

NGS Data Analysis and Galaxy

University of Pretoria
Pretoria, South Africa
14-18 October 2013

Dave Clements, Emory University

<http://galaxyproject.org/>

Fourie Joubert, Burger van Jaarsveld
Bioinformatics & Computational Biology Unit
University of Pretoria

<http://science.up.ac.za/>



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA



AFRICAN CENTRE FOR
GENE TECHNOLOGIES



This Week

Monday	Welcome, Project Intro, Basic Galaxy Usage NGS QualityControl
Tuesday	RNA-Seq - Mapping and Transcript Prediction RNA-Seq: Differential expression and Alternative Pipelines; SNP & Variant Analysis
Wednesday	SNP & Variant Analysis Chip-Seq Analysis
Thursday	Genome Assembly Install your own Galaxy on Amazon Cloud
Friday	Customizing Galaxy, Galaxy Tool Shed, and Wrapping Tools for Galaxy

Monday Agenda

- 8:30 **Welcome and Intro**
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done

Introductions

In 50 seconds or less tell us

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

Goals

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

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

What is Galaxy?

- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- These options result in several **ways to use Galaxy**

<http://galaxyproject.org>

Galaxy is available ...

As a free (for everyone) web service

<http://usegalaxy.org>

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

Galaxy is available ...

- As a free (for everyone) web service

<http://usegalaxy.org>

- As open source software

<http://getgalaxy.org>

As Open Source Software: Local Galaxy Instances

- Galaxy is designed for local installation and customization
- Easily integrate new tools
- Run jobs on existing compute clusters
- Requires a computational resource on which to be deployed

<http://getgalaxy.org>

Got your own cluster?

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



Galaxy is available ...

- As a free (for everyone) web service

<http://usegalaxy.org>

- As open source software

<http://getgalaxy.org>



- *On the Cloud*

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

We are using this right now, and you will set up your own Galaxy on AWS on Thursday

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

Galaxy is available ...

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



A ready-to-use appliance (BioTeam)

Cloud-based solutions (Appistry, ABgenomica, AIS)

Consulting & Customization (Arctix, Deena Bioinformatics)

Galaxy Project: Further reading & Resources

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

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

<http://bit.ly/gxychoices>

Monday Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done

Basic Analysis

Which genes have most overlapping
Repeats?

HG19, chr22

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

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

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

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

Genes & Repeats: A General Plan

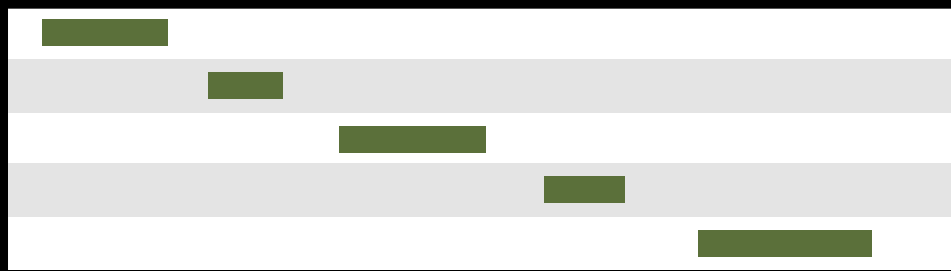
- Get some data
 - **Get Data** → **UCSC Table Browser**
- Identify which genes/exons have Repeats
- Count Repeats per exon
- Visualize, save, download, ... exons with most Repeats

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

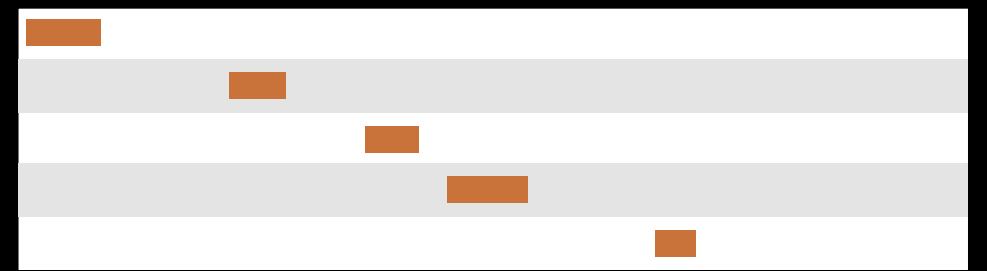
<http://cloud2.galaxyproject.org/>

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

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

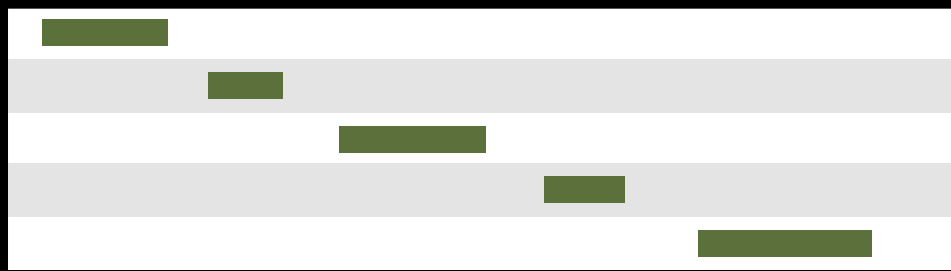


Exons

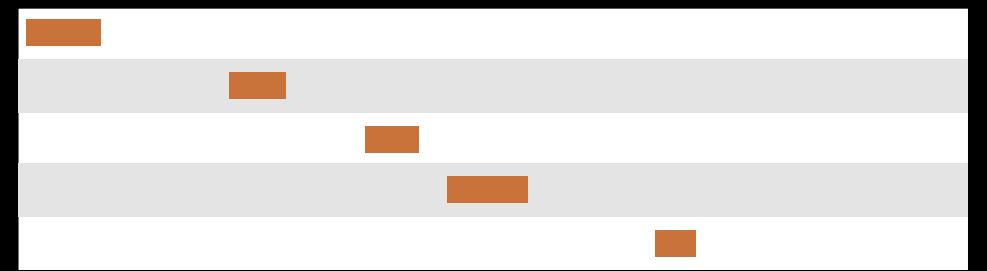


RepeatS

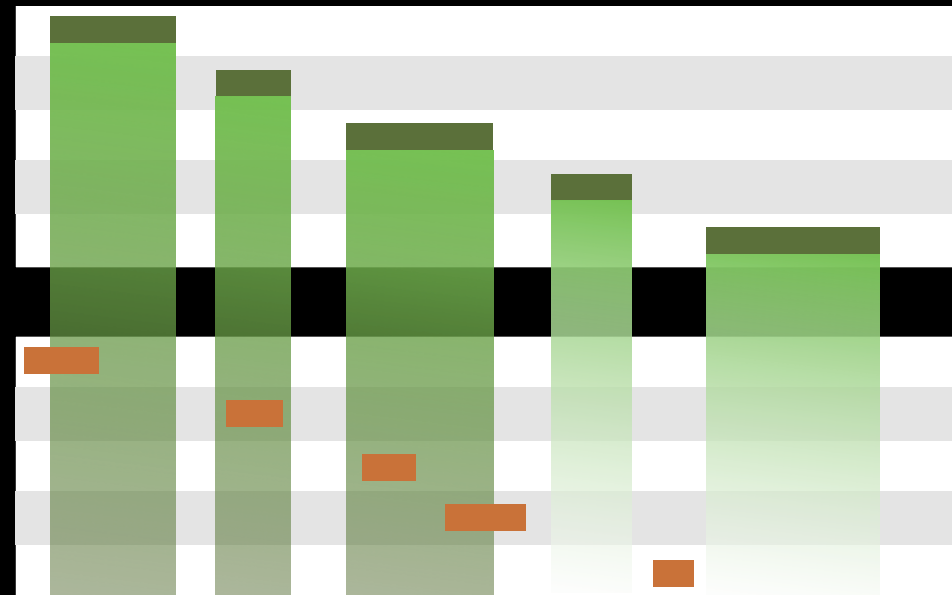
(Identify which genes/exons have Repeats)



