

Introduction to Galaxy

Northwestern University
February 27, 2015

Dave Clements
Galaxy Project
Johns Hopkins University

Matt Schipma
NGS Core Facility
Center for Genetic Medicine (CGM)
Northwestern University



Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done

Goals

Provide a basic introduction to using Galaxy for bioinformatic analysis.

Demonstrate how Galaxy can help you explore and learn options, perform analysis, and then share, repeat, and reproduce your analyses.

Not Goals

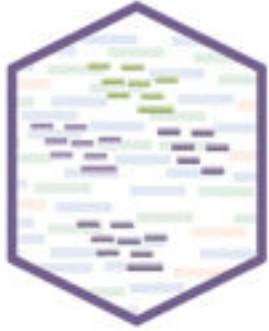
This workshop will *not* cover

- details of how tools are implemented, or
- new algorithm designs, or
- which assembler or mapper or peak caller or ... is best for you.

This workshop does cover RNA-Seq. However, our emphasis today is on learning general NGS and Galaxy principles.

It will still be enough to get you started with RNA-Seq.

Want more RNA-Seq ?



Transcriptome Analysis with RNA-seq -- Mar. 19, 2015

Speaker: Matthew Schipma, NUCATS Next-Generation Sequencing Core

[Direct registration link](#)

Next generation sequencing has made it possible to sequence the entire transcriptome of a cell or tissue type. This allows researchers to evaluate the expression of each gene in the genome, including specific isoforms. RNA-seq also has the potential to identify novel splice variants and even novel genes. The ability to gain useful information from an RNA-seq project depends of several experimental factors including the number of replicates and the sequencing depth. Here we will discuss the best practices in designing an RNA-seq experiment as well as the information that can be obtained through RNA-seq.

Part of the
Computational Skills for Informatics Seminar Series

[http://www.galter.northwestern.edu/News/
computational-skills-for-informatics-seminar-series](http://www.galter.northwestern.edu/News/computational-skills-for-informatics-seminar-series)

<http://bit.ly/CSIGalter>

What is Galaxy?

Data integration and analysis platform that emphasizes accessibility, reproducibility, and transparency

A free (for everyone) web server

Open source software

These options result in several ways to use Galaxy

<http://galaxyproject.org>

Galaxy is available online, for free

<http://usegalaxy.org>

As a free (for everyone) web server integrating a wealth of tools, compute resources, petabytes of reference data and permanent storage



However, *a centralized solution cannot support the different analysis needs of the entire world.*

Galaxy is available as Open Source Software

Galaxy is installed in locations around the world.

Some of them are free for anyone to use too.

<http://getgalaxy.org>

bit.ly/gxyServers

Galaxy is available on the Cloud



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

<http://globus.org/>

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

We are using the cloud today.

Galaxy is available **with Commercial Support**

A ready-to-use appliance
(BioTeam)

Cloud-based solutions
(ABgenomica, AIS,
GenomeCloud)

Consulting & Customization
(BioTeam, Deena
Bioinformatics)

Training
(OpenHelix)



Galaxy Project: Further reading & Resources

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

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

<http://bit.ly/gxychoices>

Agenda

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:00 Basic Analysis into Reusable Workflows

11:30 RNA-Seq Example Part I

12:20 Lunch (on your own)

1:35 RNA-Seq Example Part II

2:45 Break

3:00 Sharing, Publishing, and Reproducibility

3:25 Setting up your own Galaxy Cluster on Amazon

4:00 Done

Basic Analysis

Which exons have most overlapping
Repeats?

