

Introduction to Galaxy

The University of Edinburgh
Edinburgh, United Kingdom
12 May 2014

Dave Clements
Johns Hopkins University

Shaun Webb
Bert Overduin
University of Edinburgh

<http://galaxyproject.org/>



THE UNIVERSITY
of EDINBURGH



Edinburgh Bioinformatics



The Agenda

9:00 Welcome, Galaxy Platforms

9:20 Basic Analysis with Galaxy

10:30 Break

10:45 RNA-Seq Example

12:40 Project and Community Overview

13:00 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 data integration and analysis platform

A free (for everyone) web server

Open source software

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

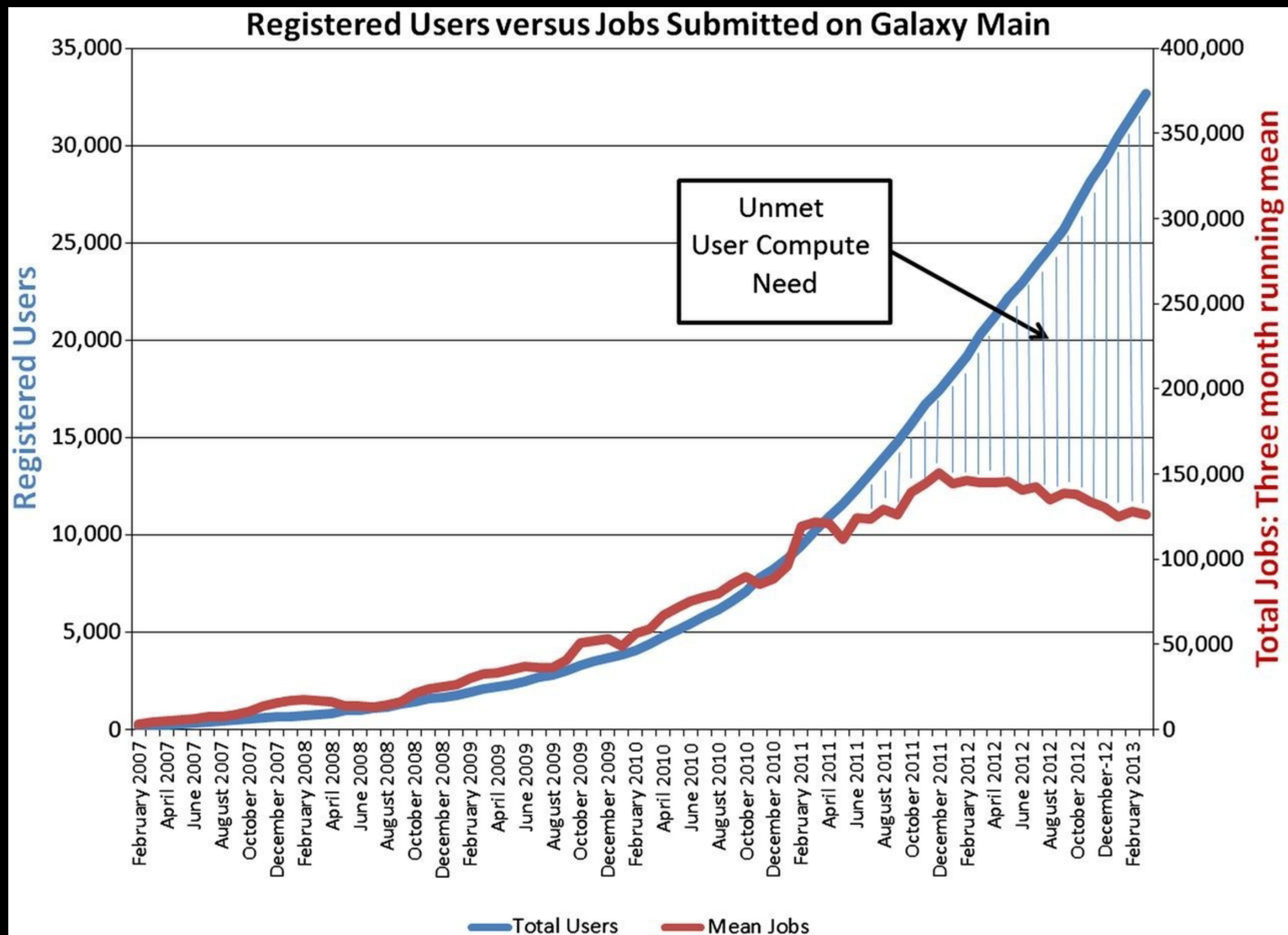
<http://galaxyproject.org>

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

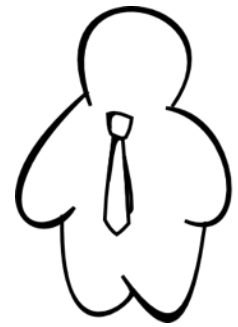
<http://usegalaxy.org>

- As open source software

<http://getgalaxy.org>

It is installed in locations around the world

Galaxy is available **on the Cloud**



CloudMan

We are using this today.

Start with a **fully configured & populated** (tools and data) Galaxy instance.

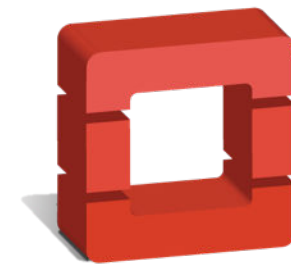
Allows you to scale up and down your compute assets as needed.

Someone else manages the data center.

<http://aws.amazon.com/education>

<http://globus.org/>

<http://wiki.galaxyproject.org/Cloud>



openstackTM
CLOUD SOFTWARE

OpenNebula.org
The Open Source Toolkit for Cloud Computing



globus
genomics

Instant CloudMan

<http://usegalaxy.org/cloudlaunch>

The image shows two overlapping screenshots of the Galaxy web interface. The top screenshot displays the main Galaxy dashboard with the 'Cloud' menu open, showing the 'New Cloud Cluster' option. The bottom screenshot shows the 'Launch a Galaxy Cloud Instance' form, which includes fields for Cluster Name, Password, Key ID, Secret Key, and Instance Share String (optional), along with an 'Instance Type' dropdown set to 'Large' and a 'Submit' button.

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

Tools

search tools

Get Data

- [Upload File](#) from your computer
- [UCSC Main](#) table browser
- [UCSC Archaea](#) table browser
- [BX main](#) browser
- [EBI SRA](#) ENA SRA
- [BioMart](#) Central server
- [GrameneMart](#) Central server
- [Flymine](#) server
- [modENCODE fly](#) server
- [modENCODE modMine](#) server

Managing Data
Store, Manage, and Share data with Libraries
An in-depth tutorial

0 bytes

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

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

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

Galaxy is available: **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

9:00 Welcome, Galaxy Platforms

9:20 Basic Analysis with Galaxy

10:30 Break

10:45 RNA-Seq Example

12:40 Project and Community Overview

13:00 Done

Basic Analysis

Which genes have most overlapping
Repeats?

