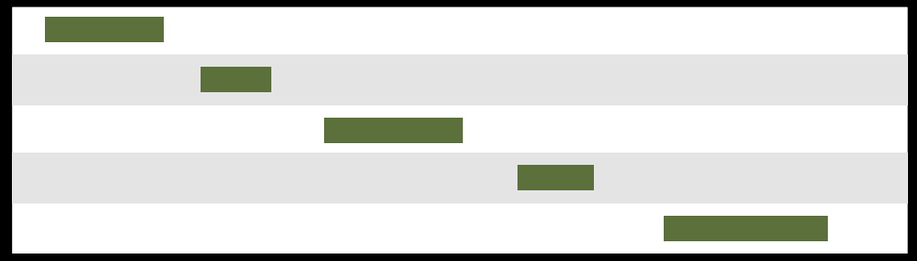
Exons

Join on exon name

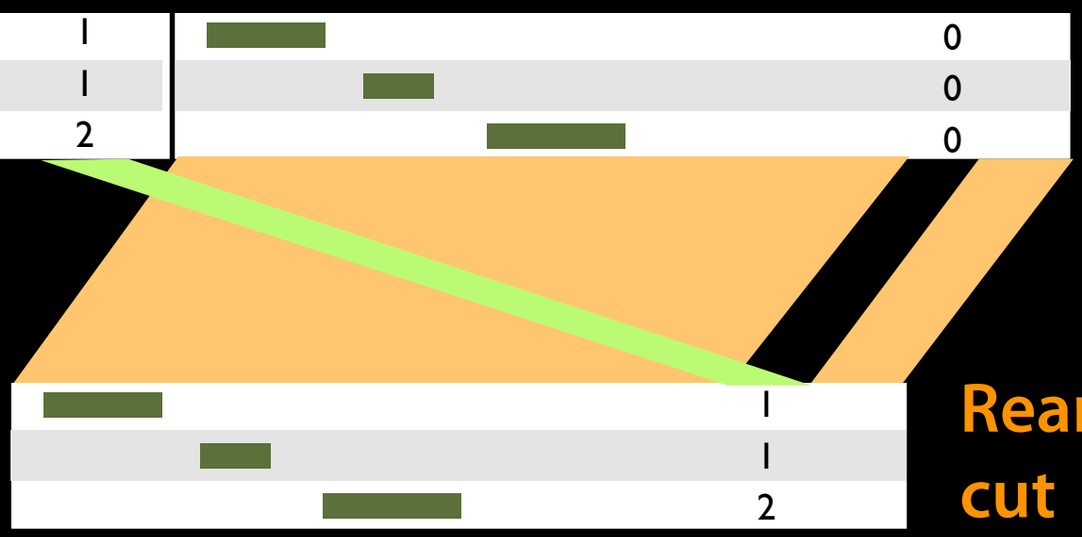
Join, Subtract, and Group → Join

(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)

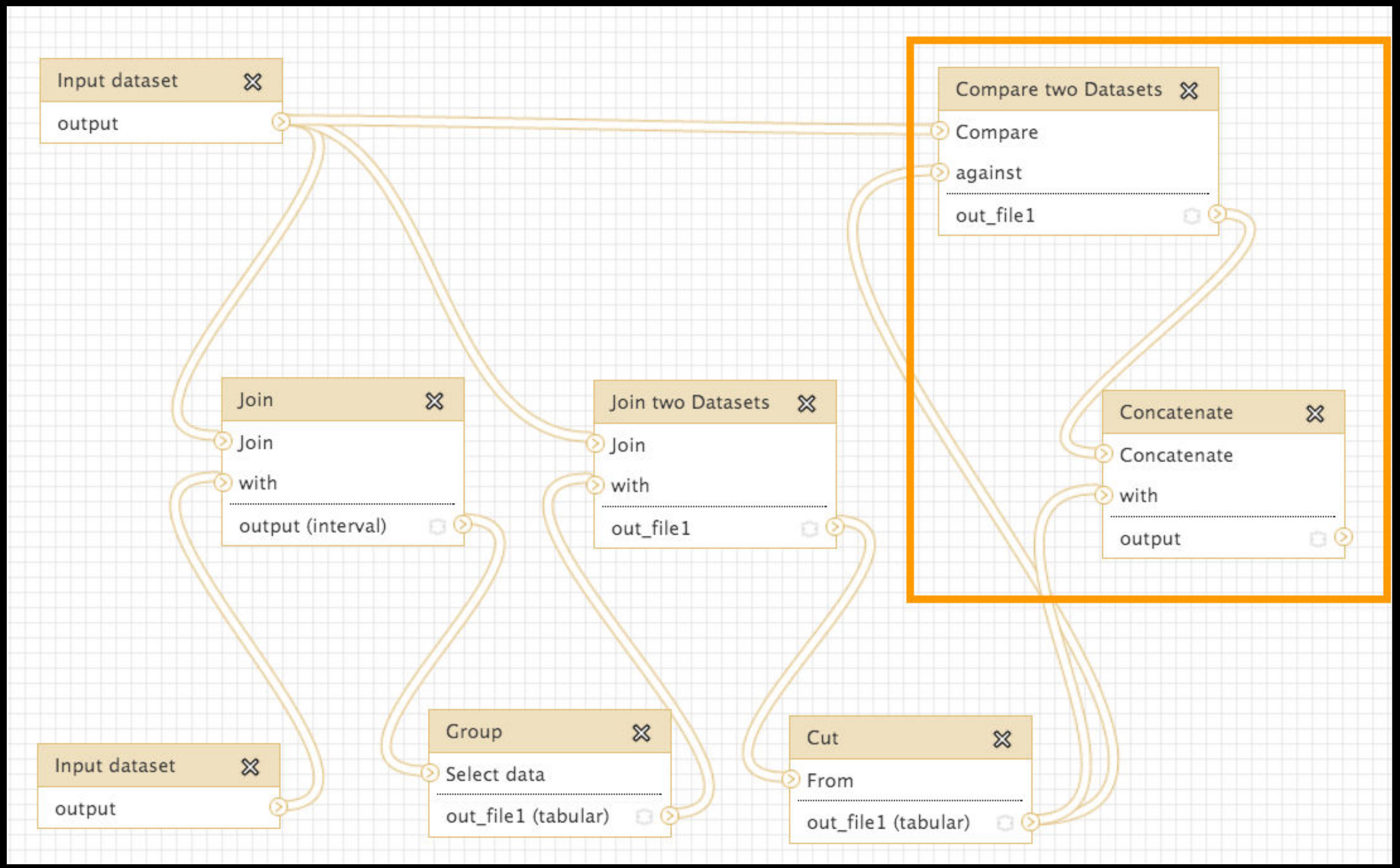
Genes & Repeats: Exercise

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

Basic Analysis: Further reading & Resources

http://usegalaxy.org/galaxy101

https://vimeo.com/76343659

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 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 Repeats *History* → Reusable *Workflow?*

- The analysis we just finished was about
  - Human chr22
  - Overlap between exons and Repeats
- But, ...
  - there is nothing inherent in the analysis about humans, exons or repeats
  - 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.