Use Human, HG19, Chromosome 22

cloud1.galaxyproject.org

cloud2.galaxyproject.org

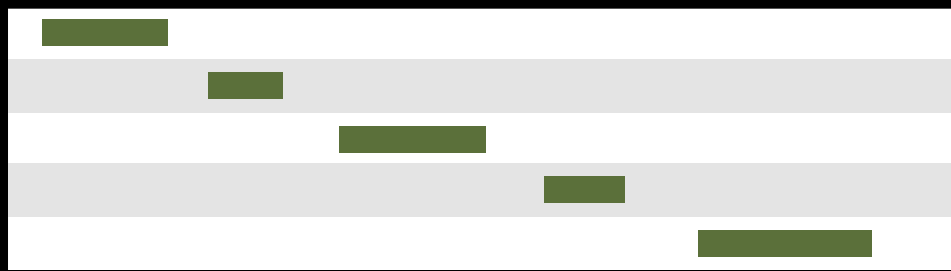
cloud3.galaxyproject.org

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

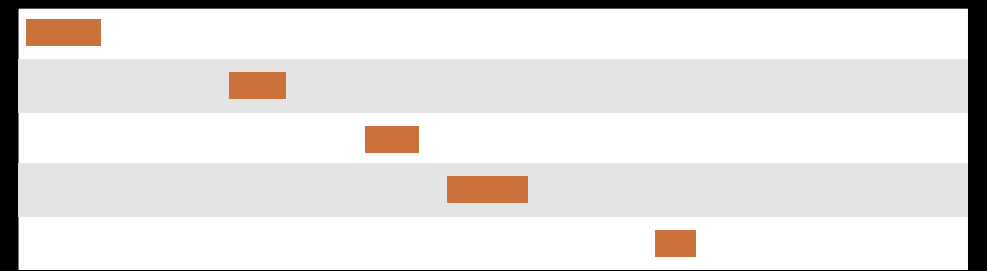
Exons & Repeats: A General Plan