Exons



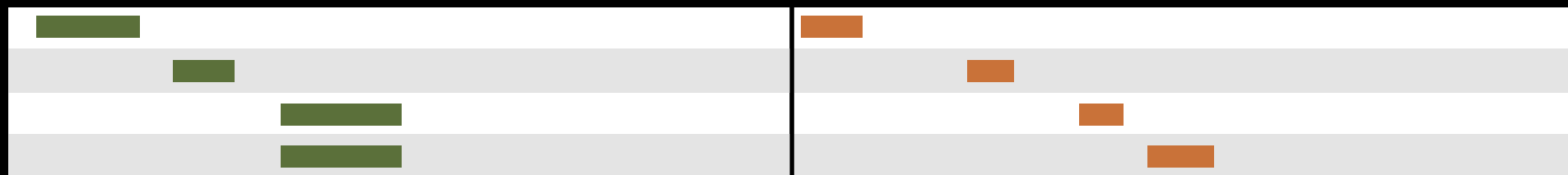
RepeatS



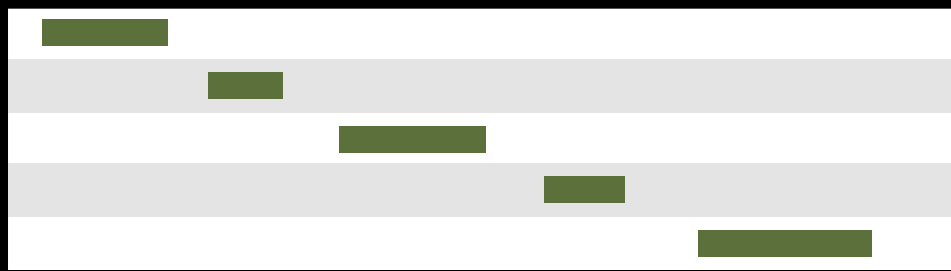
Exons

RepeatS

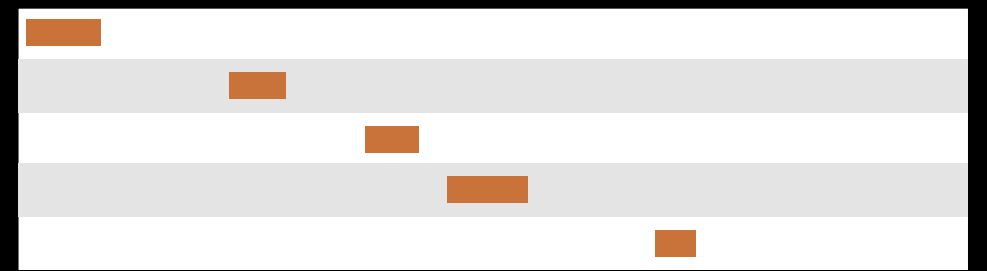
Overlap pairings



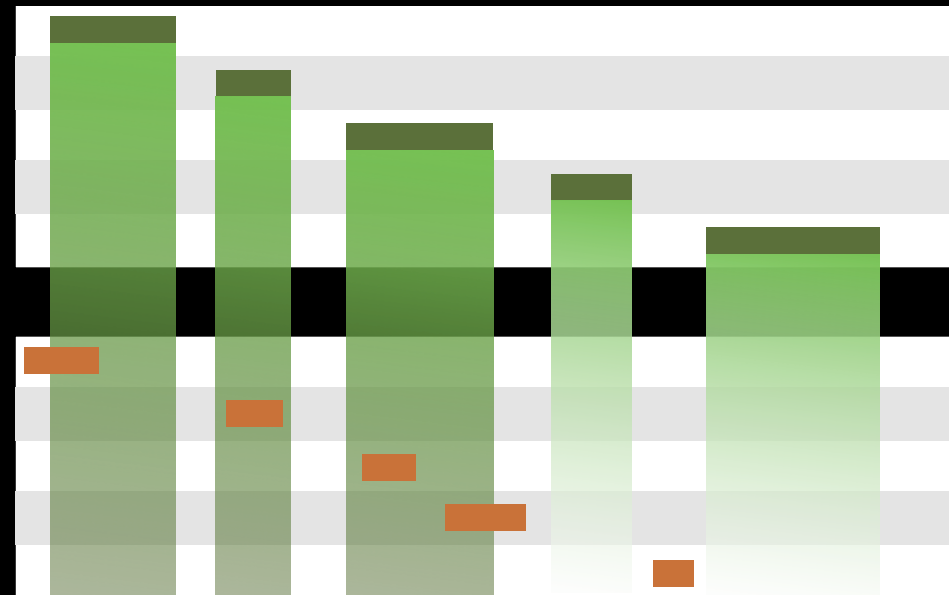
Operate on Genomic Intervals → Join
(Identify which genes/exons have Repeats)



Exons



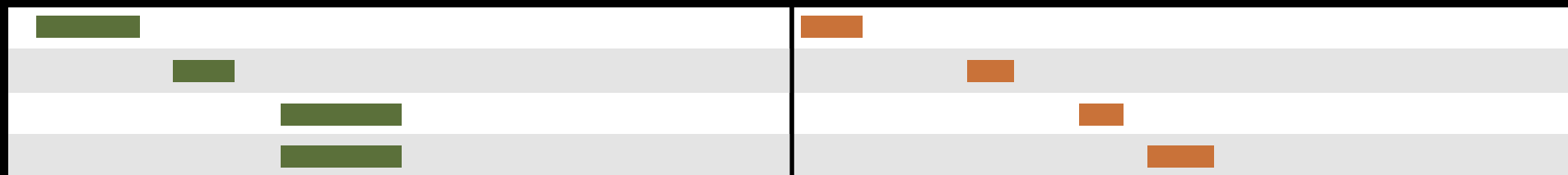
RepeatS



Exons

RepeatS

Overlap pairings

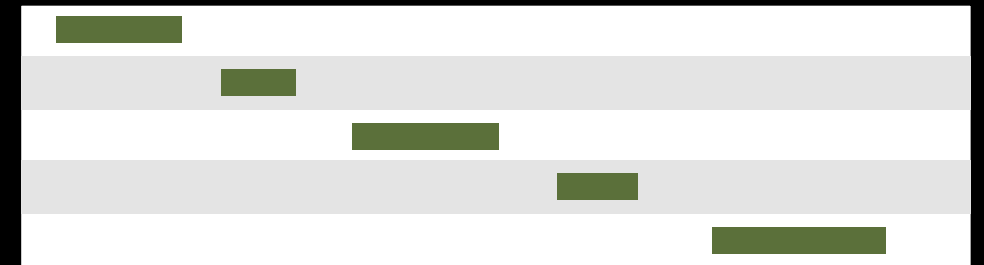


Exon overlap counts

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

	1
	1
	2

Exon overlap counts

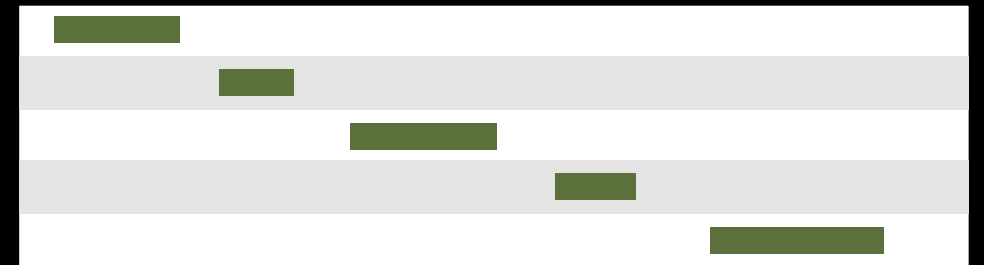


Exons

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



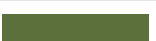


Join on exon name

Join, Subtract, and Group → Join







(Incorporate the overlap count with rest of Exon information)




	1
	1
	2

Exon overlap counts

Exons

	1		0
	1		0
	2		0

	1
	1
	2

Join on exon name

Rearrange columns w/
cut

Text Manipulation → Cut

(Incorporate the overlap count with rest of Exon information)

Basic Analysis: Further reading & Resources

<http://usegalaxy.org/galaxy101>

<https://vimeo.com/76343659>

Monday Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done



Genes & Repeats: Exercise

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

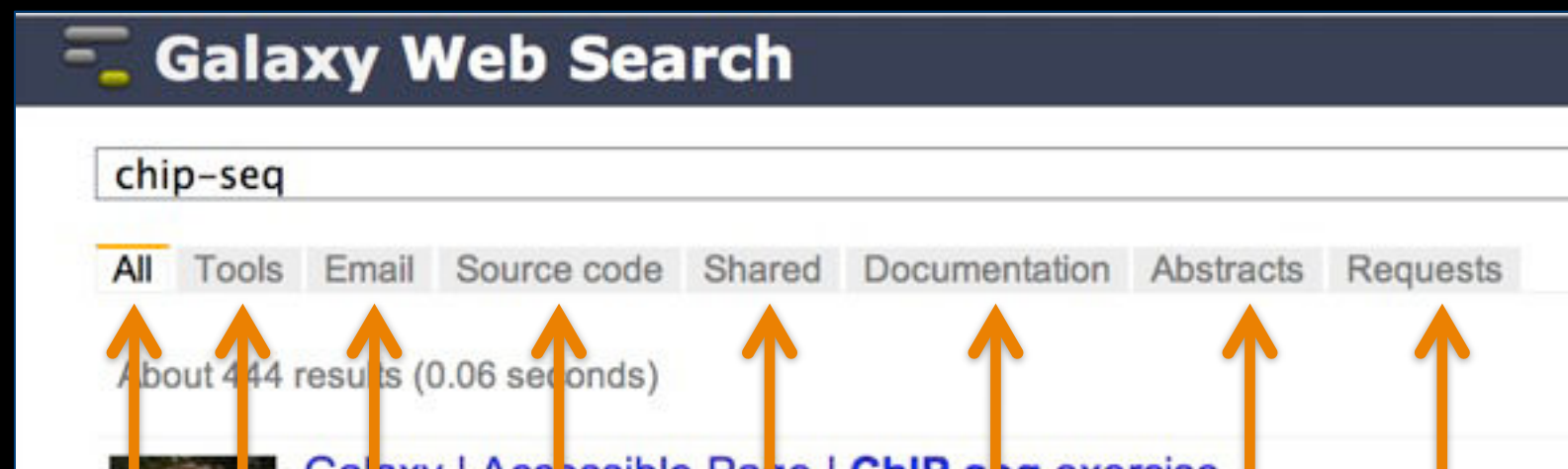
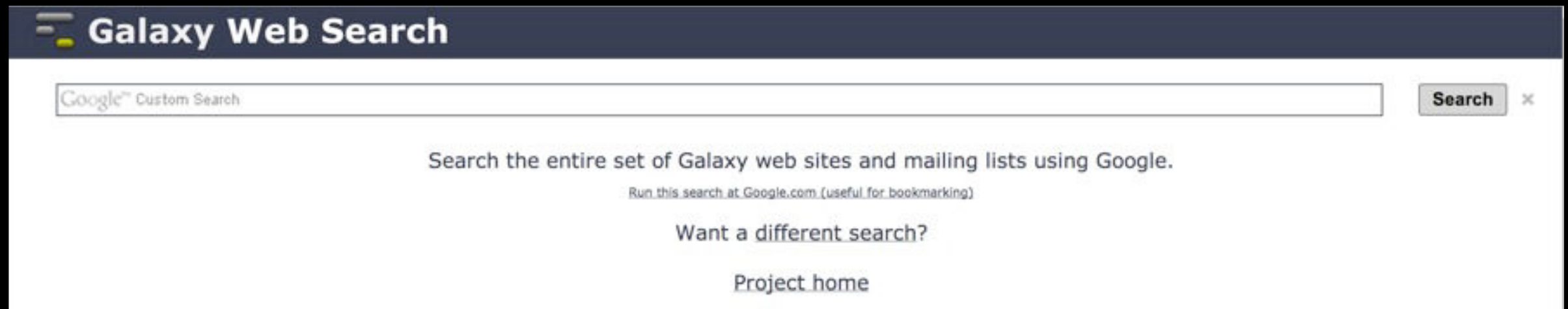
Everything you need will be in the toolboxes we used
in the first Gene/Exon-RepeatS exercise.