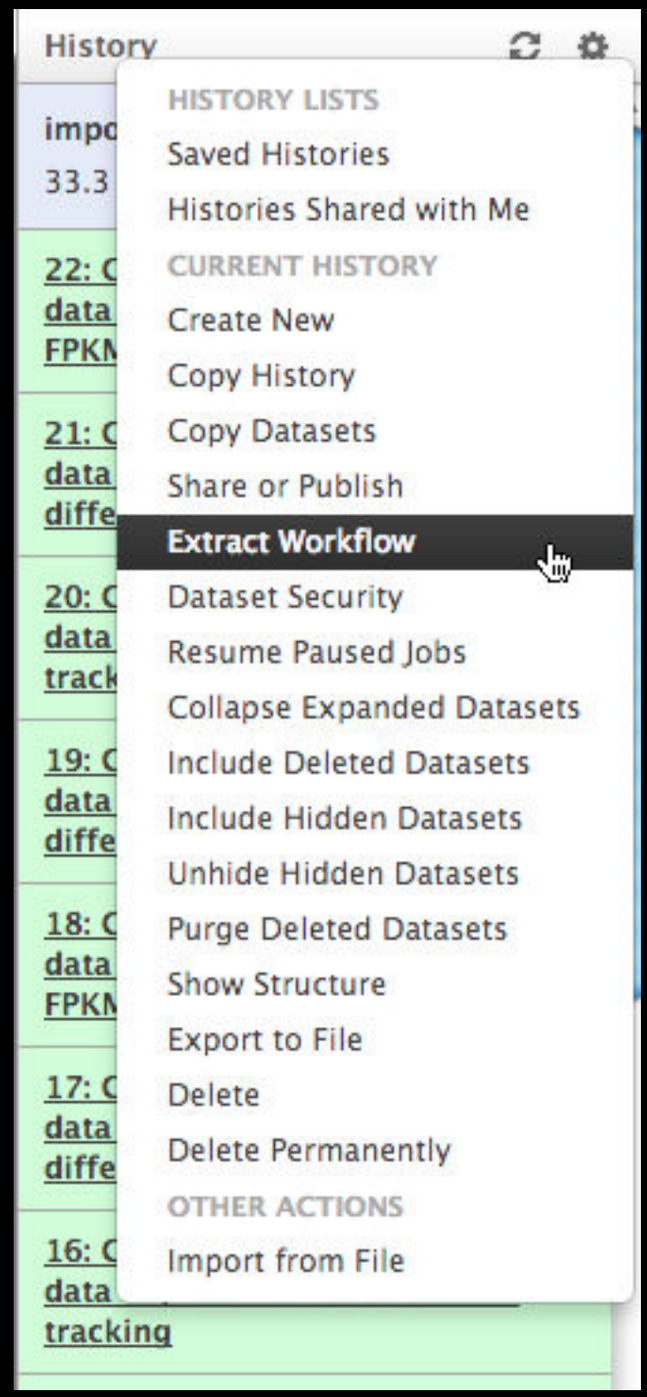
Create a Workflow from a History

Extract Workflow from history

Create a workflow from this history.

Edit it to make some things clearer.

(cog) → Extract Workflow



Run / test it

Guided: rerun with same inputs

Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 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
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  ............................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./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 Genome Center Bioinformatics Core

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

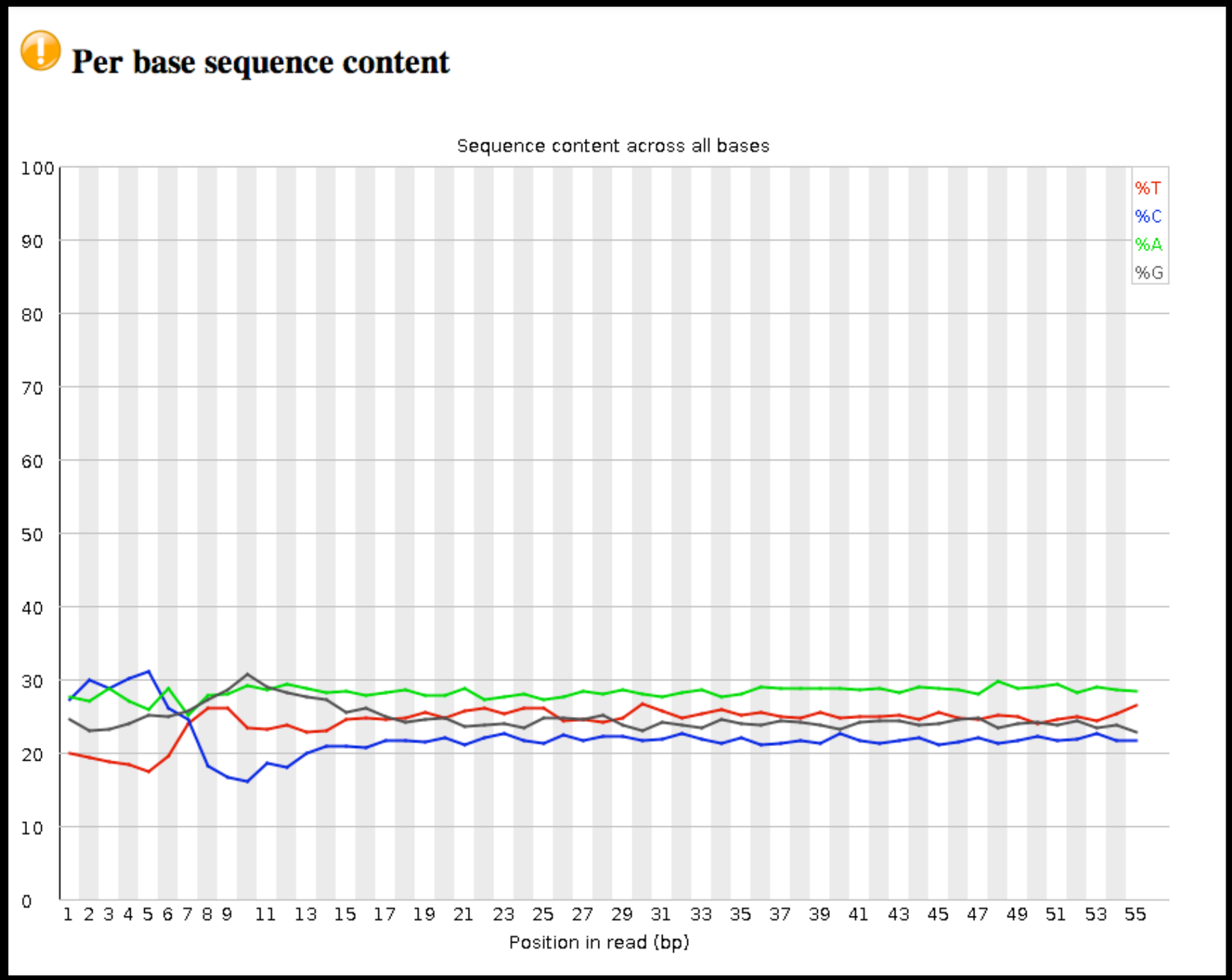
NGS Data Quality: Assessment tools

NGS QC and Manipulation → FastQC

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

http://bit.ly/FastQCBoxPlot

NGS Data Quality: Sequence bias at front of reads?