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

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

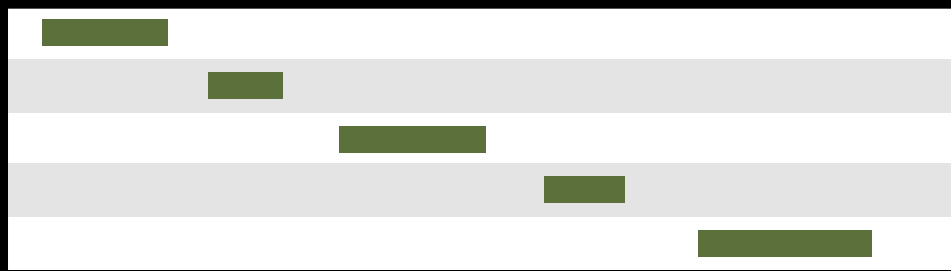


Exons

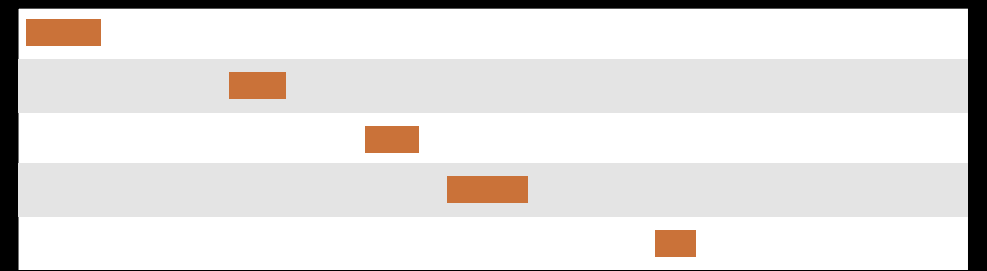


Repeats

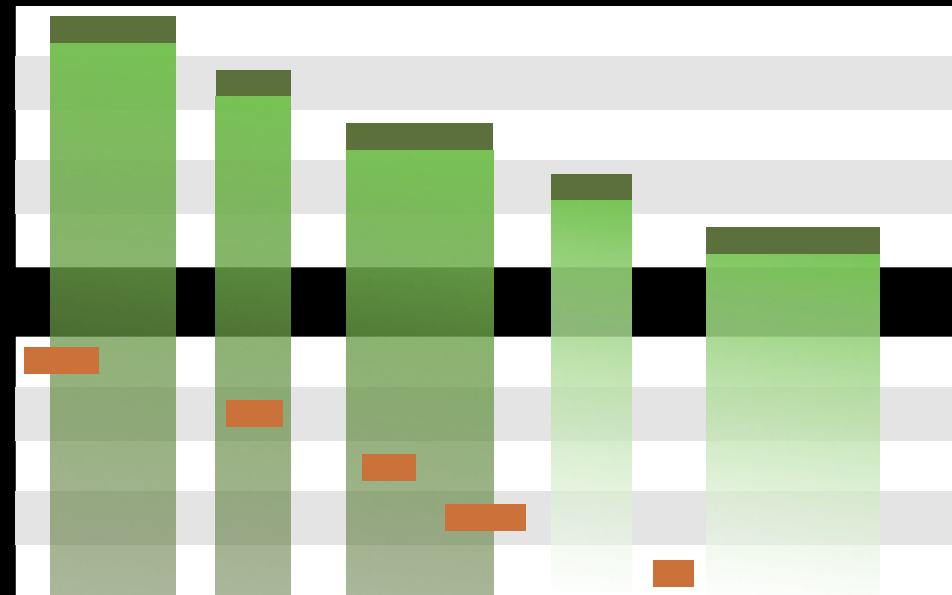
(Identify which exons have Repeats)