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

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

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

Your Friend: <http://galaxyproject.org/search>



Find

Everything on ...

Tools for ...

Email about ...

Source code for ...

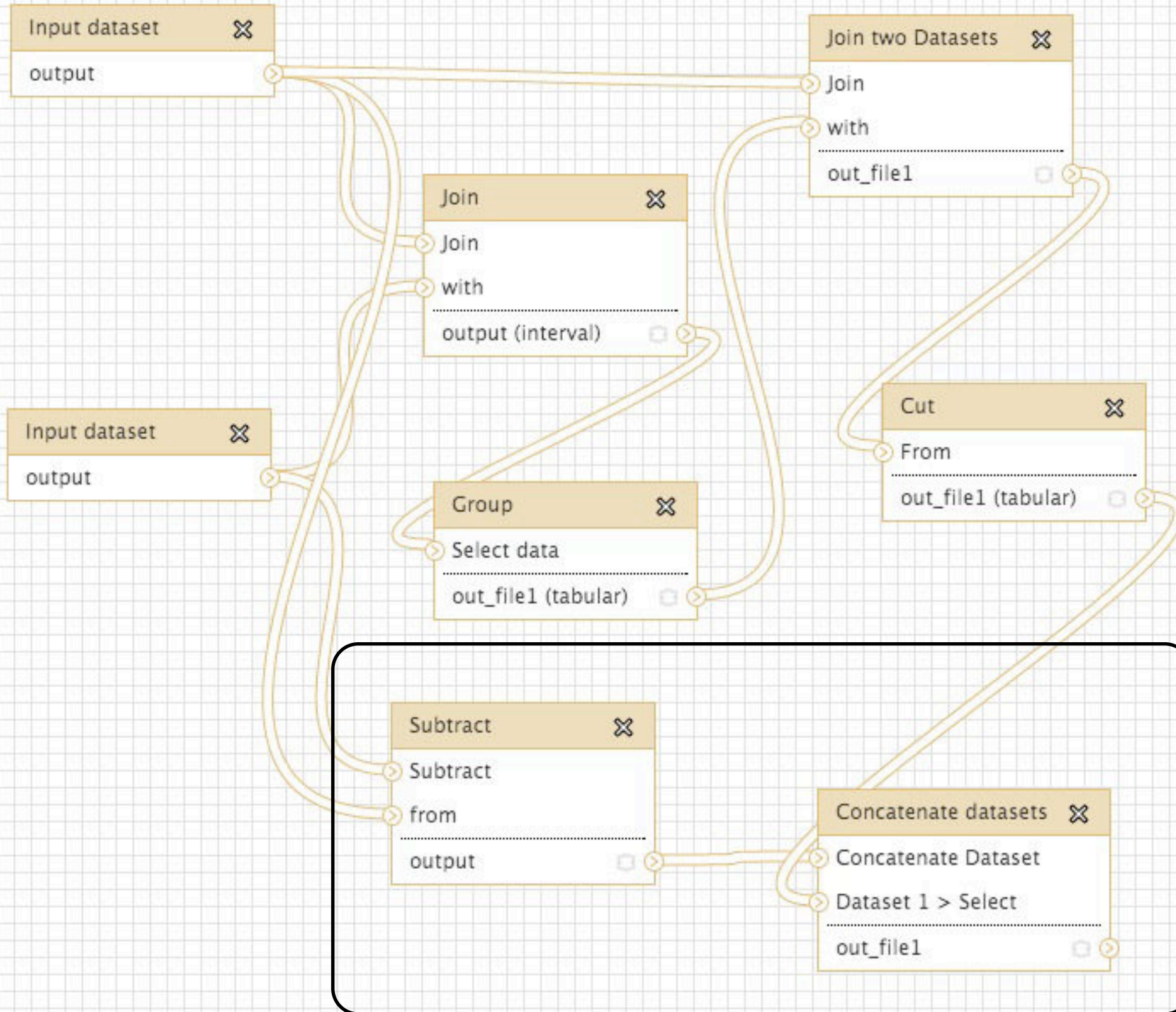
Published Histories, Pages, Workflows, about ...

Documentation on ...

Papers using Galaxy for ...

Related feature requests

One Possible Solution



Solution assumes starting with a score column value of 0.

Monday Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 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 generic *Overlap* Workflow

Extract Workflow from history

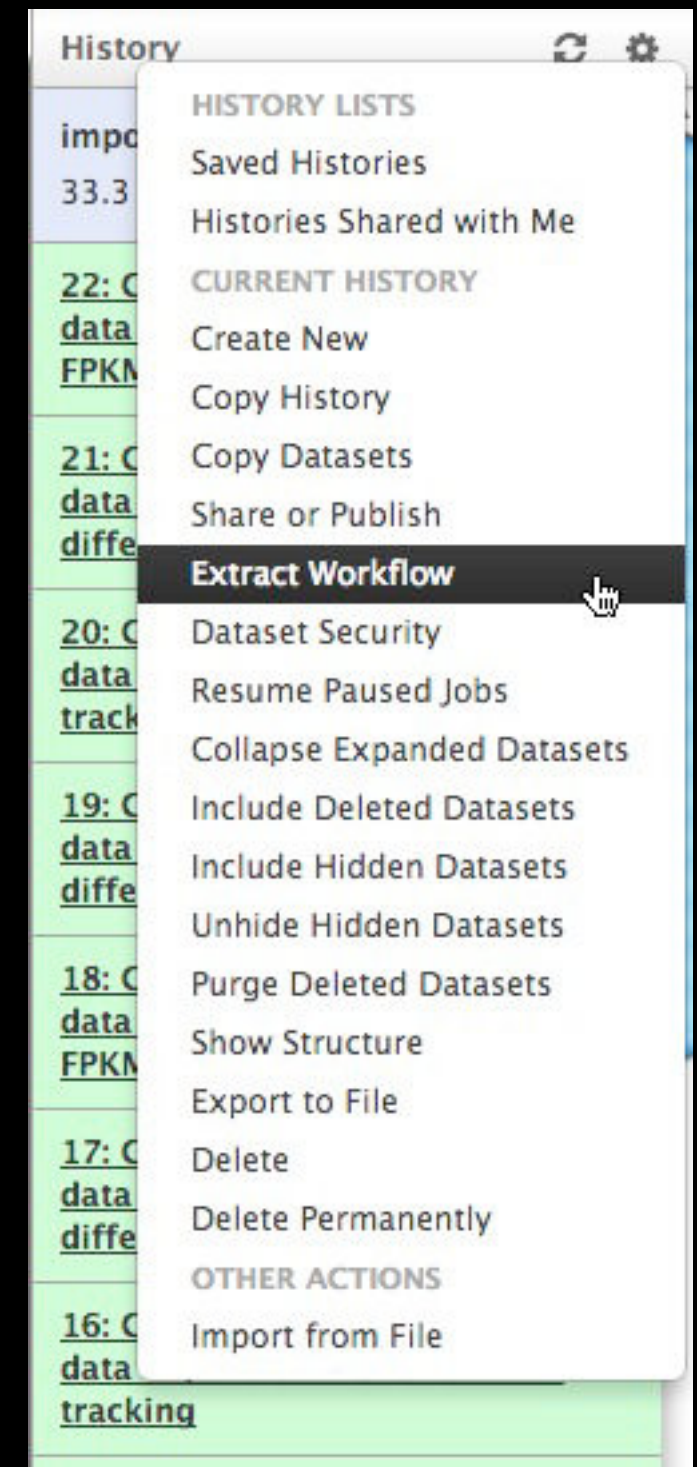
Create a workflow from this history.
Edit it to make some things clearer.

Run / test it

Guided: rerun with same inputs
Did that work?

On your own:

Count # of exons in each Repeat
Did that work? *Why not?*
Edit workflow: doc assumptions



Monday Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done

Monday Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done

Community: Local Galaxy Instances

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

Galaxy Tool Shed

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

The screenshot shows the Galaxy Tool Shed interface. On the left, there's a sidebar with 'Galaxy Tool Shed' and links to 'Browse by category', 'Browse all repositories', and 'Login to create a repository'. The main content area is titled 'Repository revision' and shows details for the 'clustalomega' repository. It includes a 'Repository revision' dropdown set to '2.bb1847455ec1', a 'repository tip' section, and a 'Detailed description' of the tool. Below this, there's a 'Preview tools and inspect metadata by tool version' section with a table of tools.