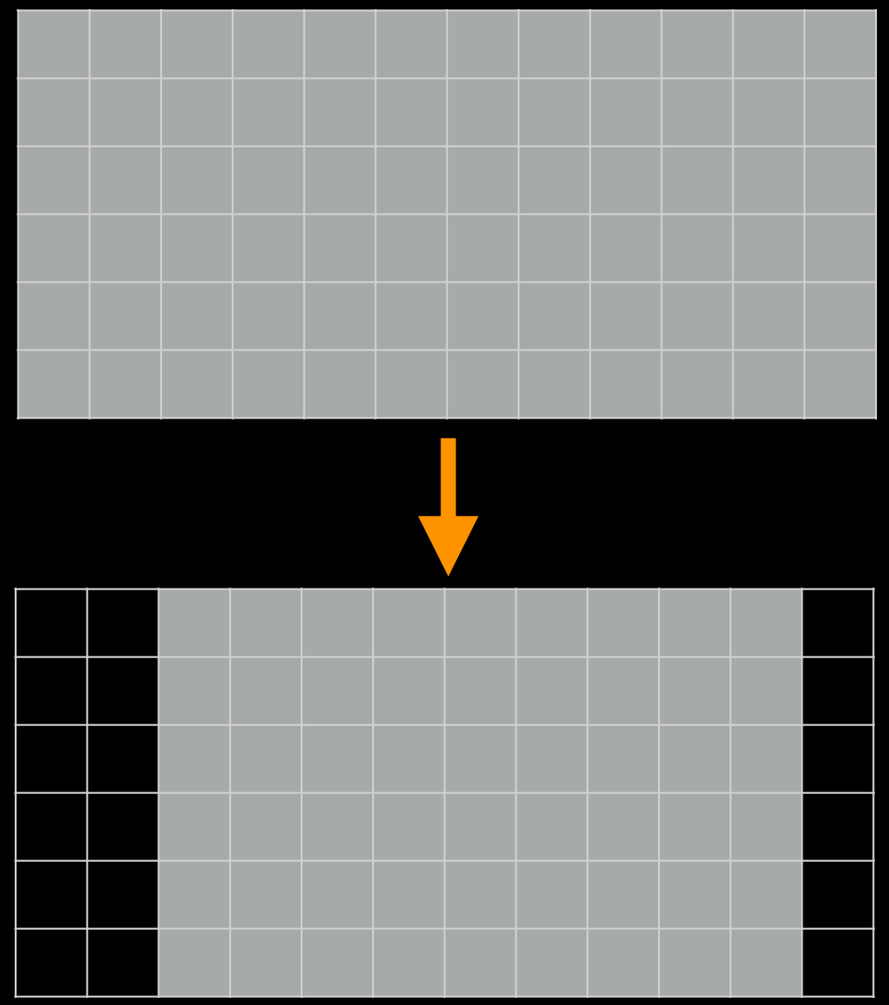


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

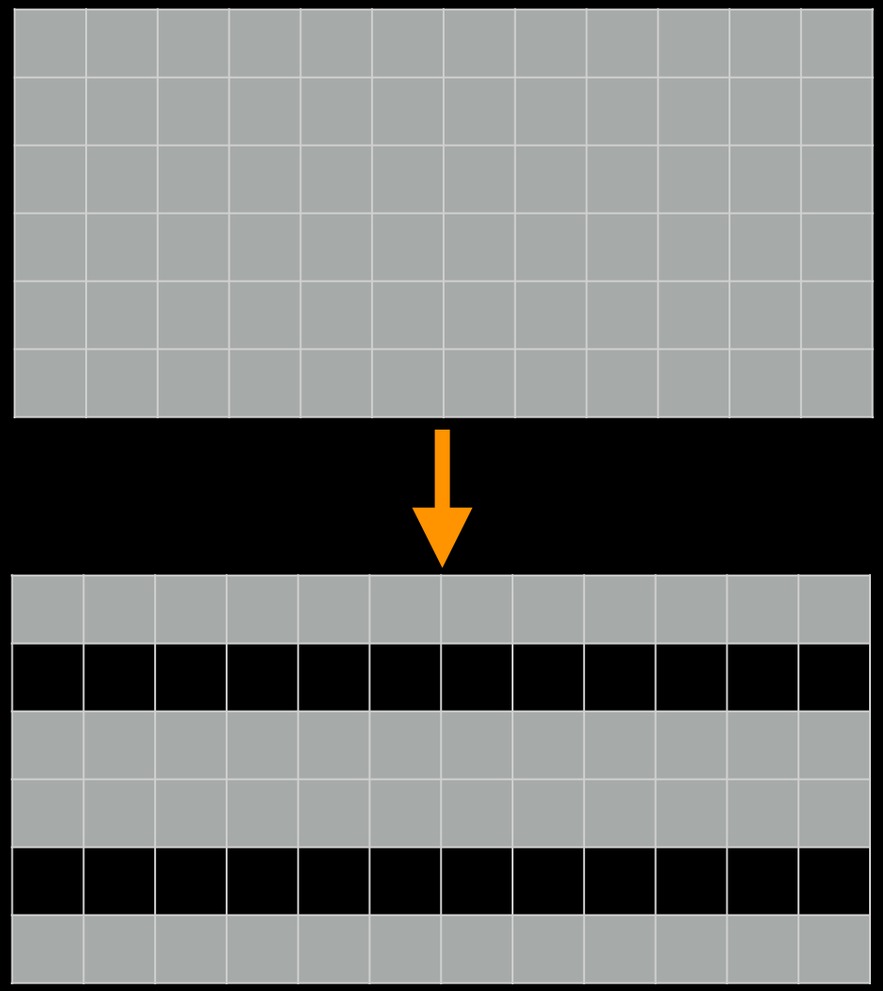
NGS Data Quality: Trim as we see fit

- 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



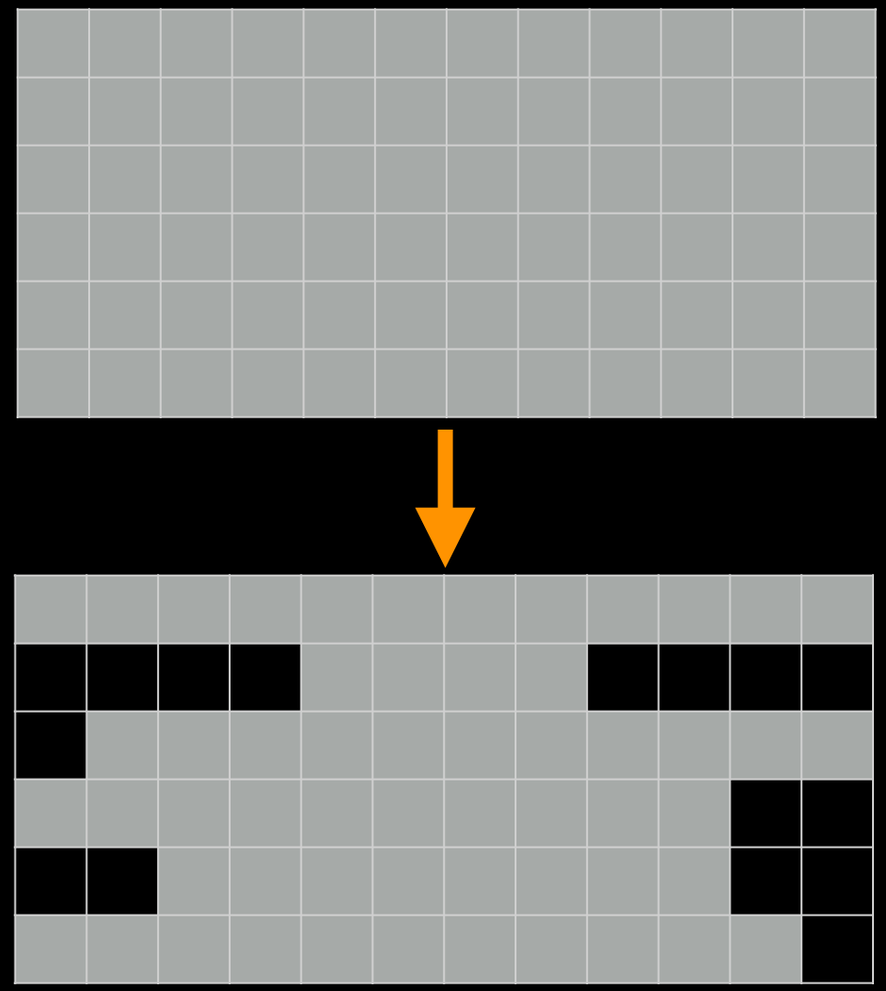