<http://cloud1.galaxyproject.org/>

<http://cloud2.galaxyproject.org/>

<http://cloud3.galaxyproject.org/>

<http://cloud4.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

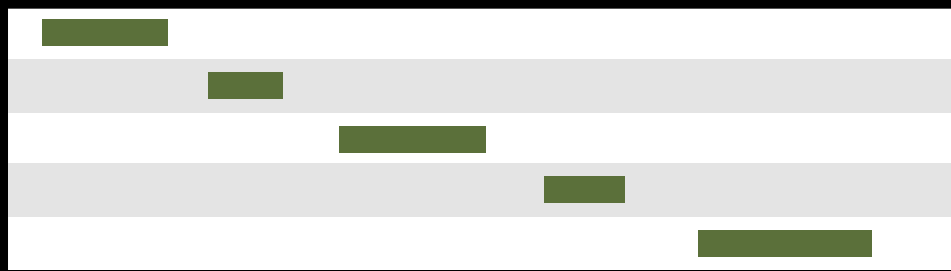
<http://cloud1.galaxyproject.org/>

<http://cloud2.galaxyproject.org/>

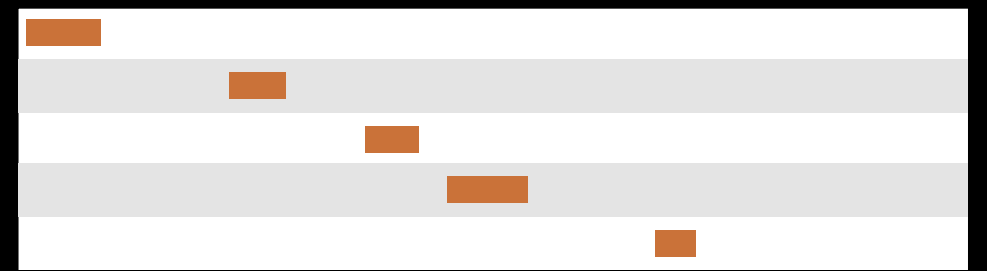
<http://cloud3.galaxyproject.org/>

<http://cloud4.galaxyproject.org/>

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

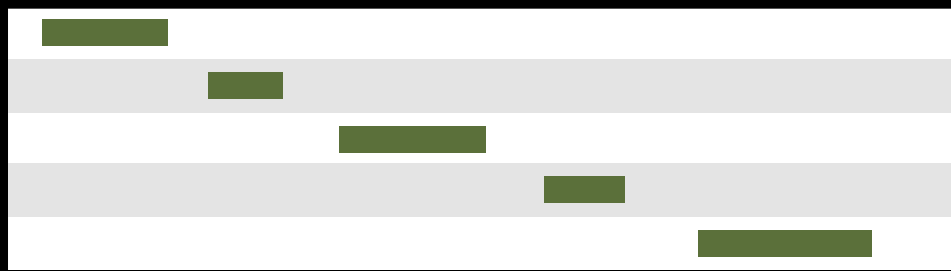


Exons

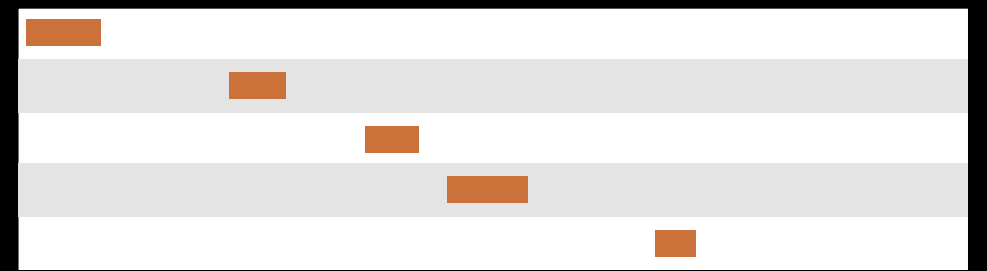


Repeats

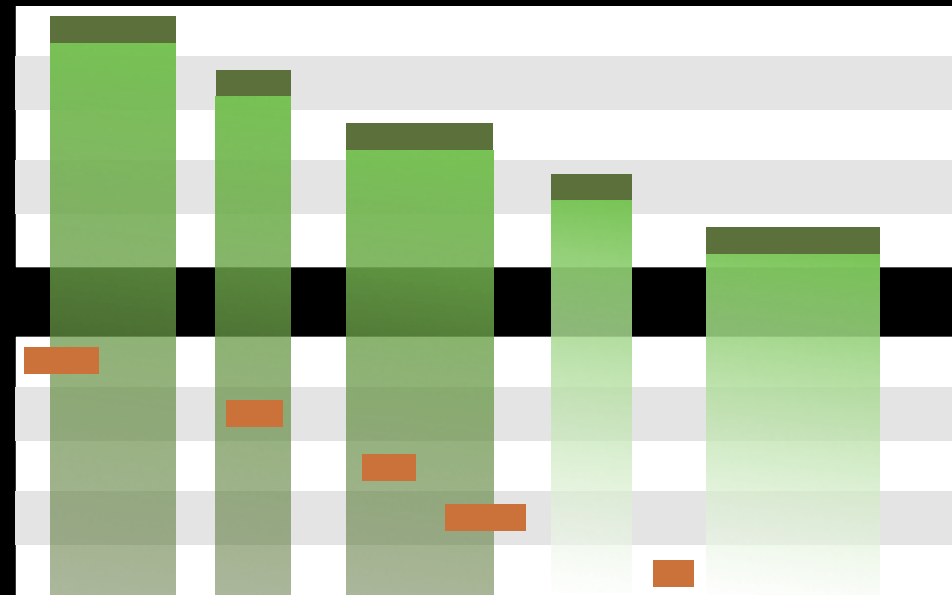
(Identify which genes/exons have Repeats)



Exons



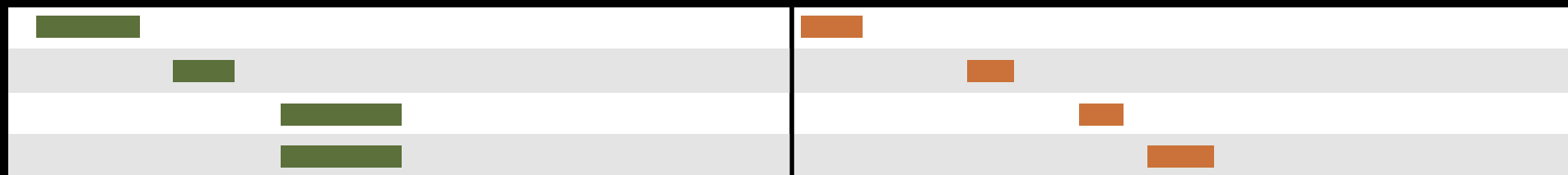
Repeats



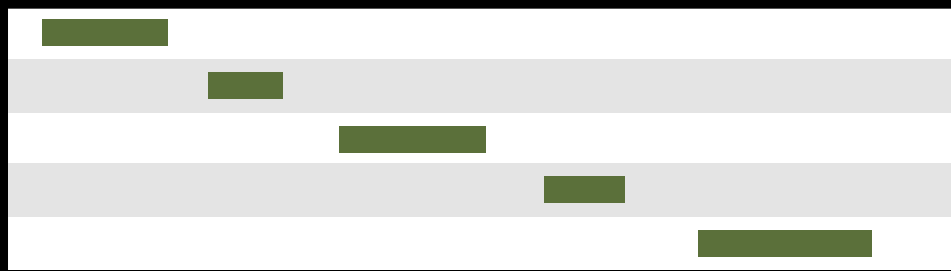
Exons

Repeats

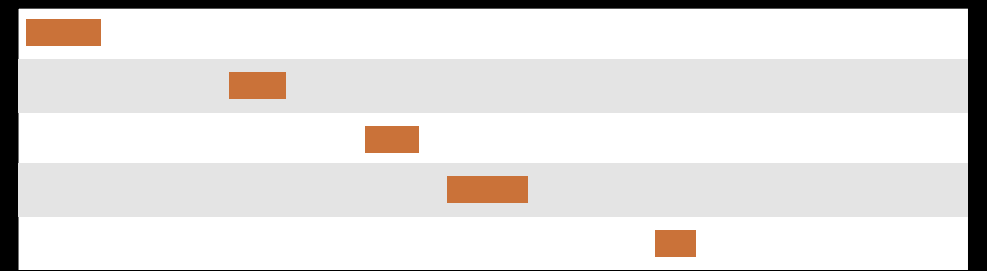
Overlap pairings



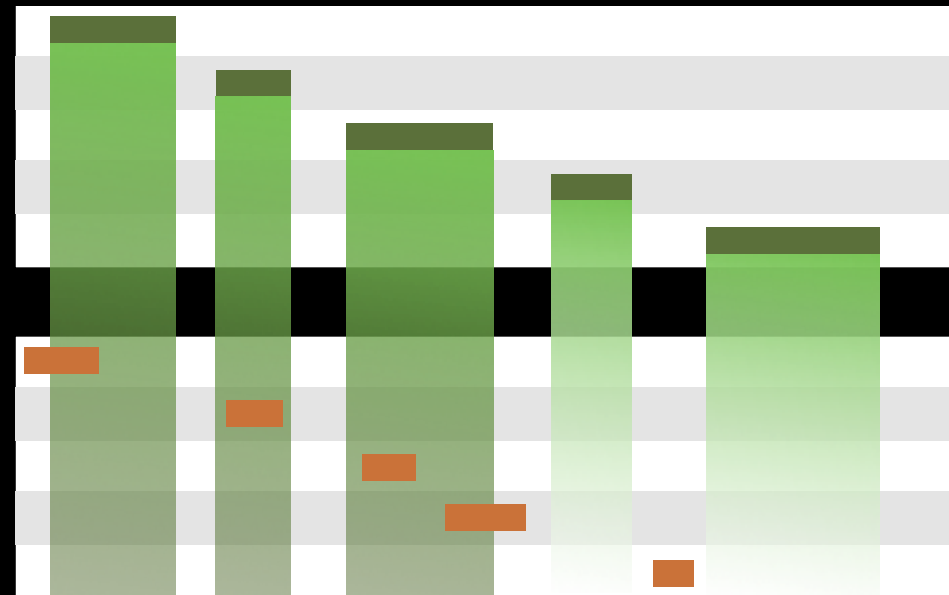
Operate on Genomic Intervals → Join
(Identify which genes/exons have Repeats)