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

The screenshot shows the Galaxy Tool Shed interface with the 'Repositories' page. It features a search bar and a table of repositories. The table has columns for 'Name', 'Synopsis', 'Revision', 'Category', and 'Owner'. The repositories listed are 'abyss.toolsuite', 'asik.wrapper', 'asdf', 'assemblystats', and 'bam.to.bigwig'.

Name	Synopsis	Revision	Category	Owner
abyss.toolsuite	This suite contains Abyss and Abyss-PE config files and wrappers for Galaxy	0-92636934a189	Assembly	edward-kirton
asik.wrapper	Quickly match reads to a reference genome or sequence file	0-d6a426afa46	Next Gen Mappers Sequence Analysis	simon
asdf	asdf	-1-000000000000	Statistics Text Manipulation	vivek
assemblystats	Summarise an assembly (e.g. NSI metrics)	0-6544228ea290	Next Gen Mappers Sequence Analysis	konradpaszkiewicz
bam.to.bigwig	Generate BigWig coverage files from BAM files. Allows gapped reads to be split (useful for RNA-Seq). Calculates	5-5b40b93ebae3	Convert Formats SAM Visualization	leatsons

Community: Public Galaxy Instances

<http://bit.ly/gxyServers>

Interested in:

ChIP-chip and ChIP-seq?

✓ Cistrome

Statistical Analysis?

✓ Genomic Hyperbrowser

Protein synthesis?

✓ GWIPS-viz

de novo assembly?

✓ CBIIT Galaxy

Reasoning with ontologies?

✓ OPPL Galaxy

Repeats!

✓ RepeatExplorer

Everything?

✓ Andromeda

Plus many more

Galaxy Resources and Community: Mailing Lists

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

Galaxy-Announce

Project announcements, low volume, moderated

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

Galaxy-User

Questions about using Galaxy and usegalaxy.org

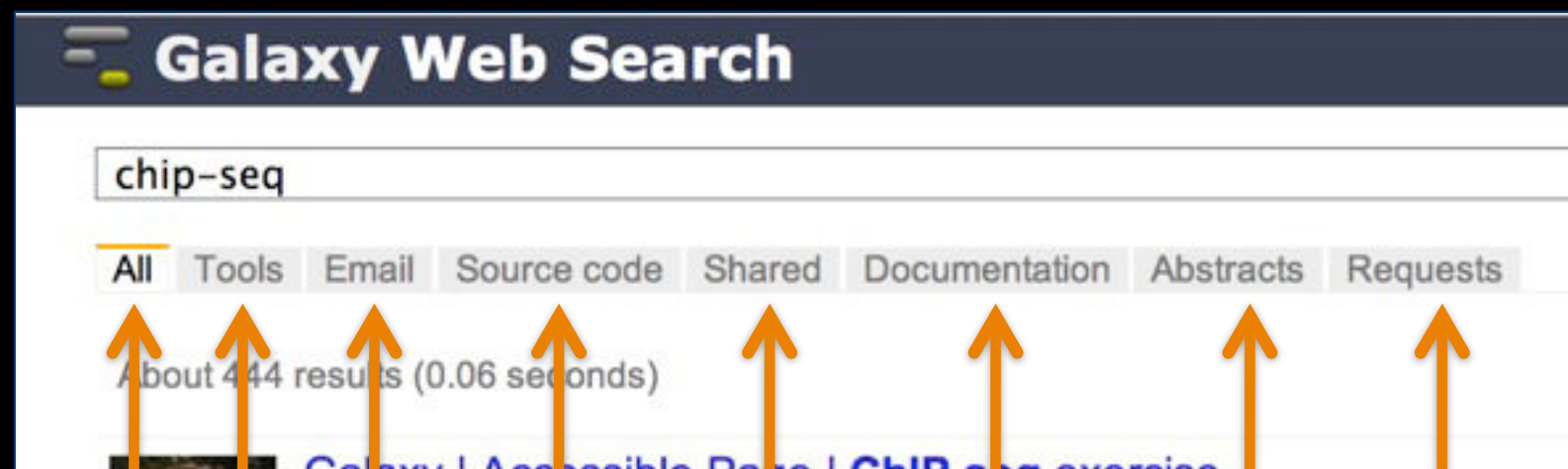
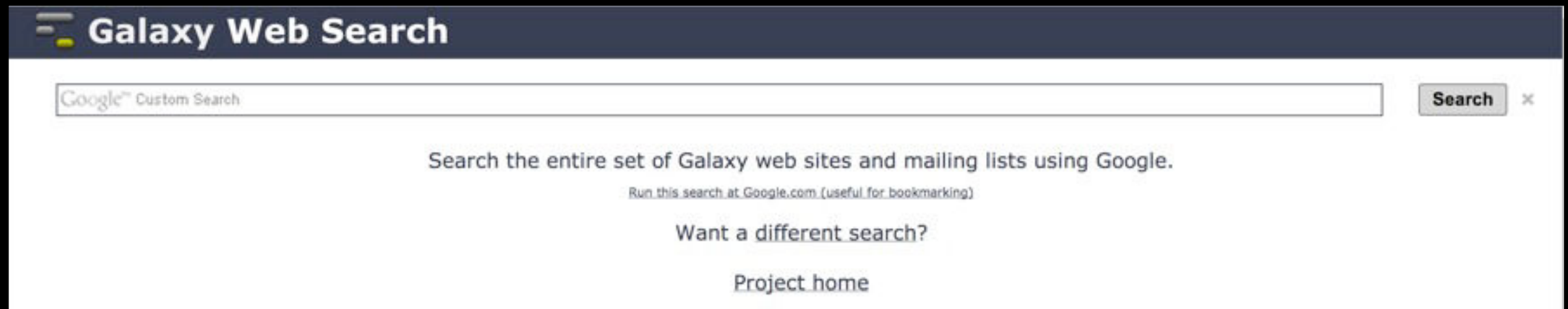
High volume (2900 posts in 2012, 2700+ members)

Galaxy-Dev

Questions about developing for and deploying Galaxy

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

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



Find

Everything on ...

Tools for ...

Email about ...

Source code for ...

Published Histories, Pages, Workflows, about ...

Documentation on ...

Papers using Galaxy for ...

Related feature requests

Community can create, vote and comment on issues

The screenshot shows a Trello board titled "Galaxy: Development Inbox" with a "Public" status. The board is organized into four main columns: "Inbox", "Developer ideas", "Bug Reports", and "Issues from Bitbucket".

- Inbox:** Contains five cards. The first card is a link to the Galaxy Project website. The second card is about a filter and sort issue. The third card is about a fastq file datatype issue. The fourth card is about a reference genome request. The fifth card is a feature request to manually hide datasets.
- Developer ideas:** Contains four cards. The first is about anonymous use of workflows. The second is a feature request to restart a failed workflow. The third is about Google Drive / Dropbox / Box integration. The fourth is a bug report about always importing deleted datasets.
- Bug Reports:** Contains four cards. The first is about workflow step hiding not persisting. The second is about a broken workflow view in Toolshed. The third is about being unable to run jobs when user job limits are set. The fourth is about a bug when using data_column.
- Issues from Bitbucket:** Contains four cards. The first is about an option to disable automatic history creation. The second is about an option to require that histories have names. The third is about more flexible output handlers. The fourth is about allowing overriding parameters when running a workflow.

On the right side of the board, there is a "Members" section showing a grid of member avatars, an "Add Members..." button, and a "Board" section with "Options", "Add List", and "Filter Cards" buttons. Below these is an "Activity" section showing a list of recent actions, including "Dannon Baker added API: Library Contents to Developer ideas and" and "g2roboto on Feature request: manually hide datasets".

<http://bit.ly/gxyissues>

http://wiki.galaxyproject.org


Galaxy Wiki

DaveClements Settings Logout | Search:

Titles Text

FrontPage

Edit History Actions



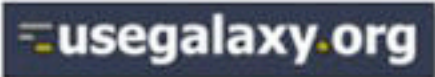
Galaxy is an open, web-based platform for *accessible, reproducible, and transparent* computational biomedical research.

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

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

Use Galaxy


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



Deploy Galaxy

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

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




Community & Project

Galaxy has a large and active user community and many ways to [Get Involved](#).

- [Community](#)
- [News](#)
- [Events](#)
- [Support](#)
- [Galaxy Project](#)

Contribute


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



Use Galaxy


[Use Main \(about\)](#)
[Use Others!](#) • [Learn](#)
[Share](#) • [Search](#)

Communication

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

Deploy Galaxy

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



Contribute

[Tool Shed](#) • [Share](#)
[Issues & Requests](#)
[Support](#)

Galaxy Project

[Home](#) • [About](#)
[Community](#)
[Big Picture](#)

Wiki

[Help](#) • [All Pages](#)

Events

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 that are relevant to the Galaxy Community. This is also available as an [RSS feed](#).

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

Upcoming Events



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

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

New CloudMan Release

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

IMPORTANT - please read

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

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

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

For more details on the new features, see the [CHANGELOG](#) and for even more details see, [all 291 commit messages from 7 contributors](#).