NGS Data Quality: Base Quality Trimming

- ~~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.

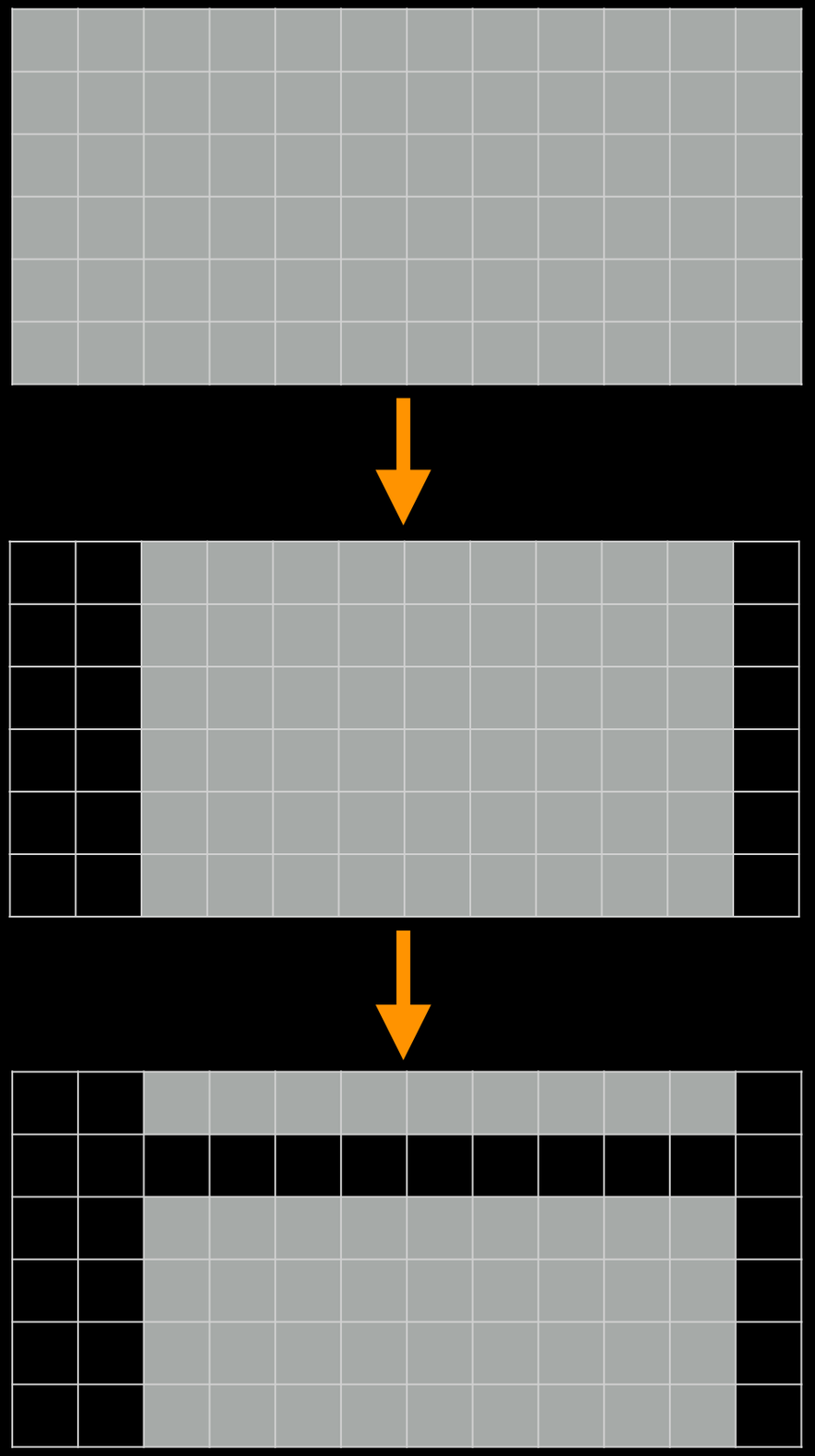


NGS Data Quality: Base Quality Trimming

- 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)