Exons



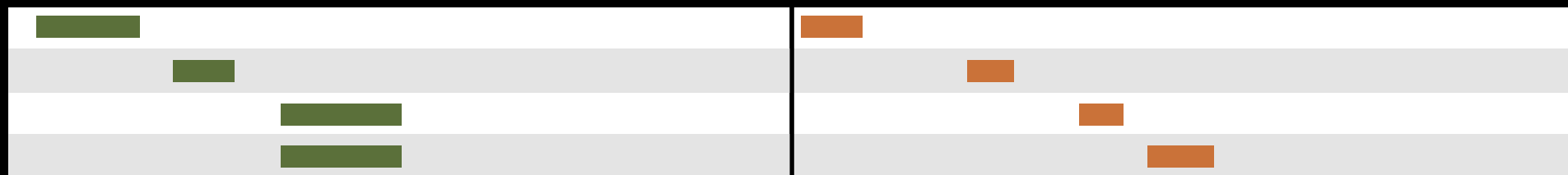
Repeats



Exons

Repeats

Overlap pairings

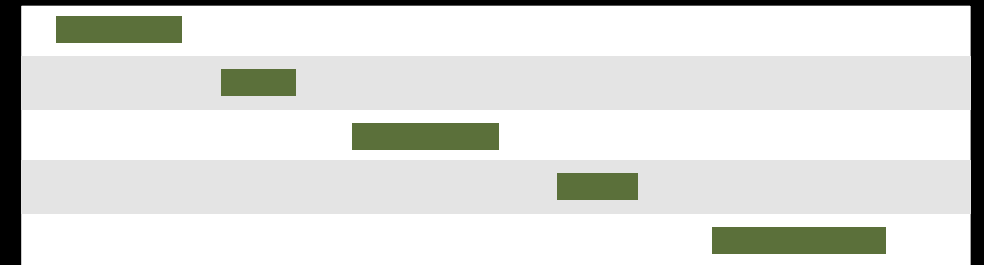


Exon overlap counts

Join, Subtract, and Group → Group
(Count Repeats per exon)



Exon overlap counts

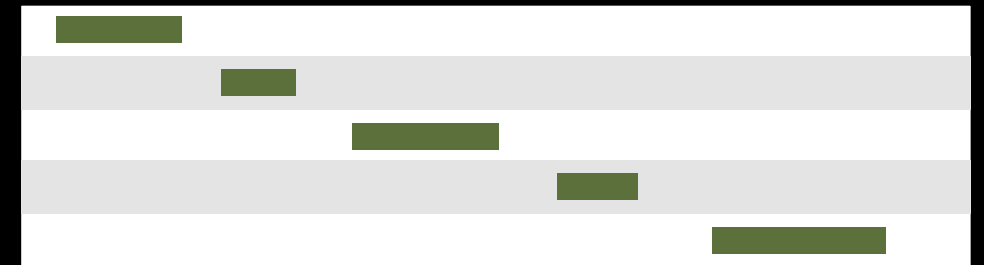


Exons

We've answered our question, but we can do better.
Incorporate the overlap count with rest of Exon information

	1
	1
	2

Exon overlap counts



Exons

	1		0
	1		0
	2		0

Join on exon name

Join, Subtract, and Group → Join

(Incorporate the overlap count with rest of Exon information)

1	1
2	1
3	2

Exon overlap counts

Device Type	Percentage of Respondents
Smartphone	100%
Tablet	95%
Laptop	85%
Desktop Computer	75%
Smartwatch	65%
Smart TV	55%

Exons

The diagram illustrates a cut in a network flow problem. It consists of two tables of node values and a network graph.

Top Table (Left):

■	1
■	1
■	2

Top Table (Right):

■	0
■	0
■	0

Bottom Table (Left):

■	1
■	1
■	2

Bottom Table (Right):

■	1
■	1
■	2

Network Graph:

The graph shows a network of nodes and edges. A red line indicates a cut. The nodes are labeled with values: 1, 1, 2, 0, 0, 0, 1, 1, 2. The edges are labeled with values: 1, 1, 2. The cut separates the nodes into two sets: those with values 1, 1, 2 and those with values 0, 0, 0, 1, 1, 2.

Real cut:

The red line indicates a cut in the network. The nodes are labeled with values: 1, 1, 2, 0, 0, 0, 1, 1, 2. The edges are labeled with values: 1, 1, 2. The cut separates the nodes into two sets: those with values 1, 1, 2 and those with values 0, 0, 0, 1, 1, 2.

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.

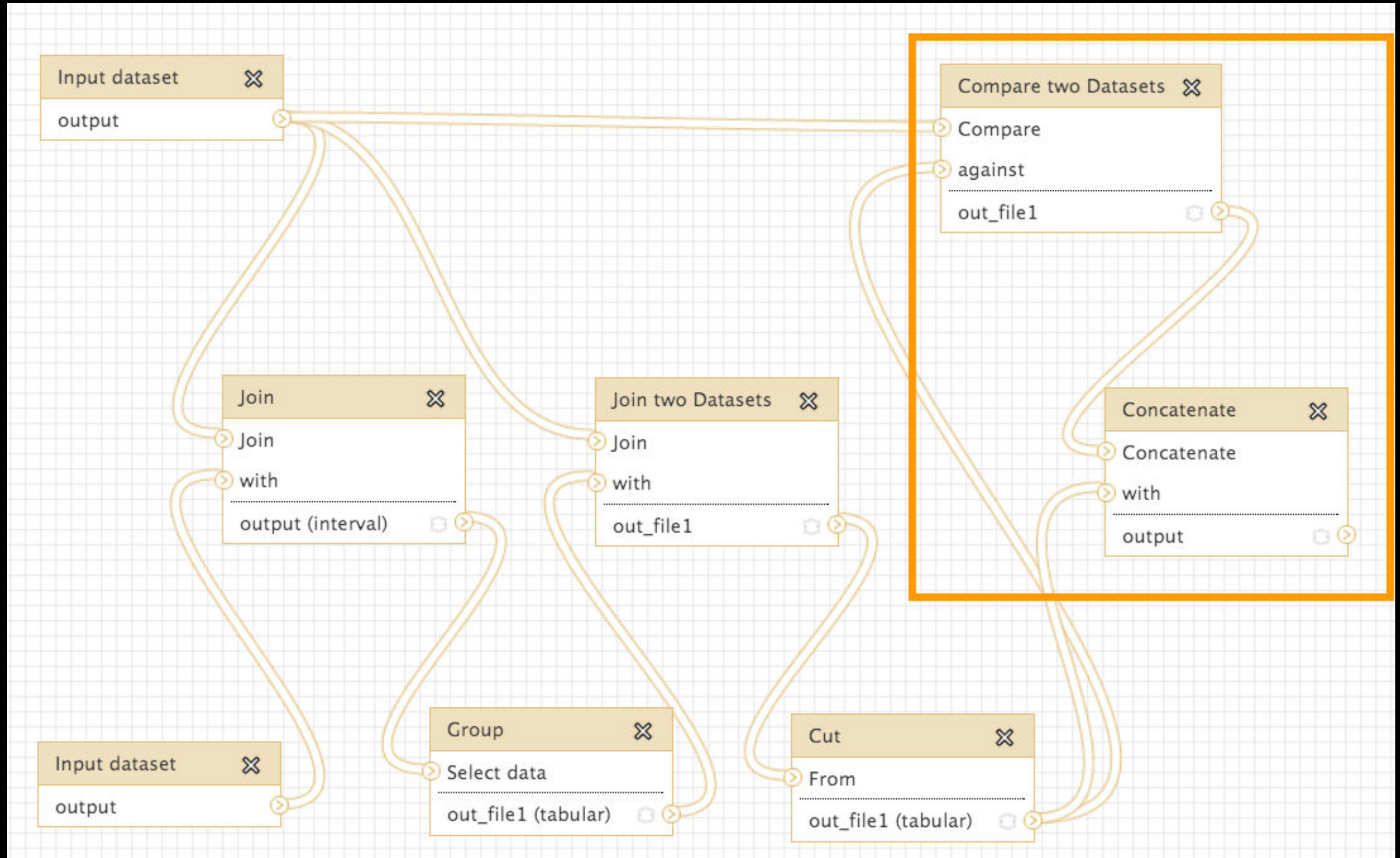
<http://cloud1.galaxyproject.org/>

<http://cloud2.galaxyproject.org/>

<http://cloud3.galaxyproject.org/>

<http://cloud4.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

9:00 Welcome, Galaxy Platforms

9:20 Basic Analysis with Galaxy

10:30 Break

10:45 RNA-Seq Example

12:40 Project and Community Overview

13:00 Done

The Agenda

9:00 Welcome, Galaxy Platforms

9:20 Basic Analysis with Galaxy

10:30 Break

10:45 RNA-Seq Example

12:40 Project and Community Overview

13:00 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!**

[illegible]