Enjoy and please let us know what you think,

Enis Afgan

Posted to the Galaxy News on 2013-07-08

SlipStream Appliance: Galaxy Edition

News Items

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

[News Archive](#)





GALAXY

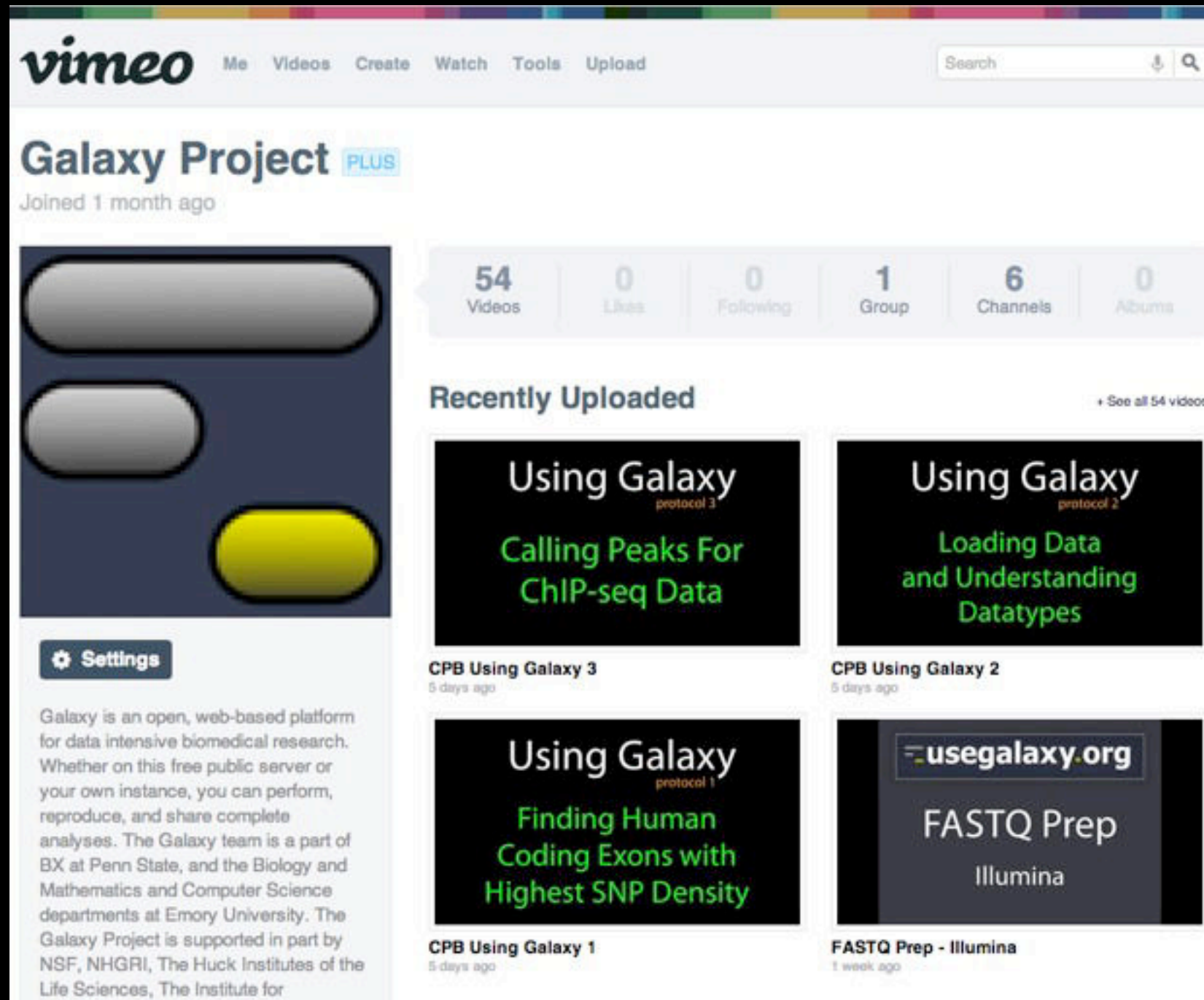
COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

<http://bit.ly/gcc2014>



Galaxy Resources & Community: Videos



“How to”
screencasts on
using and
deploying Galaxy

Talks from previous
meetings.

<http://vimeo.com/galaxyproject>

Galaxy Resources & Community: CiteULike Group



The screenshot shows the CiteULike interface for the 'Galaxy' group. At the top, the 'citeulike' logo is visible. Below it, a navigation bar includes 'CiteULike', 'Group: Galaxy', and links for 'Search', 'Register', and 'Log in'. The main content area is titled 'Group: Galaxy - library 1181 articles' and includes buttons for 'Search', 'Copy', 'Export', 'Sort', and 'Hide Details'. Below these buttons are sorting options: 'Sort by: Group rating', 'Order: Default', and 'Empty fields: Default'. The article list shows three entries, each with a checkmark, a title, a citation, and a link to the abstract. The sidebar on the right, titled 'Group Tags', lists various tags associated with the group, including 'cloud', 'howto', 'isgalaxy', 'methods', 'other project', 'republic', 'reproducibility', 'shared tools', 'unknown', 'usecloud', 'uselocal', 'usemain', 'usepublic', 'visualization', and 'workbench'.

Group: Galaxy - library 1181 articles

Search Copy Export Sort Hide Details

Sort by: Group rating Order: Default Empty fields: Default

✓ **The map-based sequence of the rice genome**
Nature, Vol. 436, No. 7052. (11 August 2005), pp. 793-800, [doi:10.1038/nature03895](https://doi.org/10.1038/nature03895)
by [Sequencing Project International Rice Genome](#)
posted to [workbench](#) by [galaxyproject](#) to the group [Galaxy](#) on 2011-12-15 17:59:32 ★★ [along with](#)
■ Abstract

✓ **Galaxy: A platform for interactive large-scale genome analysis**
Genome Research, Vol. 15, No. 10. (01 October 2005), pp. 1451-1455, [doi:10.1101/gr.4086505](https://doi.org/10.1101/gr.4086505)
by [Belinda Giardine](#), [Cathy Riemer](#), [Ross C. Hardison](#), et al.
posted to [project](#) by [galaxyproject](#) to the group [Galaxy](#) on 2011-12-02 22:40:23 ★★★★★ [along with](#)
■ Abstract

✓ **Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques**
Genome Research In *Genome Research*, Vol. 16, No. 12. (01 December 2006), pp. 1455-1464, [doi:10.1101/gr.4140006](https://doi.org/10.1101/gr.4140006)
by [Laura Elnitski](#), [Victor X. Jin](#), [Peggy J. Farnham](#), [Steven J. M. Jones](#)
posted to [workbench](#) by [galaxyproject](#) to the group [Galaxy](#) on 2013-01-10 07:48:08 ★★ [along with 18 people and 2 groups](#)
■ Abstract

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

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

Almost
1200
papers

17
different
tags

<http://bit.ly/gxycul>

Share & Publish: More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Sharing & Publishing enables **Reproducibility**

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)



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 analyzed through Galaxy using the shown 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