Exons



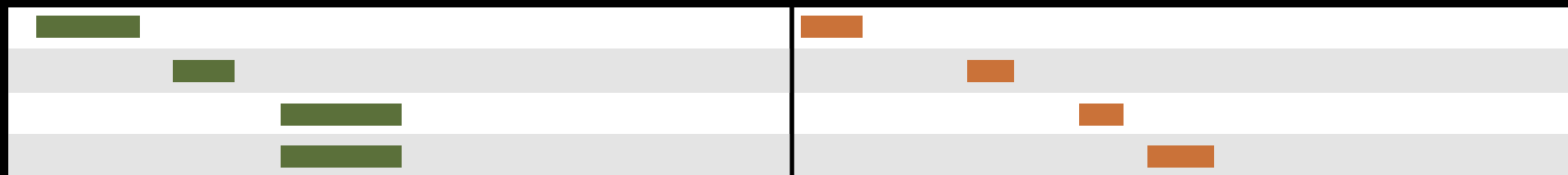
Repeats



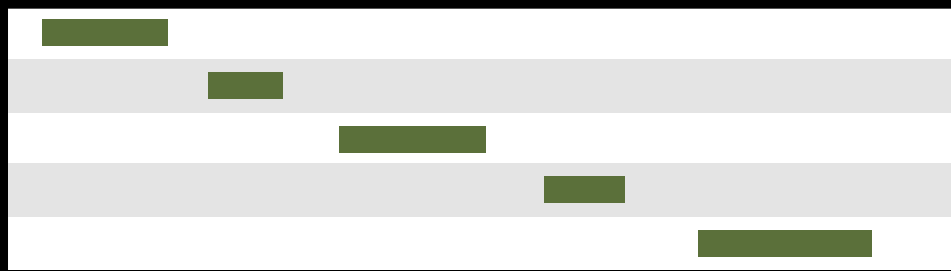
Exons

Repeats

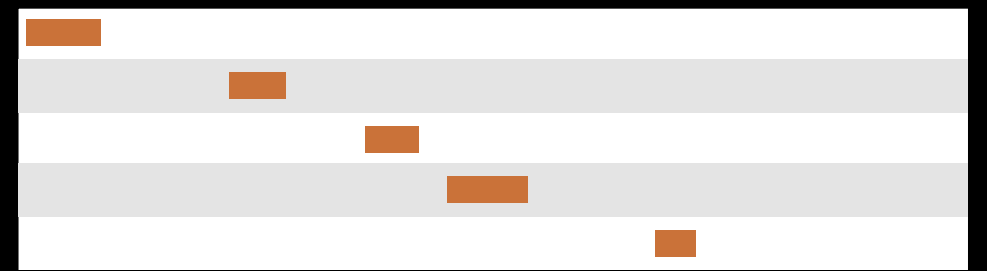
Overlap pairings



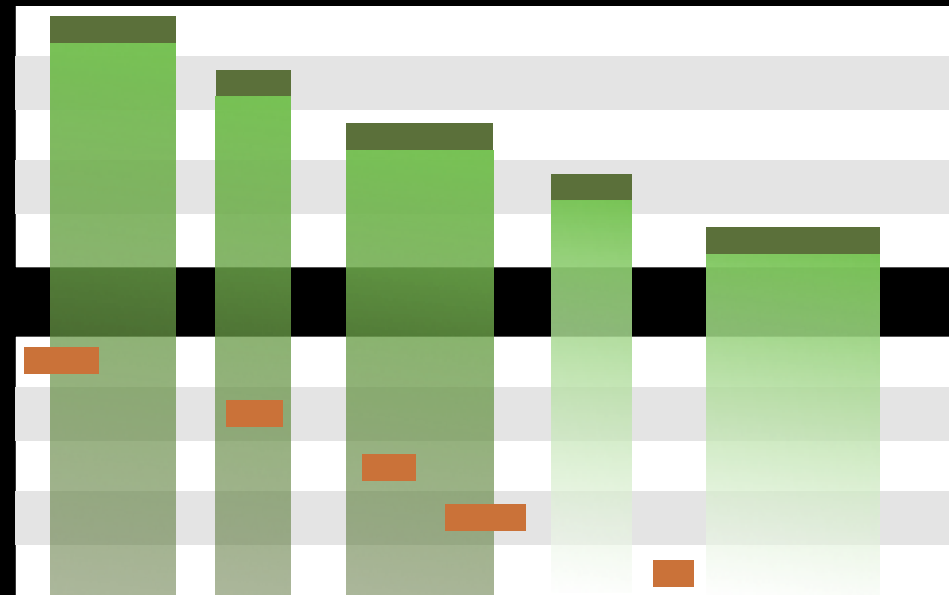
Operate on Genomic Intervals → Join
(Identify which exons have Repeats)