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



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

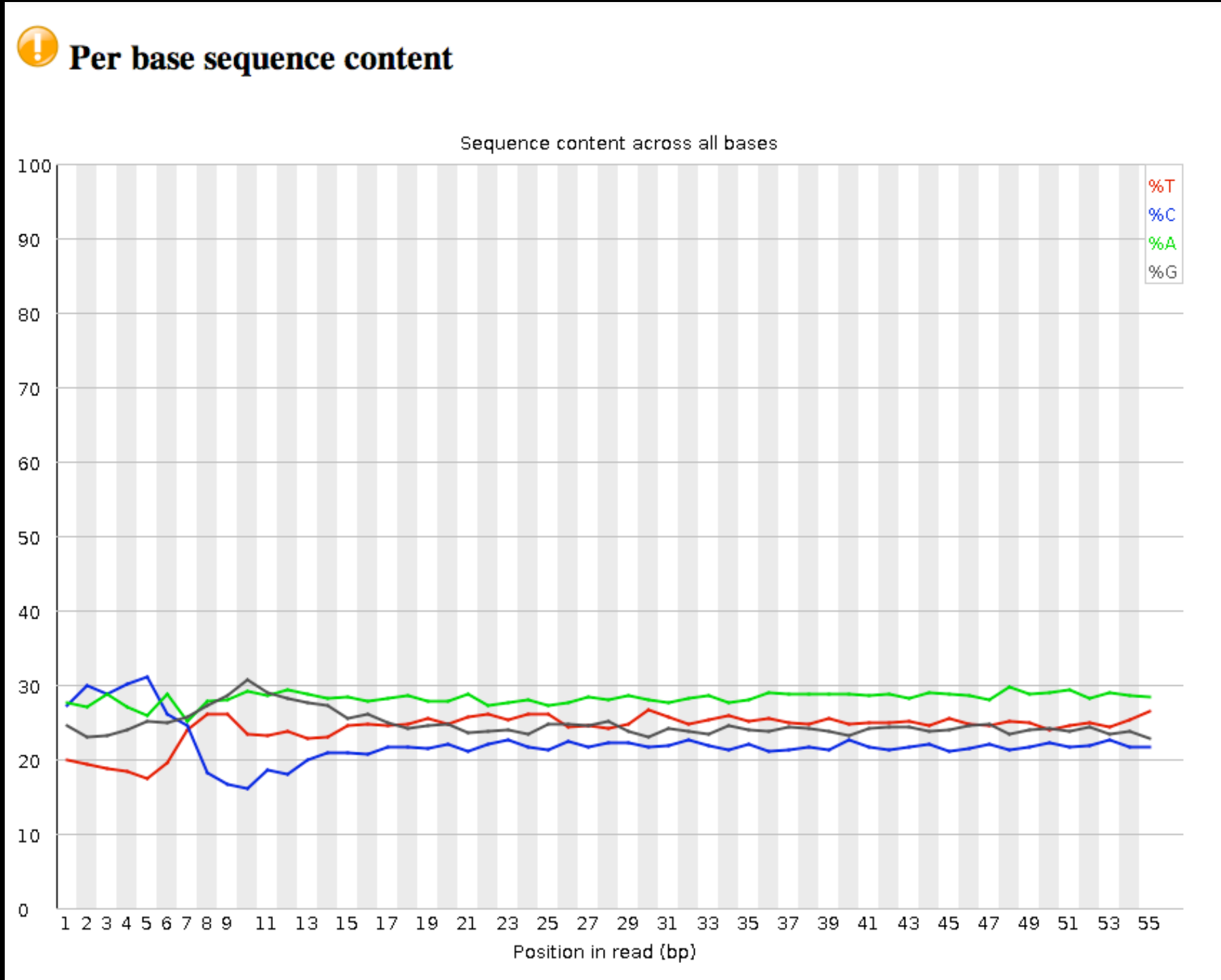
NGS Data Quality: Assessment tools

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?