The Galaxy Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Marten Cech



Dave Clements



Nate Coraor



Carl Eberhard



Dorine Francheteau



Jeremy Goecks



Sam Guerler



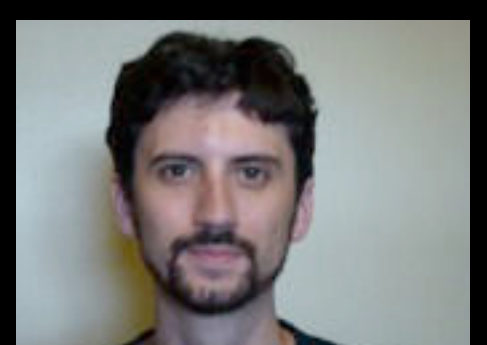
Jen Jackson



Greg von Kuster



Ross Lazarus



Nick Stoler



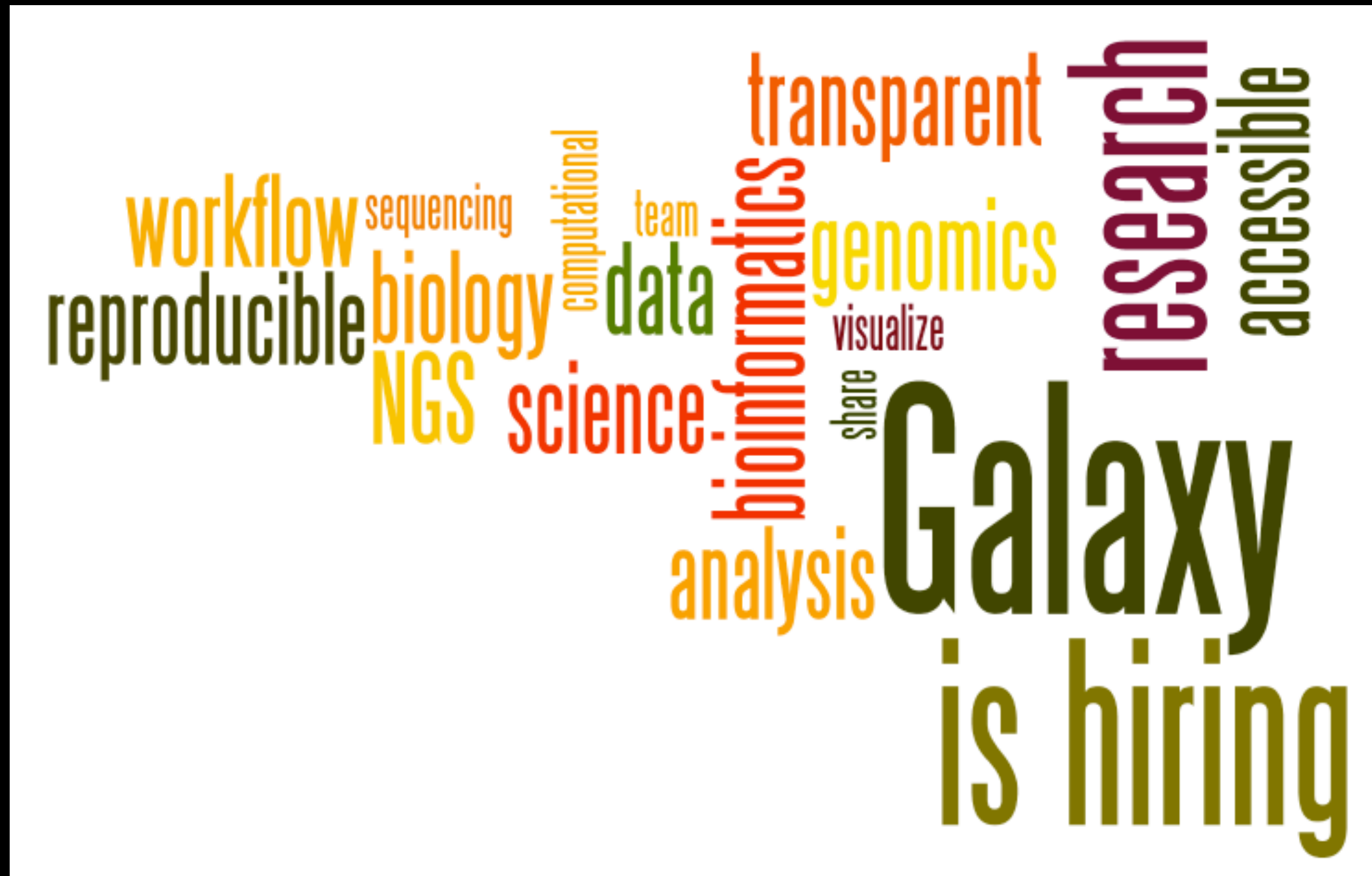
Anton Nekrutenko

James Taylor



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

Galaxy is hiring post-docs and software engineers



Please help.

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

Monday Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done

NGS Data Quality Control

- Introduce FASTQ format
- Analyze an RNA-Seq dataset
- Trim as we see fit.

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

→ UC Davis RNA-Seq Human*

→ 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

Options 1 & 2:

1. NGS QC and Manipulation → **Compute Quality Statistics**

NGS QC and Manipulation → **Draw quality score boxplot**

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

2. NGS QC and Manipulation → **FastQ Summary Statistics,**

Graph / Display Data → **Boxplot of quality statistics**

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

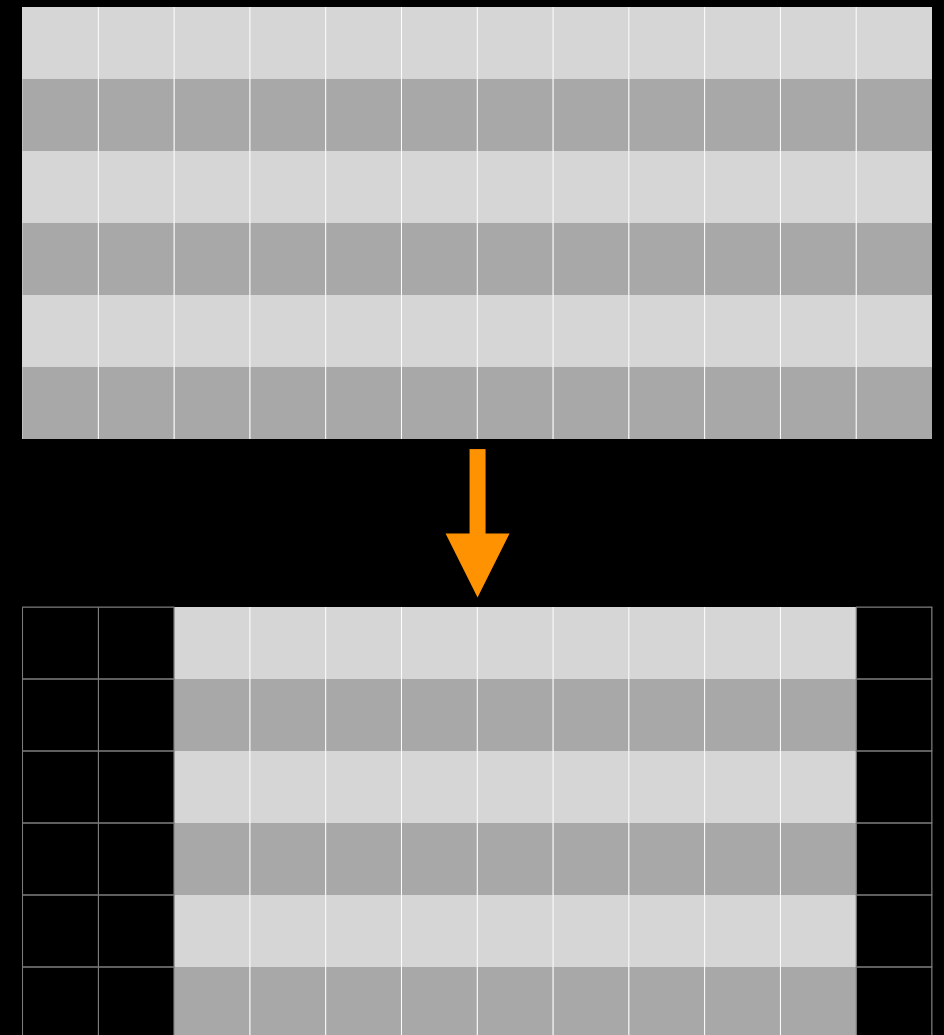
NGS Data Quality: Assessment tools

- Option 3
 - NGS QC and Manipulation → **FastQC**
 - Gives you a lot a lot more information but little control over how it is calculated or presented.