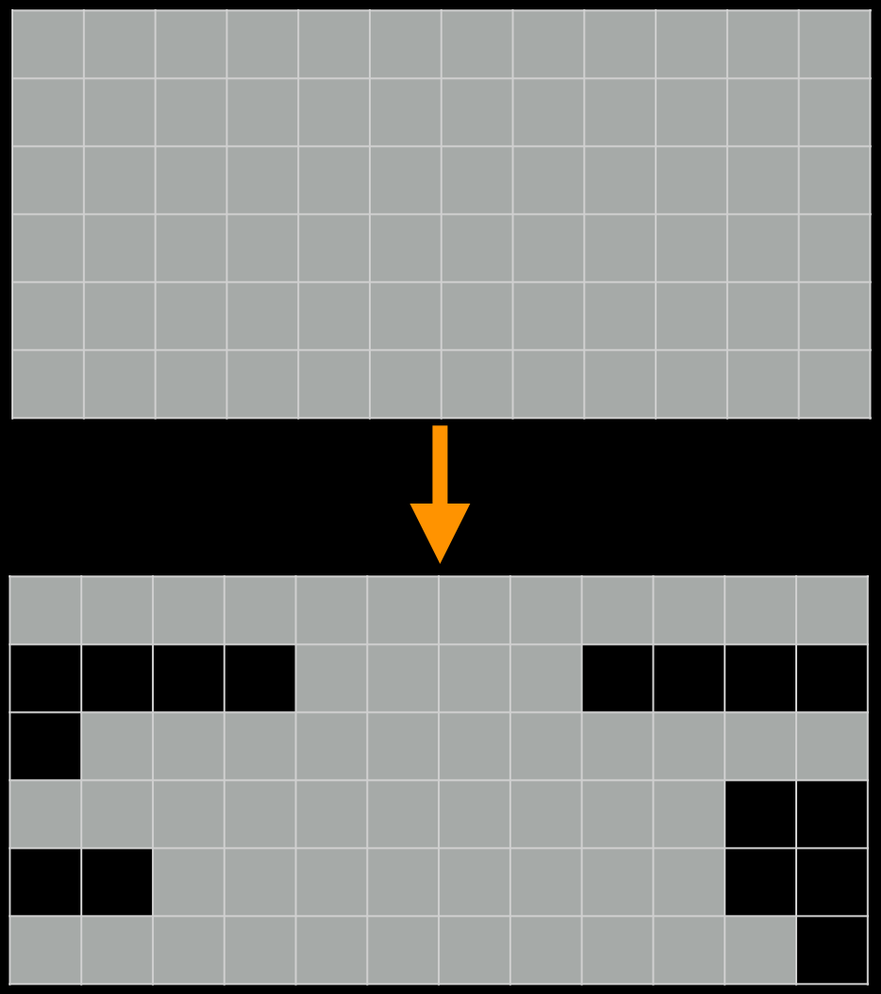
Trim? *As we see fit?*

- 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."

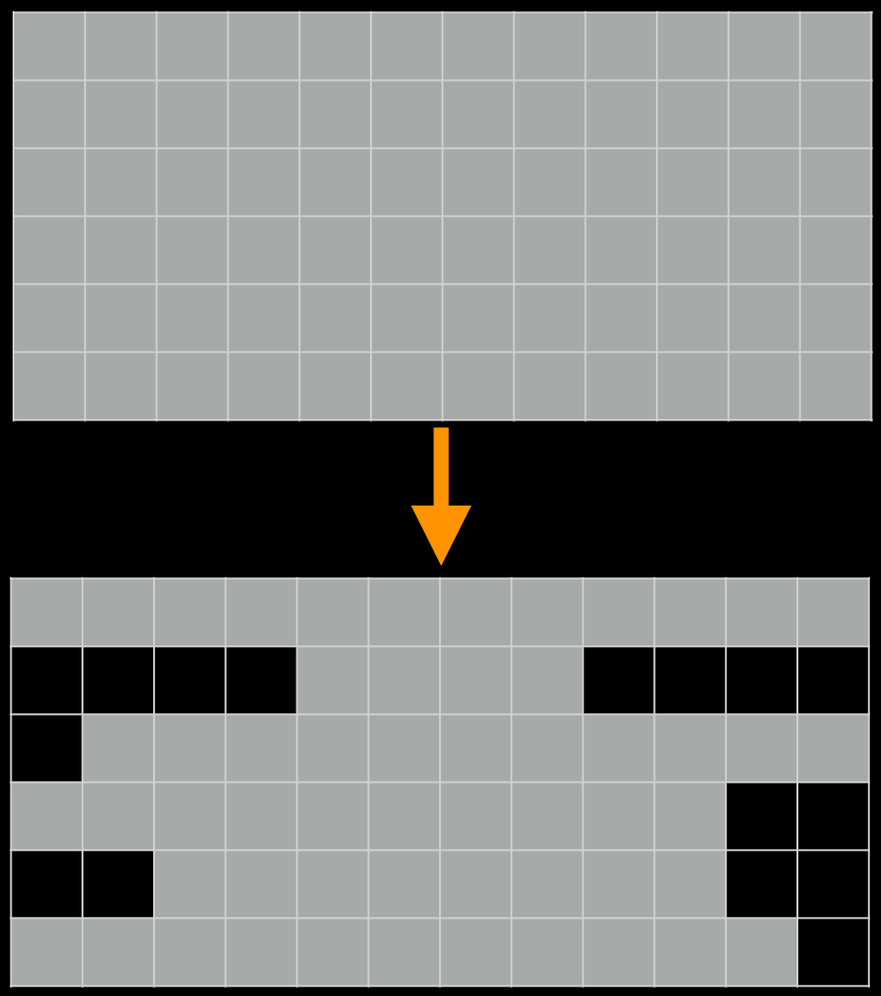
Most of us

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 ...):*

- 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 3 (or 5) more to go!

Workflows:

Create a QC workflow that does all these steps

Or, cheat and import the shared workflow.

Or, really cheat and just import the already trimmed datasets from the shared data library

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*.

RNA-seq Exercise: Mapping with Tophat

Cheat Alert!

We are going to talk about Tophat but we aren't going to run it today:

1. It takes a lot of time to run
2. Tophat2 has issues on these instances

Therefore we will talk about Tophat, and then use results of Tophat run that was run before the workshop

RNA-seq Exercise: Mapping with Tophat

Create a new history

Import all datasets from library:

RNA-Seq Example → Mapped Reads

and genes_chr12.gtf

RNA-seq Exercise: Mapping with Tophat

Yes, but how *might* we run 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 end reads

- Determined by sample prep
- 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*

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

Or not.

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

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done

Galaxy Community Resources: Galaxy Biostar

Tens of thousands of users leads to a lot of questions.

Absolutely have to encourage community support.