Exons



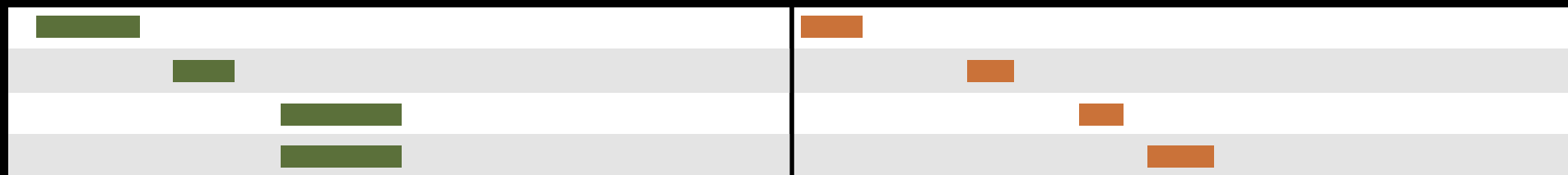
Repeats



Exons

Repeats

Overlap pairings

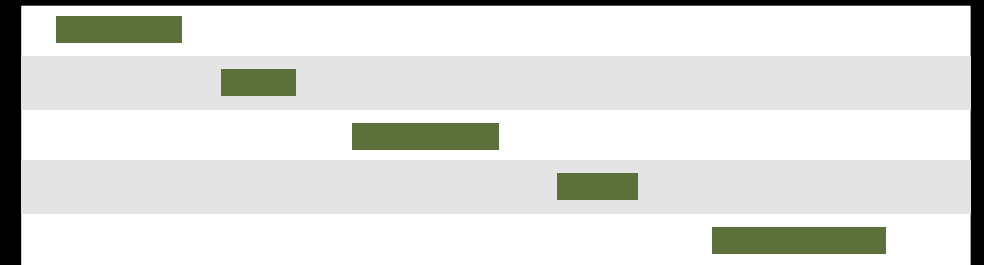


Exon overlap counts

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

	1
	1
	2

Exon overlap counts

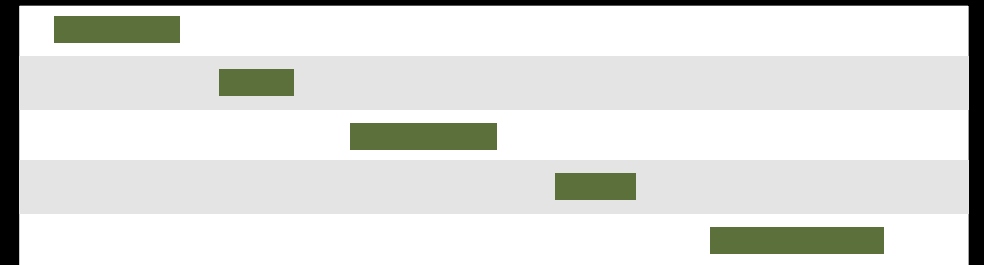


Exons

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
1	1
2	2

Exon overlap counts

Response	Percentage
Yes, the current government is responsible	45%
No, the current government is not responsible	55%

Exons

The diagram illustrates a network flow problem with a source node s and a sink node t . A red path from s to t is shown. A green line represents a cut, separating the source side (S) from the sink side (T). The cut is labeled "Real cut".

Join on exon name

Rearrange columns w/ cut

Text Manipulation → Cut

(Incorporate the overlap count with rest of Exon information)

Exons & Repeats: Exercise


Include 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 Exon-Repeats exercise.

Tools 

search tools 

[Get Data](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)
[Convert Formats](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Get Genomic Scores](#)
[Statistics](#)
[Graph/Display Data](#)
[Evolution](#)
[Motif Tools](#)
[NGS: QC and manipulation](#)
[NGS: Mapping](#)
[NGS: SAM Tools](#)
[NGS: Simulation](#)
[Phenotype Association](#)

 **Obrigado! Welcome to Galaxy on the Nuve**

- Data Libraries
- Data Libraries Beta
- Published Histories** 
- Published Workflows
- Published Visualizations
- Published Pages







Paulo

Galaxy is an open, web-based platform for data intensive biomedical research. 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 NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

101: Overlapping Exons and Repeats

3.5 MB

search datasets

Dataset		Annotation
<u>1: Exons, chr22</u>		
<u>2: Repeats, chr22</u>		
<u>3: Join on data 2 and data 1</u>		
<u>4: Group on data 3</u>		
<u>5: Join two Datasets on data 1 and data 4</u>		
<u>6: Exons with overlapping repeats</u>		

Make a copy of this history and switch to it

Autho

outreac

Relate

All publ
Publishe

Rating

Commu
(0 ratings)

Tags

Commu

Note: In your solution, you can take advantage of the fact that Exons already have 0 scores.

Agenda

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:00 Basic Analysis into Reusable Workflows

11:30 RNA-Seq Example Part I

12:20 Lunch (on your own)