<http://bit.ly/FastQCBoxPlot>

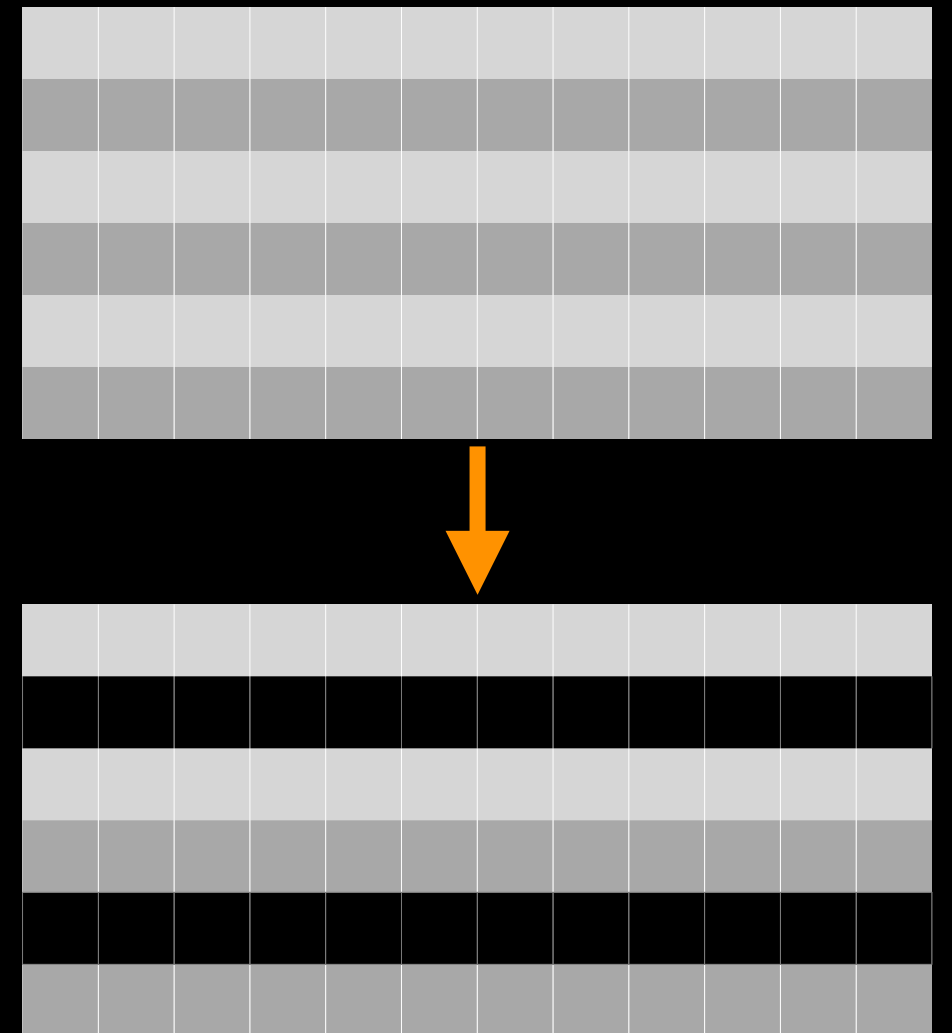
NGS Data Quality: Base Quality Trimming

- 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



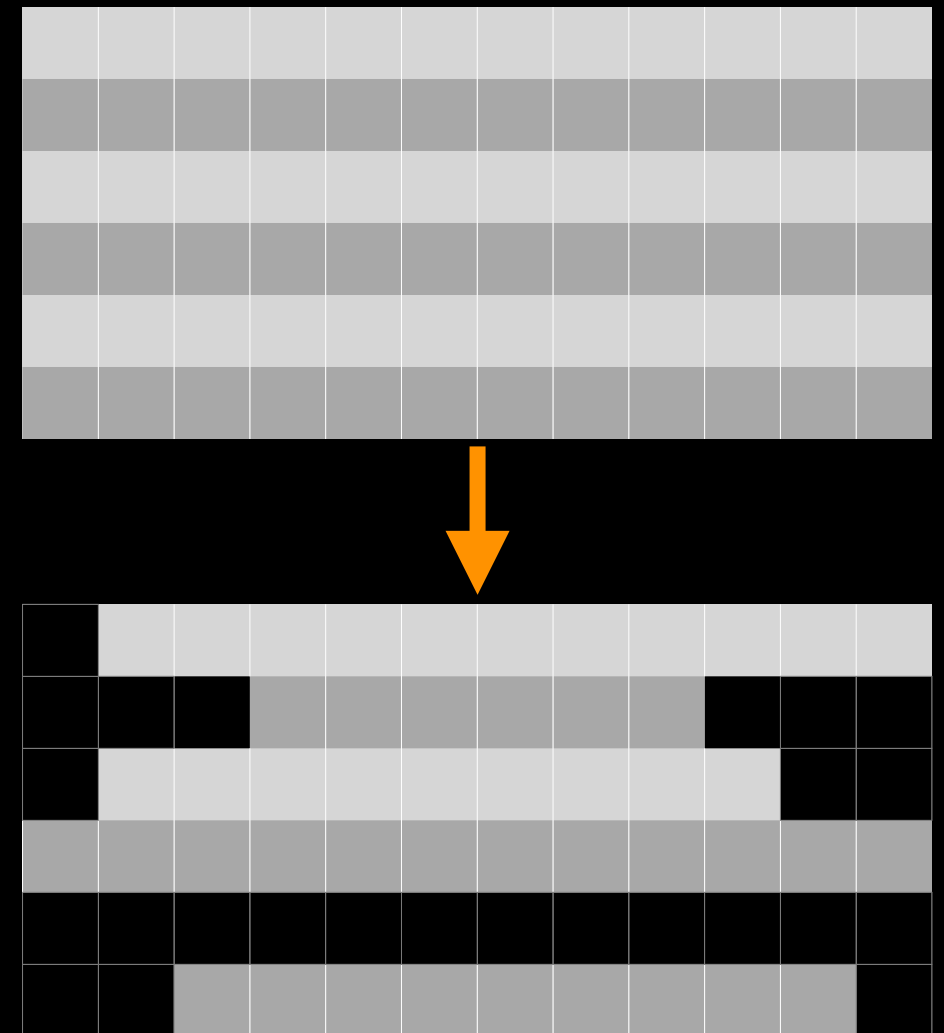
NGS Data Quality: Base Quality Trimming

- ~~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.

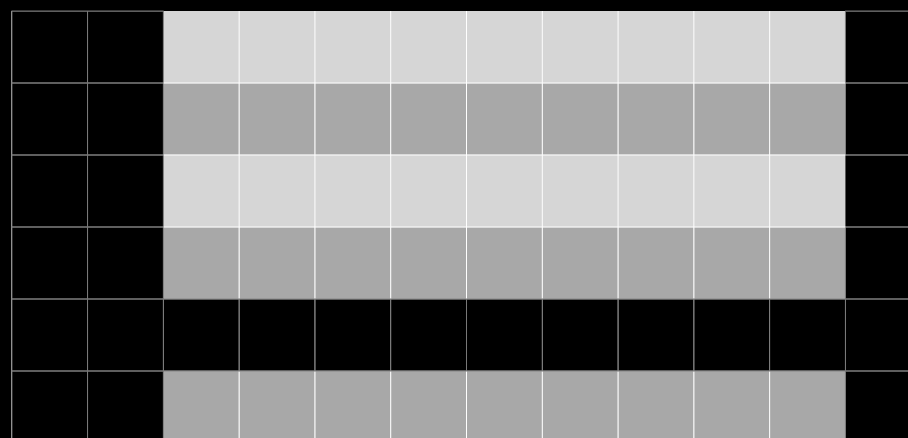
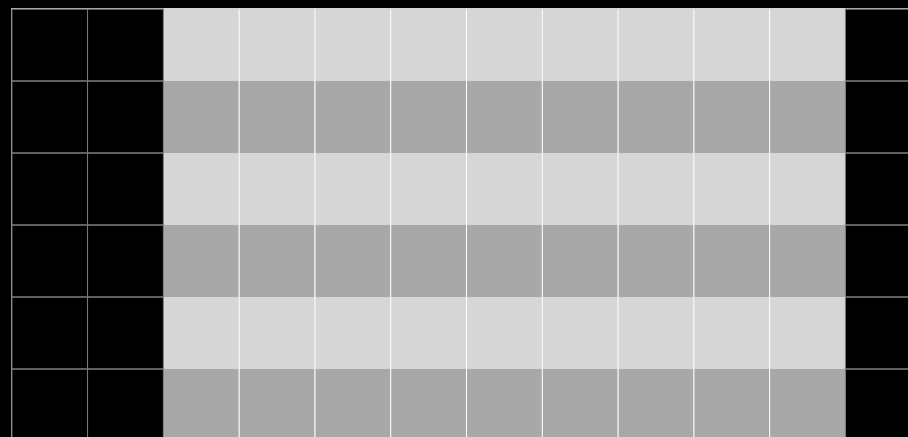
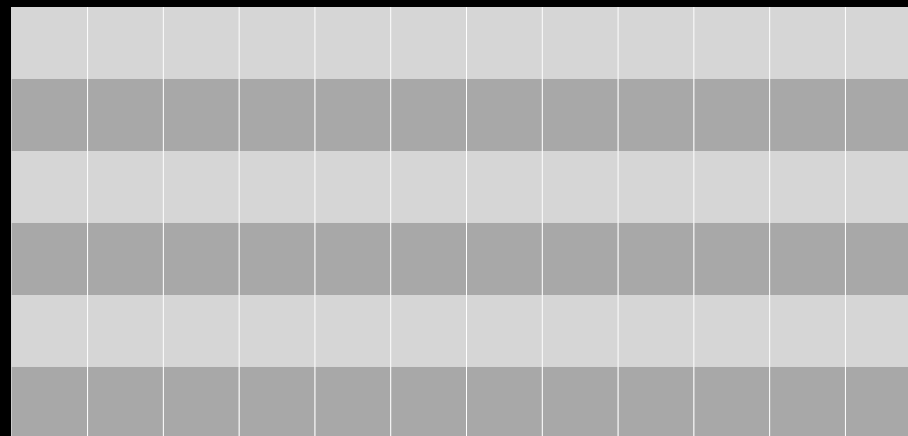


NGS Data Quality: Base Quality Trimming

- 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

+

Option 2

Trim? *As we see fit?*

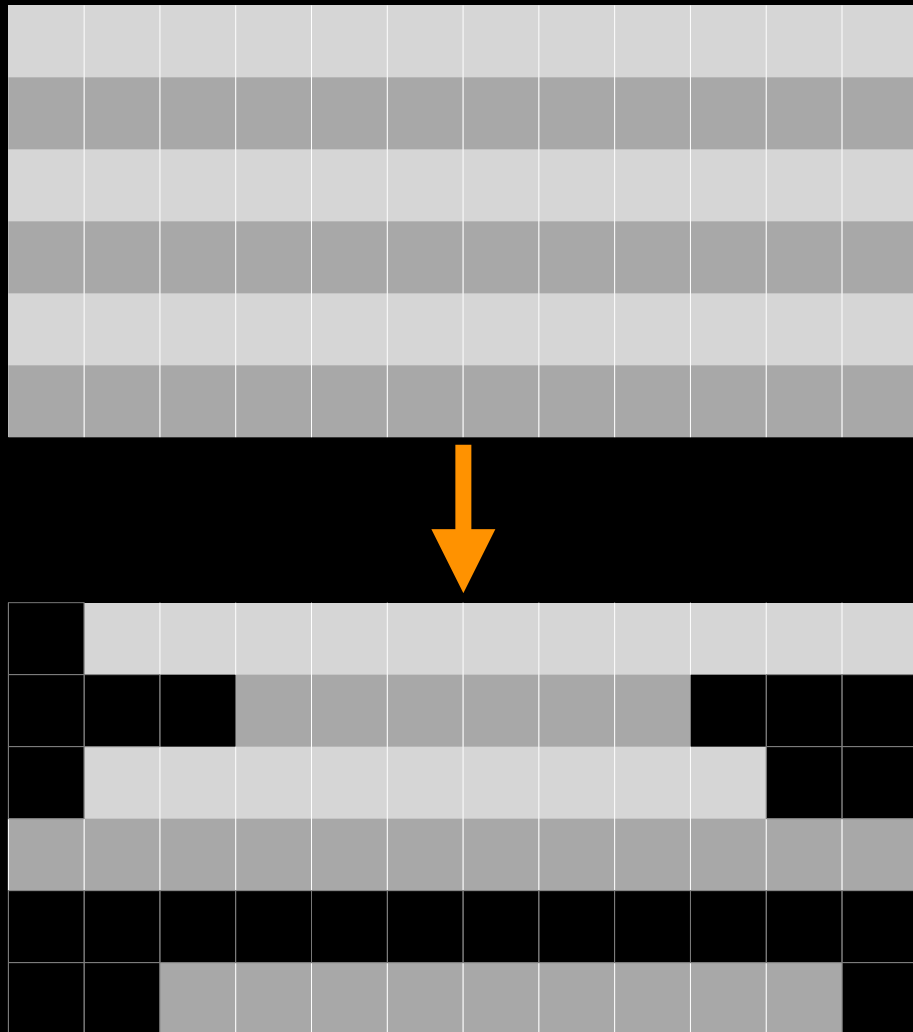
- Introduced 3 options
 - One preserves original read length, two don't
 - One preserves number of reads, two don't
 - Two keep/make every read the same length, one does not
- One preserves pairings, two don't
 - Can also trim aggressively and then restore pairings