Project traditionally uses mailing list

Just moved the user support list to Galaxy Biostar, an online forum, that uses the Biostar platform

https://biostar.usegalaxy.org/

Galaxy Community Resources: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Dev

Questions about developing for and deploying Galaxy
High volume (5200 posts in 2013, 900+ members)

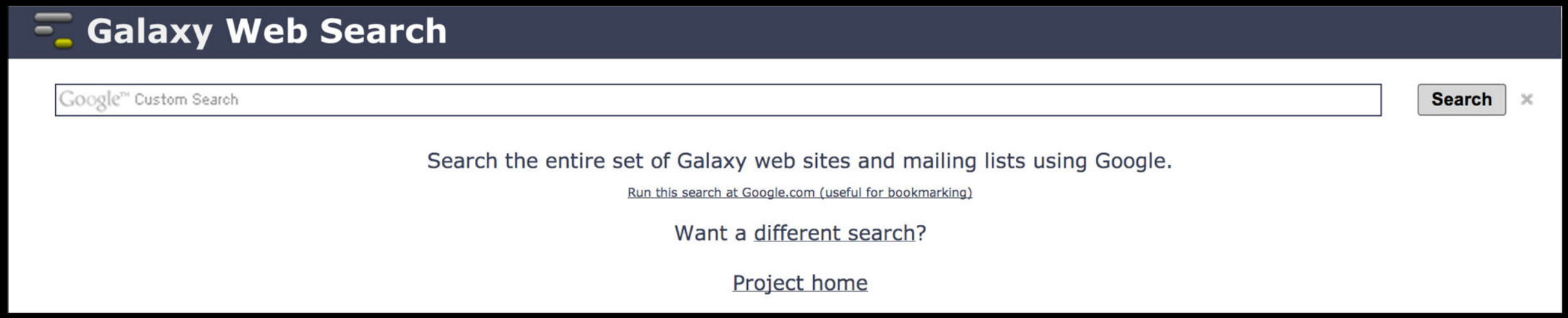
Galaxy-Announce

Project announcements, low volume, moderated
Low volume (47 posts in 2013, 3400+ members)

Galaxy-User (deprecated)

Questions about using Galaxy and usegalaxy.org
High volume (1328 posts in 2013, 2600+ members)

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: Public Galaxy Instances

http://bit.ly/gxyServers

Interested in:

ChIP-chip and ChIP-seq?

✓ Cistrome, Nebula

Statistical Analysis?

✓ Genomic Hyperbrowser

Protein synthesis?

✓ GWIPS-viz

de novo assembly?

✓ CBIIT Galaxy

Reasoning with ontologies?

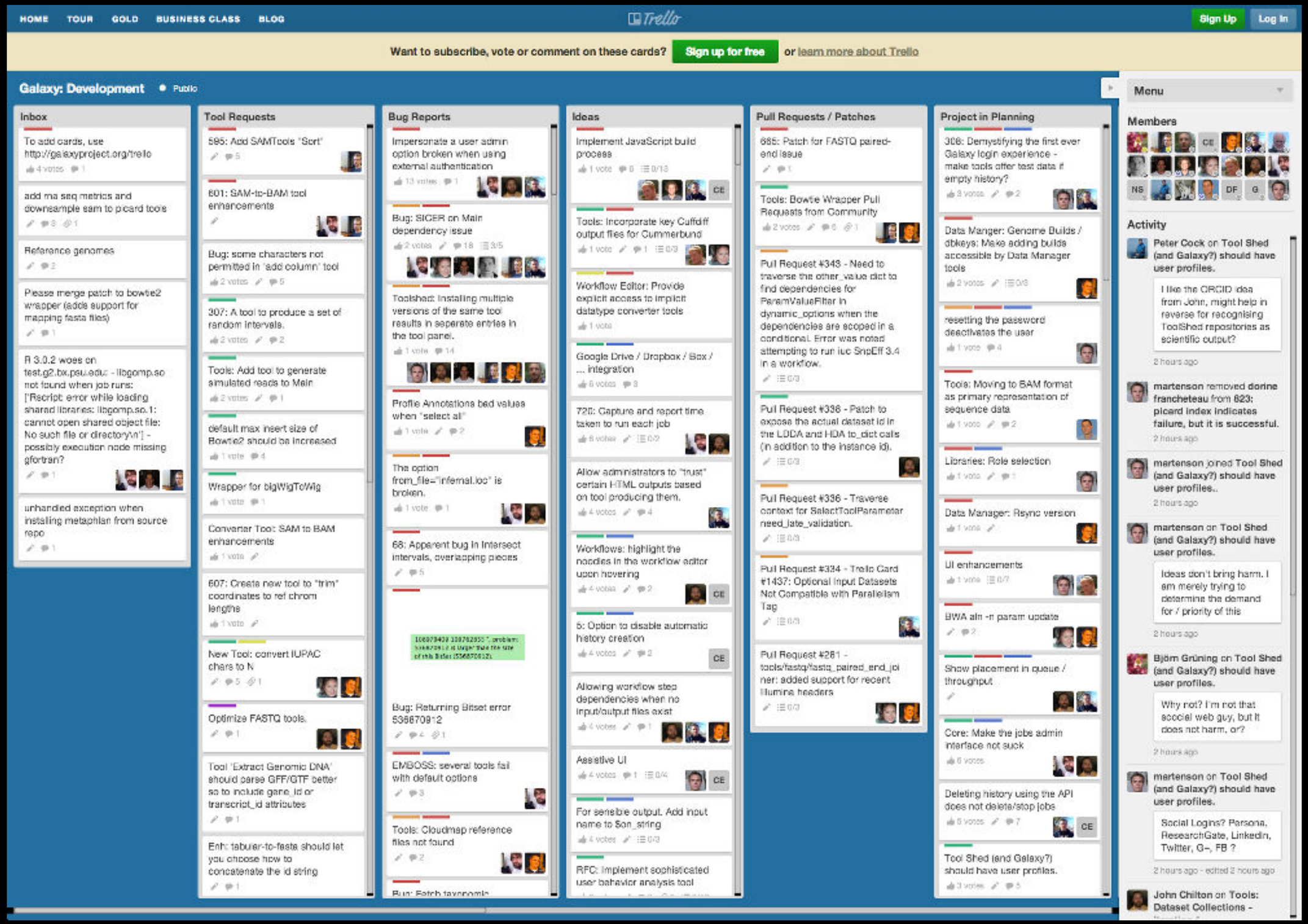
✓ GO Galaxy

Repeats!

✓ RepeatExplorer

Over 50 public Galaxy servers

Community can create, vote and comment on issues



http://bit.ly/gxytrello

http://wiki.galaxyproject.org

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
- 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!**

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

Early Registration & **Abstract Submission** are now open

Galaxy Australasia Workshop 2014

24-25 March, Melbourne

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
Teach • Support

Events

News

Galaxy Wiki — DaveClements Settings Logout | 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 a
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
send it to outreach@glaxyproject.org .

For events prior to this year, see the Events Archive.