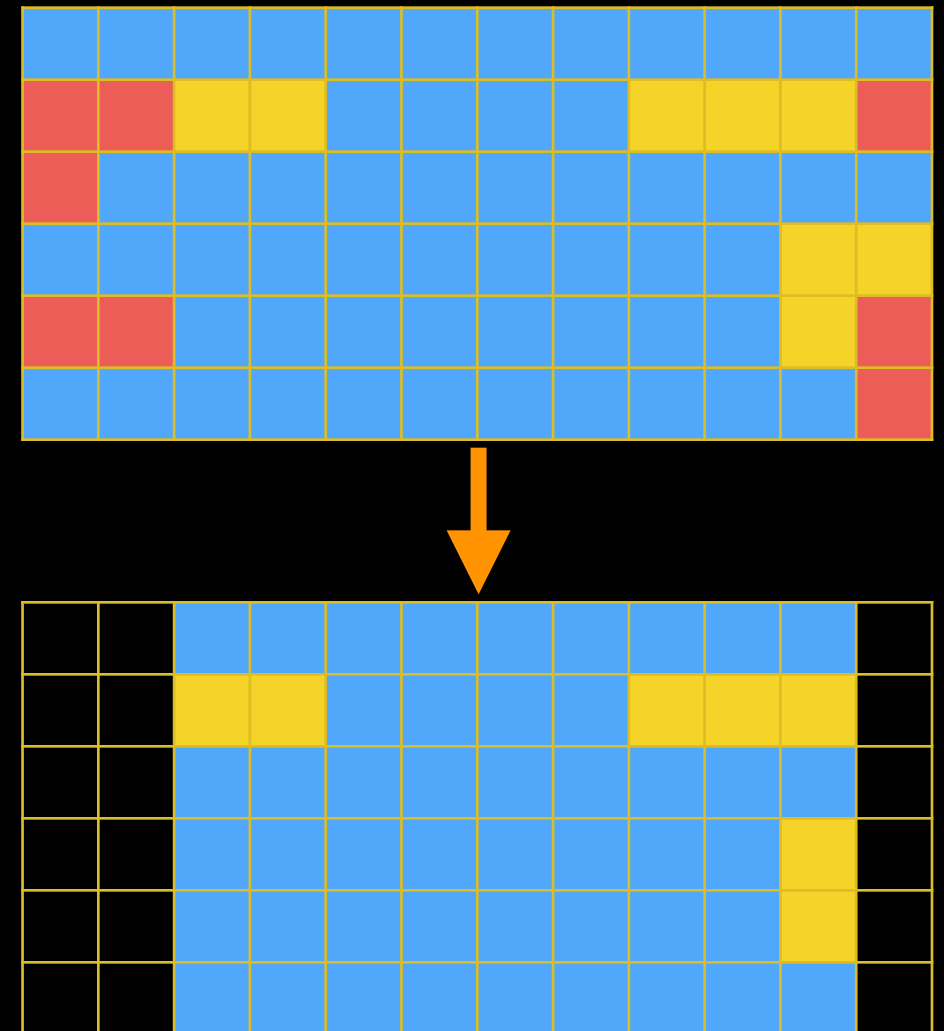


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

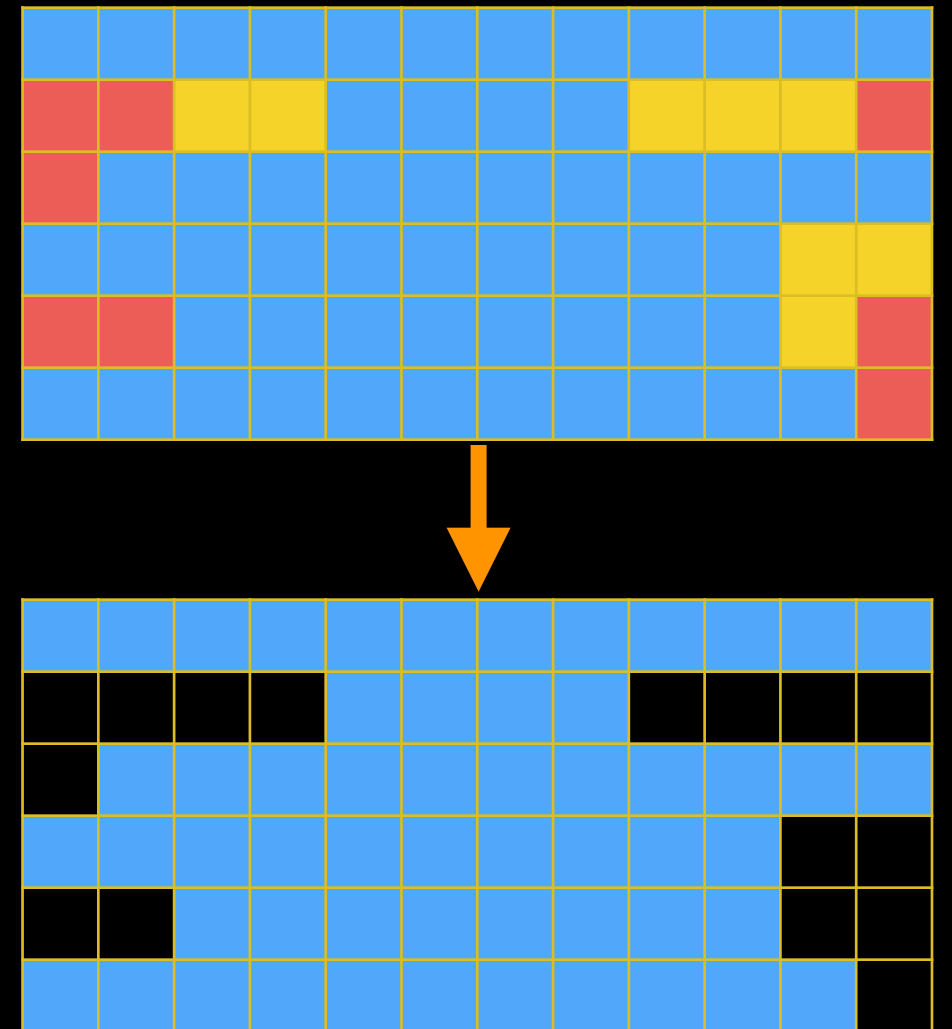
NGS Data Quality: Trim as we see fit

- 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

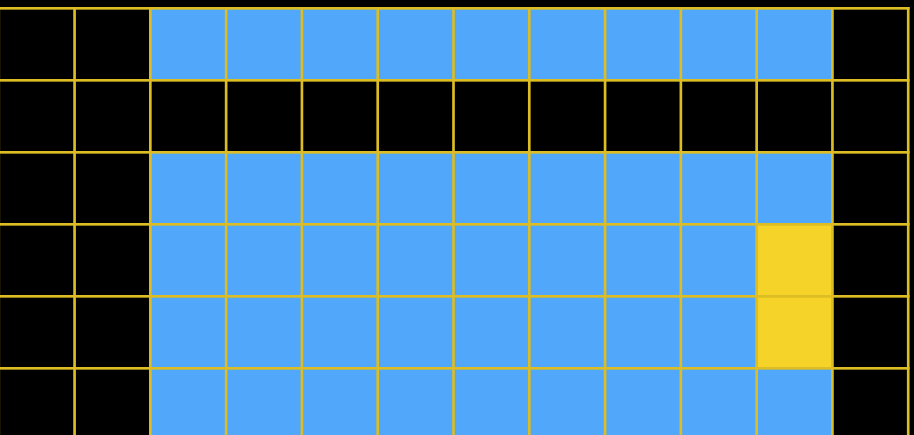
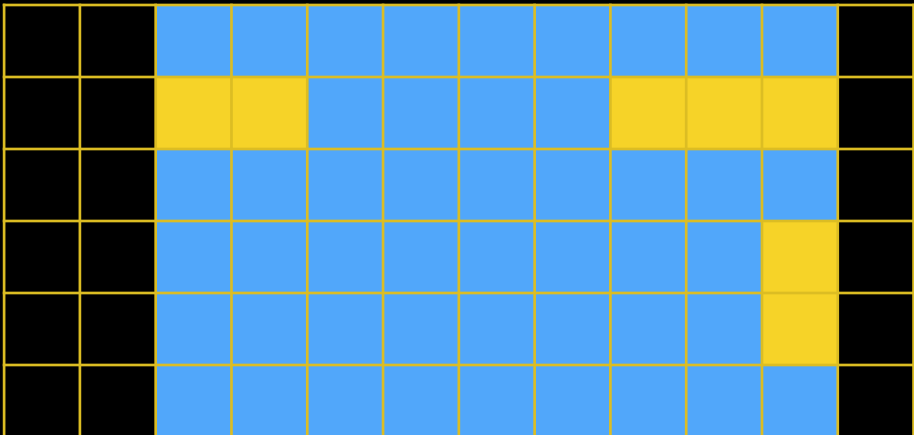
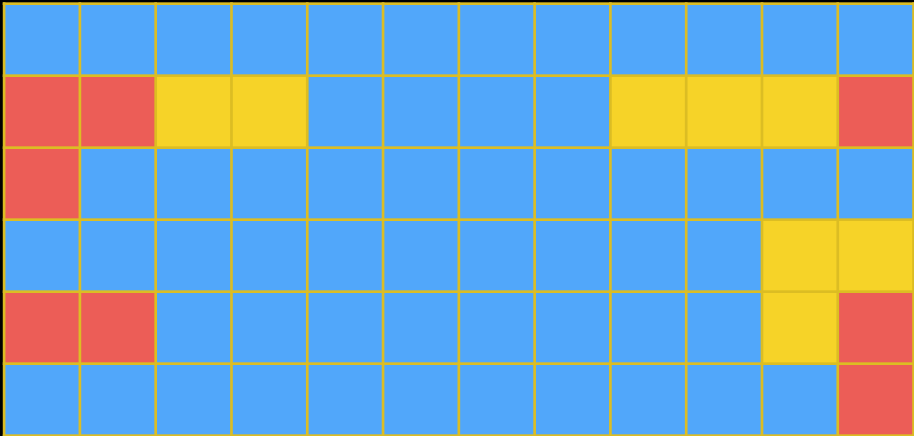


NGS Data Quality: Base Quality Trimming

- Trim 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?*

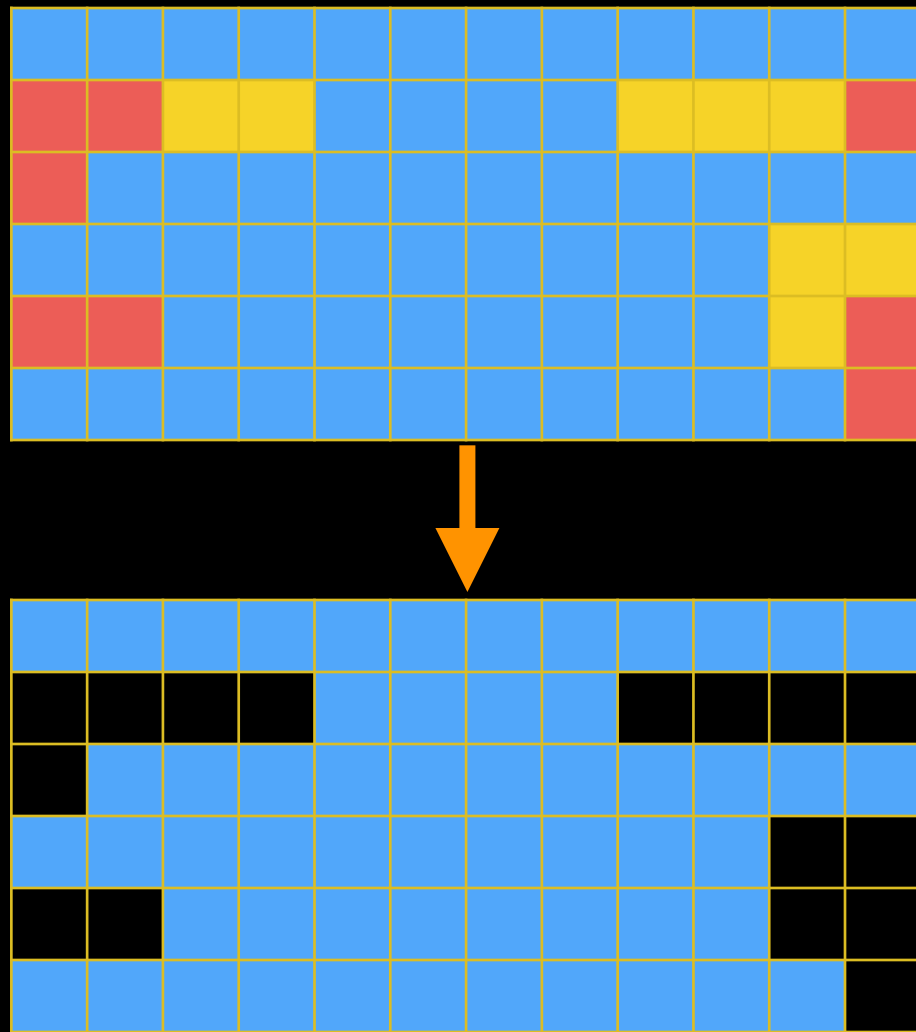
- Introduced 3 options
 - One preserves original read length, two don't
 - One preserves number of reads, two don't
 - Two keep/make every read the same length, one does not
 - One preserves pairings, two don't

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.”

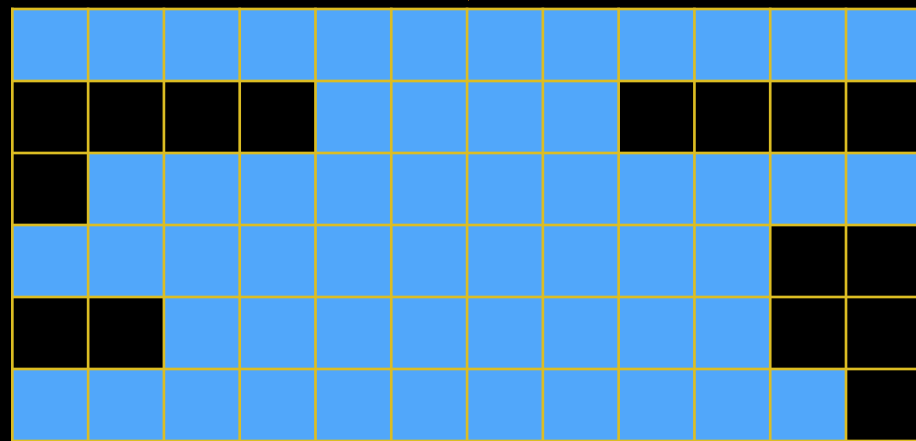
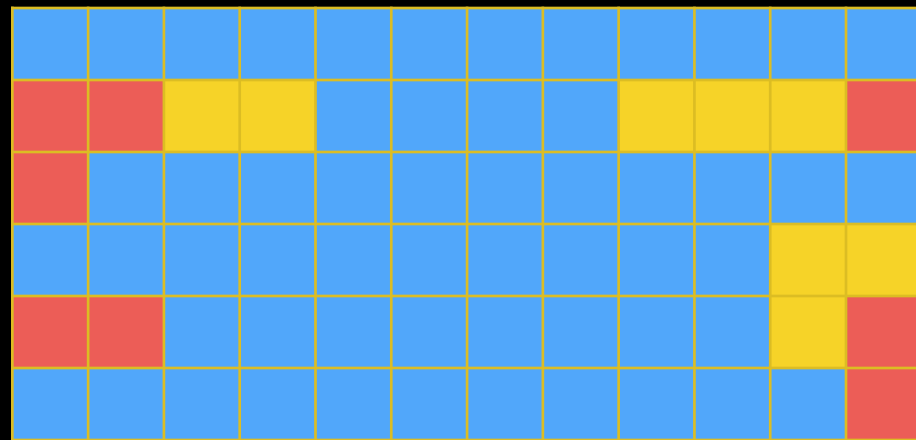
Most of us