Trim? *As we see fit?*

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



NGS Data Quality: Base Quality Trimming



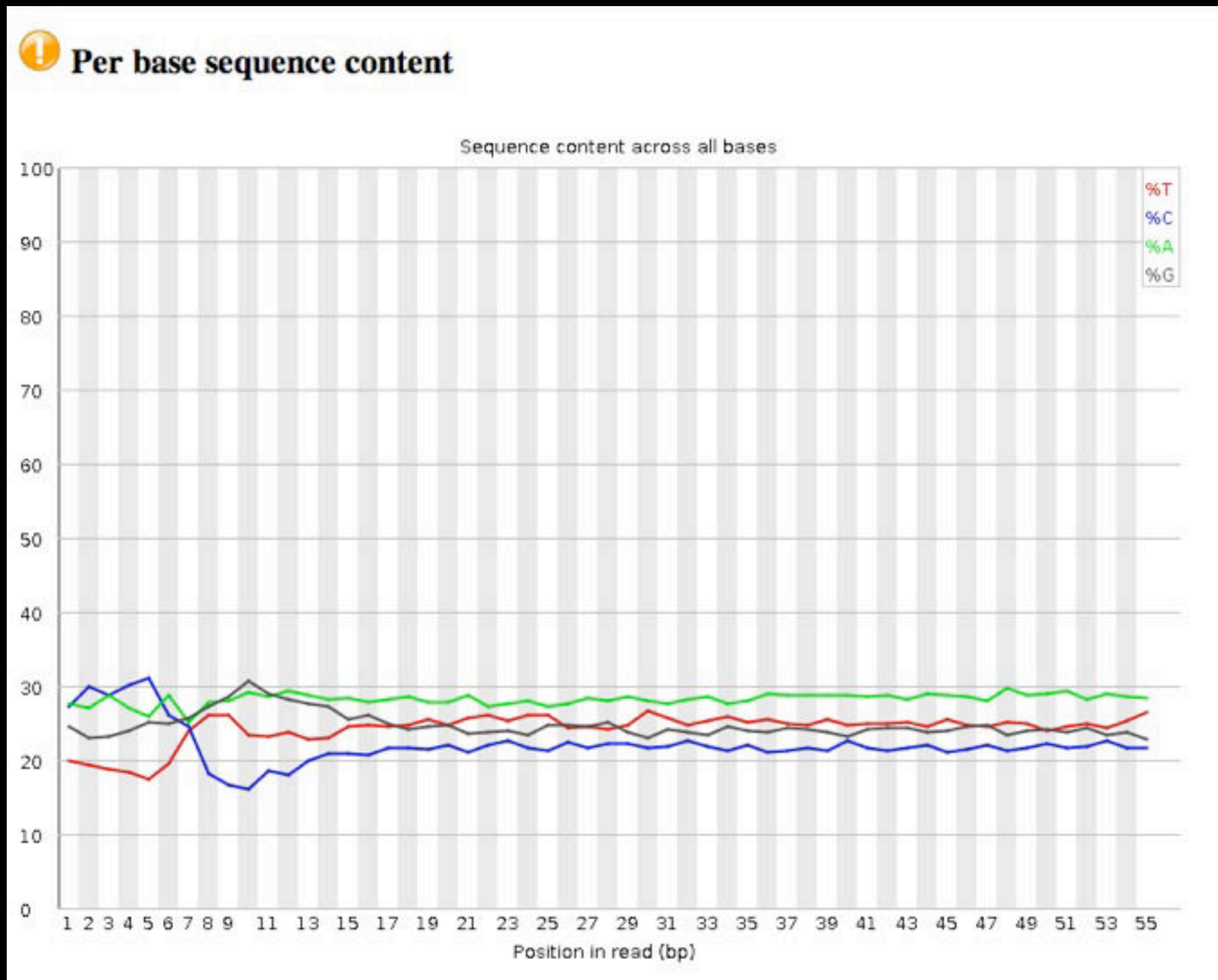
I'll use Option 3

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

Run

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

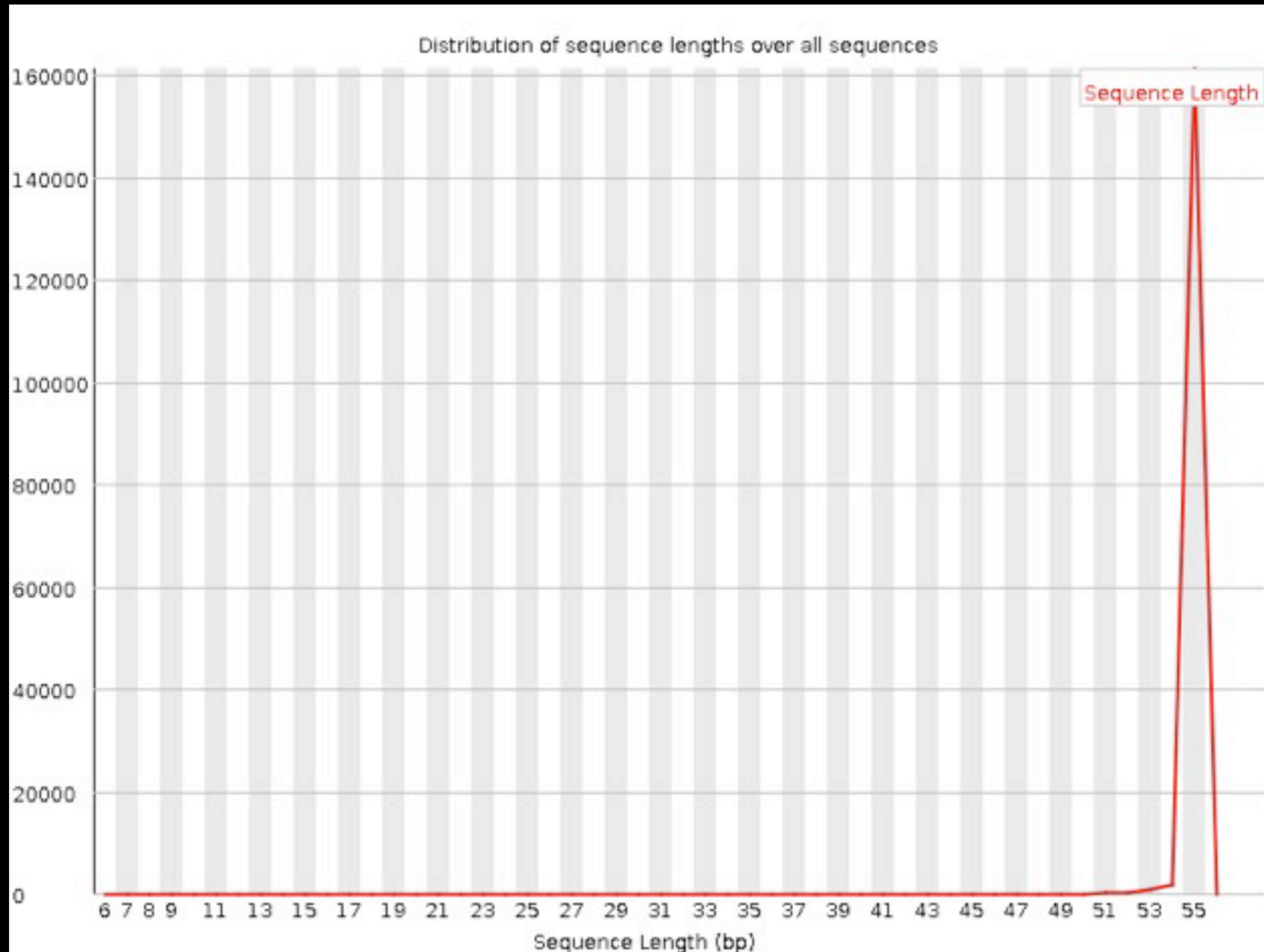
NGS Data Quality: Sequence bias as front of reads?



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: Base Quality Trimming



New Problem:

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

Option 2!


NGS QC and Manipulation →

Filter FASTQ reads by quality score and length

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

NGS Data Quality: Sequencing **Artifacts**

Repeat this process with MeOH Rep R2 (the reverse reads)
... and there's a new 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**

NGS Data Quality: Done with 1st Replicate!

Now, only ~~5~~ 3 more to go...

Exercise:

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

Create a workflow that runs a single FASTQ file through all the quality steps.

Or ...

Create a workflow that runs a pair of FASTQ files through all the quality steps

NGS Data Quality: Restoring Pairings

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

“Dang.”

Dave C

Is that really true? 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.

Run the “**Re-pair paired end reads after QC may have deleted some of them**” workflow on each set of paired end reads.

Each workflow run takes the raw and trimmed versions of the forward and reverse reads for each replicate.

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.*

Monday Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done



Agenda

- 8:30 Welcome and Intro
- 9:10 Basic Analysis with Galaxy
- 10:30 Break + an Exercise
- 11:30 Basic Analysis into Reusable Workflows
- 12:30 Lunch
- 13:30 Galaxy Community
- 14:00 NGS Data Quality Control
- 15:00 Break
- 16:30 Done

Thanks



Dave Clements

**Galaxy Project
Emory University**

clements@galaxyproject.org

Community: Further reading & Resources

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

<http://bit.ly/gxyServers>

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

<http://galaxyproject.org/search>

<http://bit.ly/gxyissues>

<http://bit.ly/gcc2014>

<http://vimeo.com/galaxyproject>

<http://bit.ly/gxycul>