Upcoming Events

Date	Topic/Event	Venue/Location
March 18	*Utilisation du Cloud pour la Biologie*	Institut de Biologie et C Protéines, CNRS-IBCP,
March 24-25	**Galaxy Australasia Workshop 2014 (GAW2014)**	Melbourne, Australia
March 26-30	*Galaxy toolset for Drosophila genomics* and one-on-one help in the Flybase Demonstrations Room	Drosophila Research Co San Diego, California,
April 15-17	*Biosciences/Genomics Program*	GlobusWorld, Chicago, United States
April 29 - May 1	*W1: Integrated Research Data Management for Next Gen Sequencing Analysis Using Galaxy and Globus Online Software-as-a-Service*	BioIT World
	W4: Analyzing NGS Data in Galaxy	
	W14: Running a Local Galaxy Instance	Adam Kraut, Nate Coraor, Anushka Brownley, Tristan Lubinski, James Reaney
	Globus Genomics: An End-to-End NGS Analysis Service on	

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

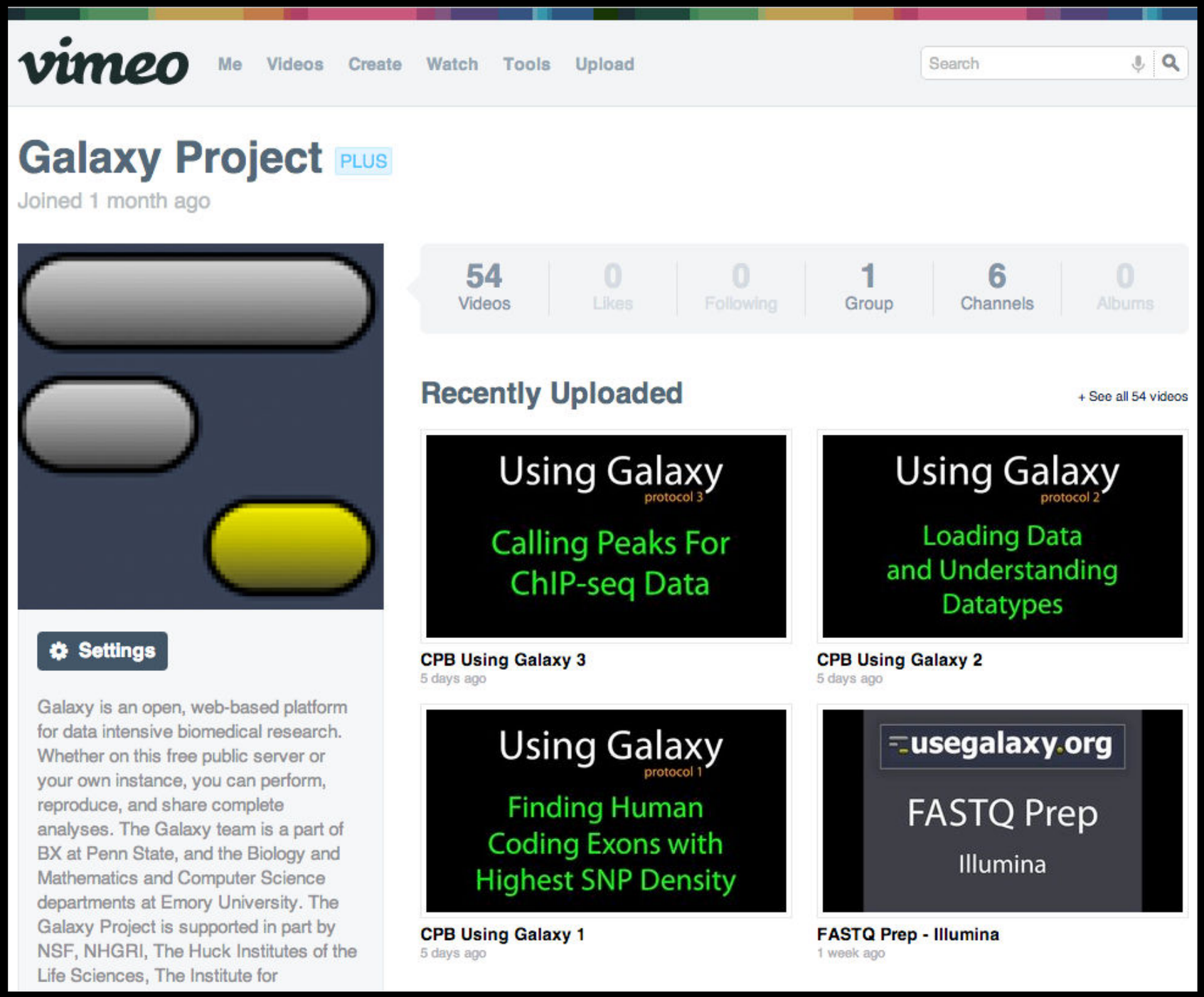
BALTIMORE, MD | JUNE 30 - JULY 2, 2014

http://bit.ly/gcc2014

Galaxy Australasia Workshop 2014

We also support community organized efforts and events.

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

citeulike

CiteULike | MyCiteULike | Group: Galaxy | Search | Logged in as galaxyproject | Log Out

Group: Galaxy - library 1437 articles

You are an administrative member of this group.
Invite other CiteULike users to join, or invite people who don't use CiteULike yet.

Search | Unwatch | Copy | Export | Sort | Hide Details

✔ **Life science data analysis workflow development using the bioextract server leveraging the iPlant collaborative cyberin**
Concurrency Computat.: Pract. Exper. (1 February 2014), pp. n/a-n/a, doi:10.1002/cpe.3237
by Carol M. Lushbough, Etienne Z. Gnimpieba, Rion Dooley
posted to workbench by galaxyproject to the group Galaxy keyed Lushbough2014Life on 2014-03-04 19:10:09 ★★/
■ Abstract ■ Copy ■ My Copy

✔ **Workshops: A Great Way to Enhance and Supplement a Degree**
PLoS Comput Biol, Vol. 10, No. 2. (27 February 2014), e1003497, doi:10.1371/journal.pcbi.1003497
by Segun Fatumo, Sayane Shome, Geoff Macintyre
posted to other by galaxyproject to the group Galaxy keyed Fatumo2014Workshops on 2014-03-04 19:08:20 ★★/
■ Abstract ■ Copy ■ My Copy

✔ **Wrangling Galaxy's Reference Data**
Bioinformatics (28 February 2014), doi:10.1093/bioinformatics/btu119
by Daniel Blankenberg, James E. Johnson, James Taylor, Anton Nekrutenko
posted to project by galaxyproject to the group Galaxy keyed Blankenberg2014Wrangling on 2014-03-04 18:55:14 ★★★★★/
■ Abstract ■ Copy ■ My Copy

✔ **Detection of PIWI and piRNAs in the mitochondria of mammalian cancer cells**
Biochemical and Biophysical Research Communications (March 2014), doi:10.1016/j.bbrc.2014.02.112
by ChangHyuk Kwon, Hyosun Tak, Mina Rho, et al.
posted to methods by galaxyproject to the group Galaxy keyed Kwon2014Detection on 2014-03-04 18:53:07 ★★/ along with 1 person
■ Copy ■ My Copy

✔ **CanSNPer: a hierarchical genotype classifier of clonal pathogens**
Bioinformatics (25 February 2014), doi:10.1093/bioinformatics/btu113
by Adrian Lärkeryd, Kerstin Myrtennäs, Edvin Karlsson, et al.
posted to tools by galaxyproject to the group Galaxy keyed Larkeryd2014CanSNPer on 2014-03-04 18:51:21 ★★/
■ Abstract ■ Copy ■ My Copy

✔ **Web-based Workflow Planning Platform Supporting the Design and Execution of Complex Multiscale Cancer Models**
pp. 1-1, doi:10.1109/jbhi.2013.2297167

Group Tags

All tags in the group Galaxy

Filter:

[Display as Cloud]

methods	697
workbench	466
usemain	108
tools	91
isgalaxy	80
cloud	50
shared	50
unknown	47
uselocal	37
project	32
howto	30
reproducibility	28
other	23
usepublic	19
refpublic	12
visualization	7
usecloud	3

Over 1500 papers

17 tags

http://bit.ly/gxycul

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done

Cuffdiff?