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 ...*):

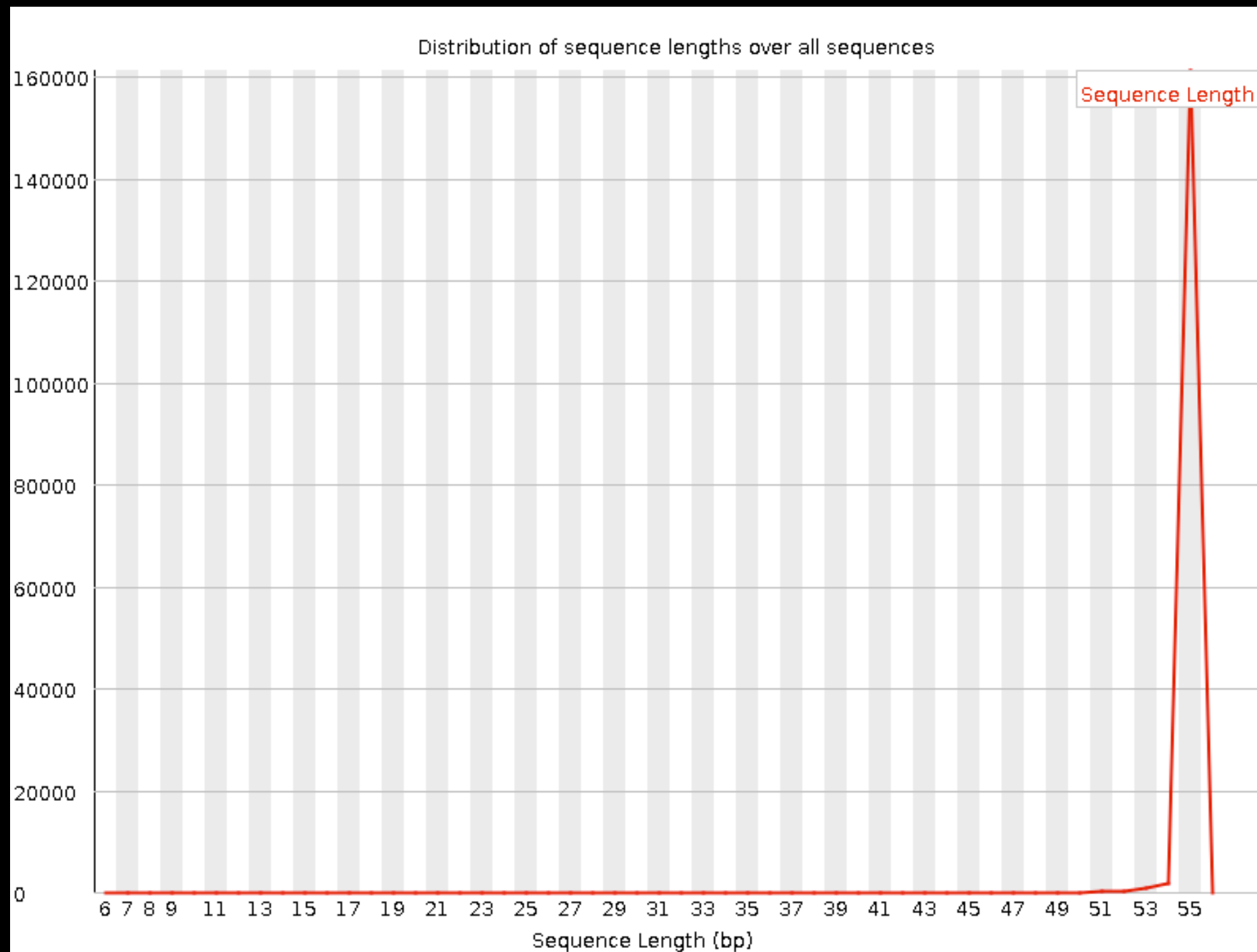
- NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**

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
TT	342	0.2052981325073385	No Hit
CGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	325	0.19509325457568719	No Hit
CTGCATTATAAAAAGGACAGCCAGATATCAACTGTTACAGAAATGAAATAAGACG	230	0.13806599554587093	No Hit
CGGCCGCAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	199	0.11945710049403614	No Hit
GTCAGCTCAACTTGTAGGCCCCAAAAGAAAACAGCGTCTTACTGGGGAGGGATAT	197	0.11825652661972422	No Hit

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

But this will break pairings.

NGS Data Quality: Done with 1st Replicate!

Now, only 5 more to go!

Workflows?

Create a QC workflow that does the trimming

Or, cheat and import the Sliding window QC, paired end, keep empties published workflow

Or, really cheat and just import the already trimmed datasets from the RNA-Seq Example → Trimmed FASTQ shared data library

NGS Data Quality: Further reading & Resources

FastQC Documentation

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

- 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*
- Things like mean inner distance, how much guidance to give it when predicting junctions, indel and "coverage" search, tie-breaking, ...

(See 5 pages on this at the end of these slides)

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

Differential Gene Expression: 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 HTSeq datasets together on gene name
Cut out the duplicate gene name columns
OR, just use the 6x DESeq Prep workflow

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

9:00 Welcome, Galaxy Platforms

9:20 Basic Analysis with Galaxy

10:30 Break

10:45 RNA-Seq Example

12:40 Project and Community Overview

13:00 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>

 **Galaxy Web 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](#)

Find

Everything on ...

Tools for ...

Email about ...


Source code for ...

Published Histories, Pages, Workflows, about ...

Documentation on ...

Papers using Galaxy for ...

Related feature requests

 **Galaxy Web Search**

About 444 results (0.06 seconds)

[Galaxy | Accessible Page | ChIP-seq exercise](#)

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

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 usegalaxy.org 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 for your convenience.

 usegalaxy.org

Community & Project

Galaxy has a large and active user community and many ways to get involved.

- [Community](#)

Deploy Galaxy

Galaxy is a free and open source project available to all. Local Galaxy servers can be set up by [downloading](#) the Galaxy application.

- [Admin](#)
- [Cloud](#)
- [Galaxy Appliance](#)

 getgalaxy.org

Contribute

- **Users:** [Share](#) your histories, workflows, visualizations, data libraries, and [Galaxy Pages](#), enabling others to use and learn from them



Use Galaxy

[Servers](#) • [Learn](#)
[Main](#) • [Share](#) • [Search](#)

Communicate

[Support](#) • [Biostar](#)
[Events](#) • [Mailing Lists](#)
[News](#)  • [Twitter](#)

Deploy Galaxy

[Get Galaxy](#) • [Cloud](#)
[Admin](#) • [Tool Config](#)
[Tool Shed](#) • [Search](#)



Contribute

[Develop](#) • [Share](#)
[Issues & Requests](#)
[Teach](#) • [Support](#)

Galaxy Project

[Home](#) • [About](#)

Events

News

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. 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 Horizon, send it to outreach@galaxyproject.org.

For events prior to this year, see the [Events Archive](#).

Upcoming Events



Date	Topic/Event	Venue
May 6-7	Scaling Galaxy for Big Data	NGS TGAC
May 9	Introduction to Galaxy Workshop	The C (TGA)
May 12	Galaxy Workshop	Unive UK
	Galaxy Project Update	5th E Meet Edinb
May 13	Galaxy Workshop	Instit Medic
May 12-14	Short course on RNA-seq and ChIP-seq	Unive Norw
May 16	Galaxy Initiation	Form Plate Biolog
May 19	Initiation au traitement et à l'analyse des données métabolomiques sur la plateforme scientifique web Galaxy IFB-MetaboHUB	8e Jo Lyon

News Items

May 2014 Galaxy News



The [May 2014 Galaxy Update Newsletter](#) is out! There's a lot going on in the project and the community right now. The big news in the past month is the move from the [Galaxy-User mailing list](#) to [Galaxy Biostar](#) for user support. This has been running for a week now, and has been very well received.



The other big news is upcoming events. [Early registration for GCC2013 closes May 23](#). Register now and save more than 70% on registration costs, and [Training Day](#) registration is an additional 55% off if you register for both at the same time. We are also pleased to [announce this year's keynote speaker](#) and the first ever [GCC Hackathon](#).



There's also a [Galaxy UK Tour](#) which is visiting Norwich and Edinburgh in May, and [there are at least 17 other Galaxy related events](#) in the next 70 days in Norway, France, *online*, Croatia, Thailand, Canada, the US, the Netherlands, and Australia



As always, there are [new papers](#) (47 of them, including four we highlighted), [new public Galaxy Servers](#) (Globus Genomics Proteomics and SunLab Galaxy), [new jobs](#) (7 postings in 6 countries), [new tools](#) in the project ToolShed (um, lots), and a [new public ToolShed](#) (at the [Dutch Techcentre for Life Sciences \(DTL\)](#)).