- Part of the Tuxedo RNA-Seq Suite (as are Tophat and Bowtie)
- Widely used and widely installed on Galaxy instances

NGS: RNA Analysis → Cuffdiff

Cuffdiff?

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.

Cuffdiff Alternatives

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.

Cuffdiff Alternatives: DESeq

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

Cuffdiff Alternatives: DESeq

NGS: SAM Tools → htseq-count
once for each BAM file

Join the 4 (or 6) HTSeq datasets together on gene name

Cut out the duplicate gene name columns

NGS: RNA Analysis → DE Seq

Cuffdiff Alternatives: DESeq

DESeq output is a list of genes,
sorted by adjusted P value,
with lowest P values listed first

How many genes have an adjusted P value < 0.05 ?

Differential Expression: 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

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 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

Let's all share...

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

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)

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

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 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.**

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

http://aws.amazon.com/education

Could follow the step by step instructions on the wiki, but ... AWS just revamped it's interface.

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.

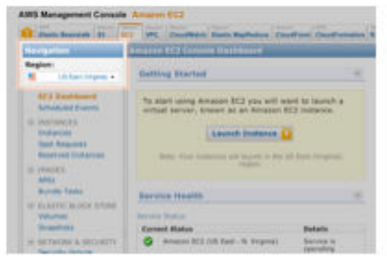
AWS
Get Started
Capacity Planning
AMIs
↑ CloudMan

Contents

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

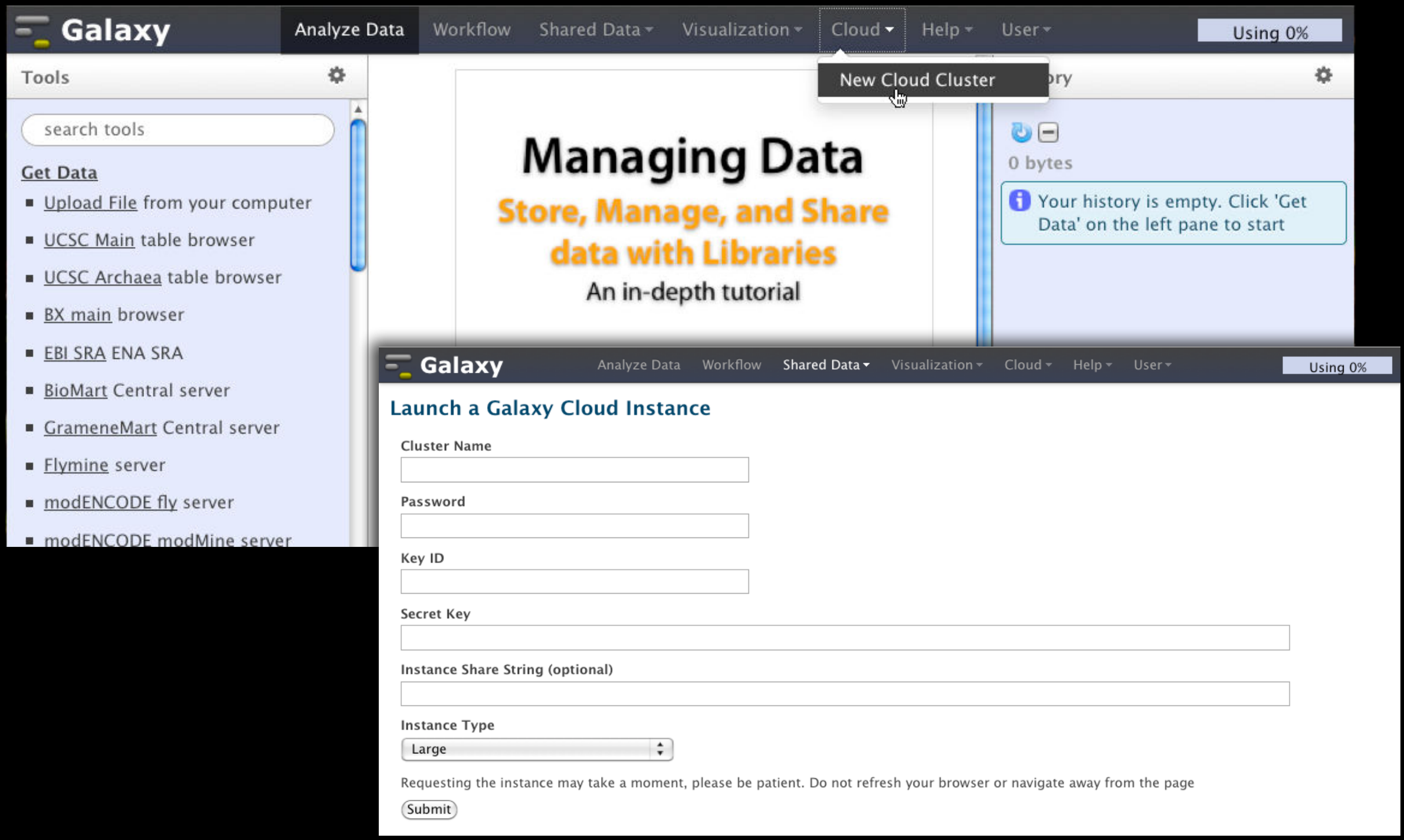


1.2. Set region

AWS Credentials

Instant CloudMan

http://usegalaxy.org/cloudlaunch



The Agenda

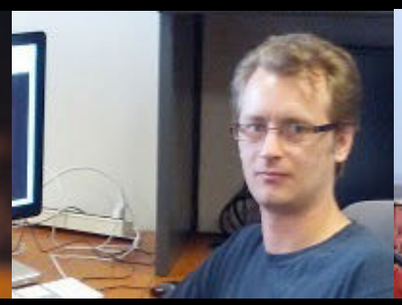
8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done (almost)

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

Also Thanks To

TGAC

The Genome Analysis Centre™

Matt Drew

Vicky Schneider-Gricar

Helen Tunney

THE UNIVERSITY of EDINBURGH

igmm

INSTITUTE OF GENETICS & MOLECULAR MEDICINE

edinburgh genomics

design execution analysis

ROSLIN

Wellcome Trust Centre for Cell Biology WTCCB

Edinburgh Bioinformatics

Feedback

We need it!

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org

The Agenda

8:30 Registration
9:00 Welcome
9:20 Basic Analysis with Galaxy
10:20 Basic Analysis into Reusable Workflows
10:40 Break
11:00 RNA-Seq Example Part I
12:00 Galaxy Community Resources
12:20 Lunch
13:05 RNA-Seq Example Part II
13:55 Sharing, Publishing and Reproducibility
14:15 Break
14:35 Setting up your own Galaxy Cluster on AWS
16:00 Done (really)