[Dave Clements](#) and the [Galaxy Team](#)

Posted to the [Galaxy News](#) on 2014-04-30

Galaxy Biostar Launched

Galaxy has teamed up with [Biostar](#) to create a [Galaxy User support forum](#) at <https://biostar.usegalaxy.org>!



We want to create a space where researchers using Galaxy can come together and share both scientific advice and practical tool help. Whether on [usegalaxy.org](#), a [CloudMan](#) instance, or any other Galaxy ([public](#) or [local](#)), if you have something to say about *Using Galaxy*, this is the place to do it!

Current integration with usegalaxy.org

- We imported the **whole history** of the [galaxy-user@bx.psu.edu](#) mailing list into [Galaxy Biostar](#). Your prior posts are automatically claimed when you login!
- If you access [Galaxy Biostar](#) from <http://usegalaxy.org> (Menu: **Help** → **Galaxy Biostar**) you will be automatically logged in. A Galaxy Biostar account will be created for you if it did not previously exist. To obtain this account's password please use the [password reset feature](#) of Galaxy Biostar.
- When you have a question, search Galaxy Biostar directly from any Galaxy tool page.

Read more about how to get started on the [Biostar wiki page](#).



GALAXY

COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

<http://bit.ly/gcc2014>



Galaxy Resources & Community: Videos

The screenshot shows the Vimeo profile for the Galaxy Project. The header includes the Vimeo logo and navigation links: Me, Videos, Create, Watch, Tools, Upload. A search bar is located in the top right. The profile name is "Galaxy Project" with a "PLUS" badge, and it notes "Joined 1 month ago". On the left sidebar, there are three video thumbnails (two grey, one yellow) and a "Settings" button. The main content area displays statistics: 54 Videos, 0 Likes, 0 Following, 1 Group, 6 Channels, and 0 Albums. Below this is a "Recently Uploaded" section with a link to "See all 54 videos". Four video thumbnails are shown in a 2x2 grid:

- Using Galaxy protocol 3**: Calling Peaks For ChIP-seq Data. CPB Using Galaxy 3, 5 days ago.
- Using Galaxy protocol 2**: Loading Data and Understanding Datatypes. CPB Using Galaxy 2, 5 days ago.
- Using Galaxy protocol 1**: Finding Human Coding Exons with Highest SNP Density. CPB Using Galaxy 1, 5 days ago.
- usegalaxy.org**: FASTQ Prep Illumina. FASTQ Prep - Illumina, 1 week ago.

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

“How to”
screencasts on
using and
deploying
Galaxy

Talks from
previous
meetings.

<http://vimeo.com/galaxyproject>

The Agenda

9:00 Welcome, Galaxy Platforms

9:20 Basic Analysis with Galaxy

10:30 Break

10:45 RNA-Seq Example

12:40 Project and Community Overview

13:00 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



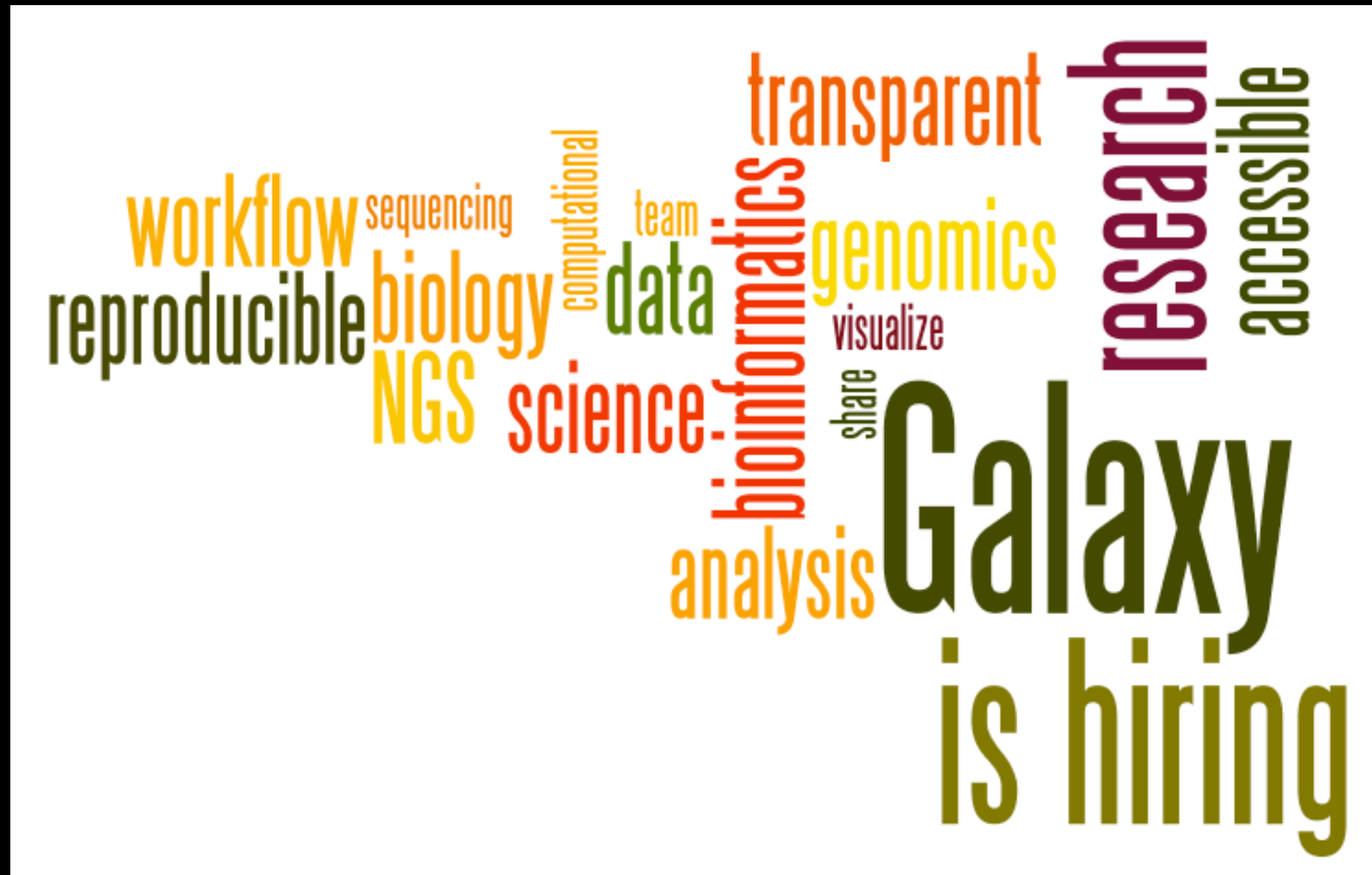
Nick Stoler



James Taylor

<http://wiki.galaxyproject.org/GalaxyTeam>

Galaxy is hiring post-docs and software engineers



Please help.

<http://wiki.galaxyproject.org/GalaxyIsHiring>

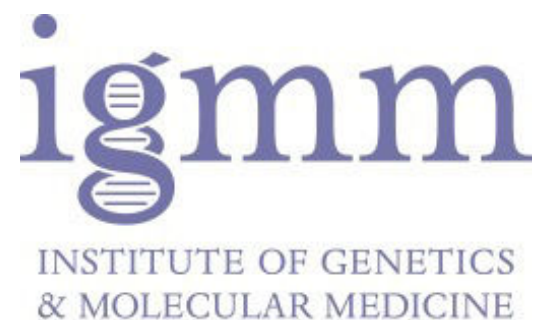
Also Thanks To



THE UNIVERSITY
of EDINBURGH



Al Kerr
Mick Watson
James Pendergast



Feedback

We need it!

<http://bit.ly/gxyedifedback>

Thanks



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

Shaun Webb
WTCCB
University of Edinburgh
shaun.webb@ed.ac.uk

Bert Overduin
Edinburgh Genomics
University of Edinburgh
Bert.Overduin@ed.ac.uk

The Agenda

9:00 Welcome, Galaxy Platforms

9:20 Basic Analysis with Galaxy

10:30 Break

10:45 RNA-Seq Example

12:40 Project and Community Overview

13:00 Done (really)

Sharing, Publishing, and Reproducibility

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**





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:
Advanced Search

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7},
Francesca Chiaromonte⁴, Guruprasad Ananda^{1,3}, Wen-Yu Chung^{1,3,8},
James Taylor^{1,5,9}, Anton Nekrutenko^{1,3,9} and The Galaxy Team¹

OPEN ACCESS ARTICLE

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**

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 CHIAROMONTE⁴, 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**):

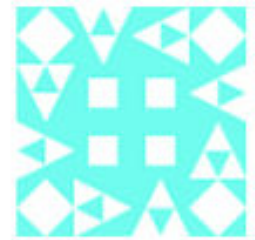
 **Galaxy History | metagenomic analysis**  

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 there, they can be re-analyzed through Galaxy using the above workflows or downloaded.



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

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

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.

Mapping with Tophat: Lets do it some more!

NGS: RNA Analysis → Tophat

for the remaining replicates

Or not.

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

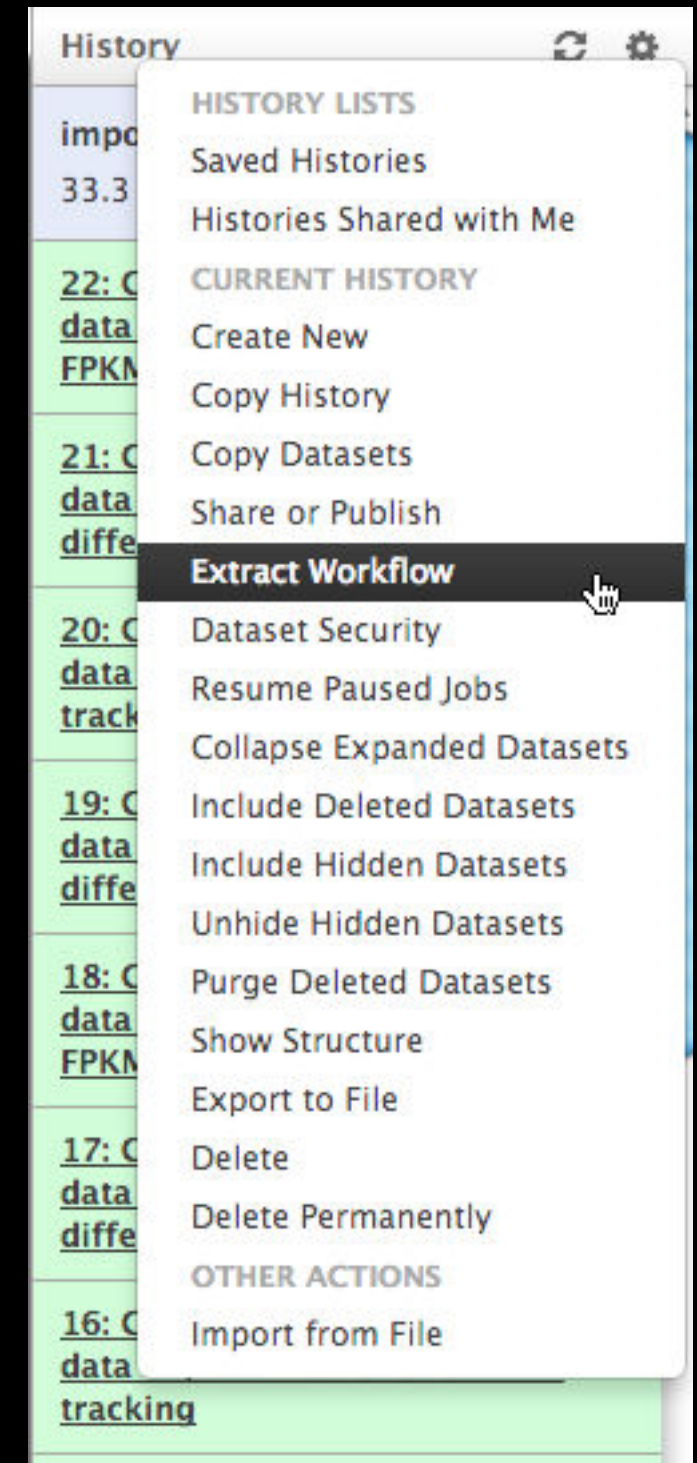
Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



Community can create, vote and comment on issues

Galaxy: Development

Inbox

To add cards, use <http://galaxyproject.org/trello>
4 votes 1

add ma seq metrics and downsample sam to picard tools
3 votes 1

Reference genomes
2

Please merge patch to bowtie2 wrapper (adds support for mapping fasta files)
1

R 3.0.2 woes on test.g2.bx.psu.edu - libgomp.so not found when job runs: [Recrpt: error while loading shared libraries: libgomp.so.1: cannot open shared object file: No such file or directory\n"] - possibly execution node missing gfortran?
1

unhandled exception when installing metaphlan from source repo
1

Tool Requests

595: Add SAMtools "Sort"
5

601: SAM-to-BAM tool enhancements
1

Bug: some characters not permitted in 'add column' tool
2 votes 5

307: A tool to produce a set of random intervals.
2 votes 2

Tools: Add tool to generate simulated reads to Main.
2 votes 1

default max insert size of Bowtie2 should be increased
1 vote 4

Wrapper for bigWigToWig
1 vote 1

Converter Tool: SAM to BAM enhancements
1 vote

607: Create new tool to "trim" coordinates to ref chrom lengths
1 vote

New Tool: convert IUPAC chars to N
5 votes 1

Optimize FASTQ tools.
1

Tool 'Extract Genomic DNA' should parse GFF/GTF better so to include gene_id or transcript_id attributes
1

Enh: tabular-to-fasta should let you choose how to concatenate the id string
1

Bug Reports

Impersonate a user admin option broken when using external authentication
13 votes 1

Bug: SICER on Main dependency issue
2 votes 18 3/5

Toolshed: Installing multiple versions of the same tool results in separate entries in the tool panel.
1 vote 14

Profile Annotations bad values when "select all"
1 vote 2

The option from_file="infernal.loc" is broken.
1 vote 1

68: Apparent bug in Intersect intervals, overlapping pieces
5

106979429 138792955: problem: STAR/STAR.LD is using more RAM than size of this folder (336870812).

Bug: Returning Bitset error S36870912
4 1

EMBOSS: several tools fail with default options
3

Tools: Cloudmap reference files not found
2

Bug: Fetch taxonomic

Ideas

Implement JavaScript build process
1 vote 0 0/13

Tools: Incorporate key Cuffdiff output files for Cumberbund
1 vote 1 0/3

Workflow Editor: Provide explicit access to implicit datatype converter tools
1 vote

Google Drive / Dropbox / Box / ... integration
6 votes 3

720: Capture and report time taken to run each job
6 votes 0/2

Allow administrators to "trust" certain HTML outputs based on tool producing them.
4 votes 4

Workflows: highlight the noodles in the workflow editor upon hovering
4 votes 2

6: Option to disable automatic history creation
4 votes 2

Allowing workflow step dependencies when no input/output files exist
4 votes 1

Assistive UI
4 votes 1 0/4

For sensible output. Add input name to Son_string
4 votes 0/3

RFC: Implement sophisticated user behavior analysis tool

Pull Requests / Patches

685: Patch for FASTQ paired-end issue
1

Tools: Bowtie Wrapper Pull Requests from Community
2 votes 0 1

Pull Request #343 - Need to traverse the other_value dict to find dependencies for ParamValueFilter in dynamic_options when the dependencies are scoped in a conditional. Error was noted attempting to run iuc SnEff 3.4 in a workflow.
0/3

Pull Request #338 - Patch to expose the actual dataset id in the LODA and HDA to_dict calls (in addition to the instance id).
0/3

Pull Request #336 - Traverse context for SelectToolParameter need_date_validation.
0/3

Pull Request #334 - Trello Card #1437: Optional Input Datasets Not Compatible with Parallelism Tag
0/3

Pull Request #261 - tools/fastq/fasta_paired_end_joiner: added support for recent Illumina headers
0/3

Project in Planning

308: Demystifying the first ever Galaxy login experience - make tools offer test data if empty history?
3 votes 2

Data Manager: Genome Builds / dbkeys: Make adding builds accessible by Data Manager tools
2 votes 0/3

resetting the password deactivates the user
1 vote 4

Tools: Moving to BAM format as primary representation of sequence data
1 vote 2

Libraries: Role selection
1 vote 1

Data Manager: Rsync version
1 vote

UI enhancements
1 vote 0/7

BWA aln -n param update
2

Show placement in queue / throughput
1

Core: Make the jobs admin interface not suck
6 votes

Deleting history using the API does not delete/stop jobs
5 votes 7

Tool Shed (and Galaxy?) should have user profiles.
3 votes 5

Menu

Members

Activity

- Peter Cock on Tool Shed (and Galaxy?) should have user profiles.**
I like the ORCID idea from John, might help in reverse for recognising ToolShed repositories as scientific output?
2 hours ago
- martenson removed dorine francheteau from 623: picard index indicates failure, but it is successful.**
2 hours ago
- martenson joined Tool Shed (and Galaxy?) should have user profiles..**
2 hours ago
- martenson on Tool Shed (and Galaxy?) should have user profiles.**
Ideas don't bring harm. I am merely trying to determine the demand for / priority of this
2 hours ago
- Björn Grüning on Tool Shed (and Galaxy?) should have user profiles.**
Why not? I'm not that social web guy, but it does not harm, or?
2 hours ago
- martenson on Tool Shed (and Galaxy?) should have user profiles.**
Social Logins? Persona, ResearchGate, LinkedIn, Twitter, G+, FB ?
2 hours ago - edited 2 hours ago
- John Chilton on Tools: Dataset Collections -**

<http://bit.ly/gxytrello>

Galaxy Australasia Workshop

2
0
1
4

We also support
community
organized efforts
and events.



Galaxy Resources & Community: CiteULike Group



[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 Myrtenäs](#), [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>