1:35 RNA-Seq Example Part II

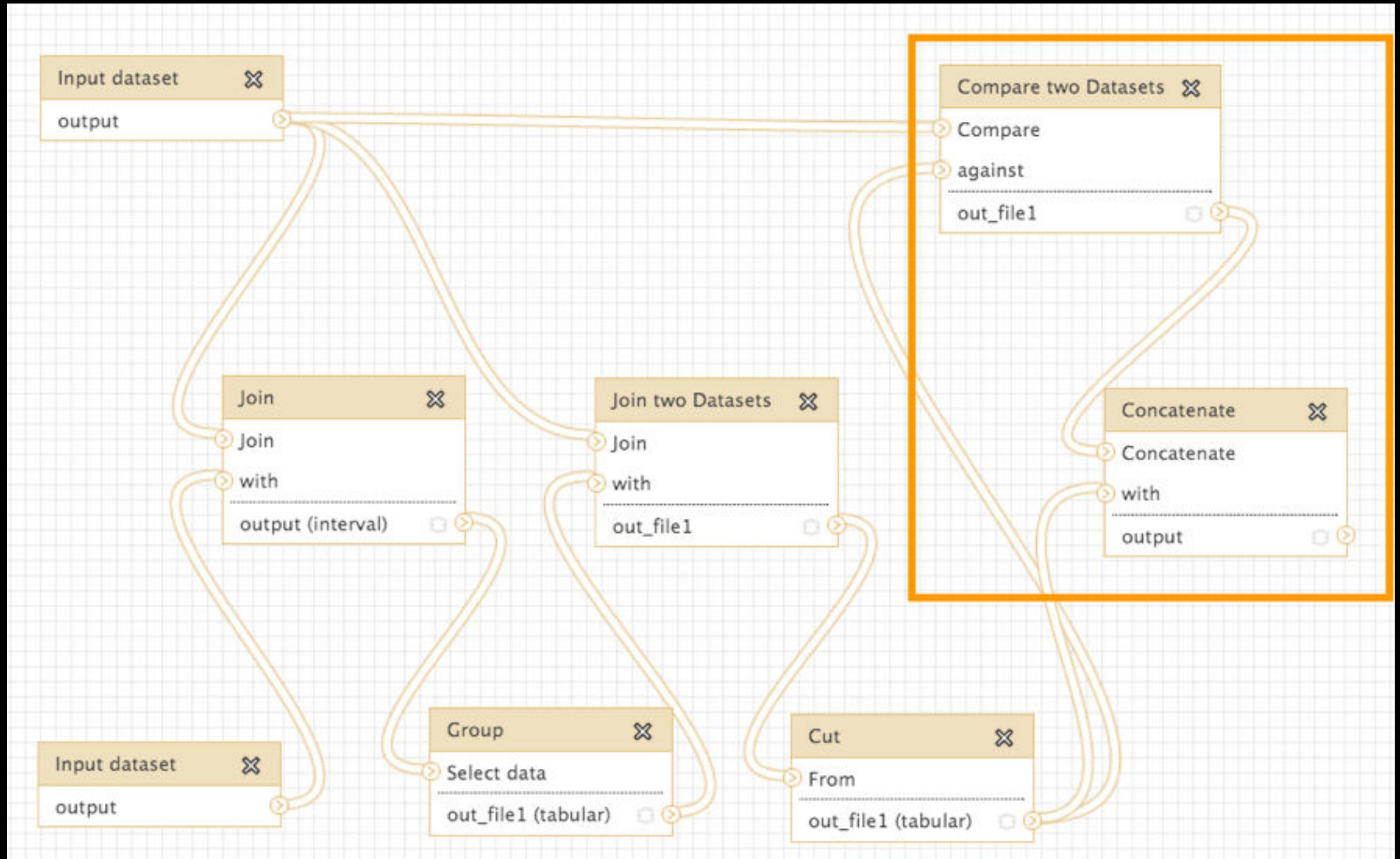
2:45 Break

3:00 Sharing, Publishing, and Reproducibility

3:25 Setting up your own Galaxy Cluster on Amazon

4:00 Done

One Possible Solution



Solution from Stanford Kwenda and Caron Griffiths, Pretoria.
Takes advantage of the fact that Exons already have 0 scores.

Agenda

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:00 Basic Analysis into Reusable Workflows

11:30 RNA-Seq Example Part I

12:20 Lunch (on your own)

1:35 RNA-Seq Example Part II

2:45 Break

3:00 Sharing, Publishing, and Reproducibility

3:25 Setting up your own Galaxy Cluster on Amazon

4:00 Done

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

A series of inputs, analysis steps, intermediate datasets, and outputs

Workflow:

A series of analysis steps

Can be repeated with different data

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

- The analysis we just finished was about
 - Human chr22
 - Overlap between exons and Repeats
- But, ...
 - there is **nothing inherent** in the analysis **about humans, exons or repeats**
 - It is a series of steps that **sets the score of one set of features to the number of overlaps from another set of features.**

Create a Workflow from a History

Extract Workflow from history

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



(cog) → Extract Workflow

Run / test it

Guided: rerun with same inputs

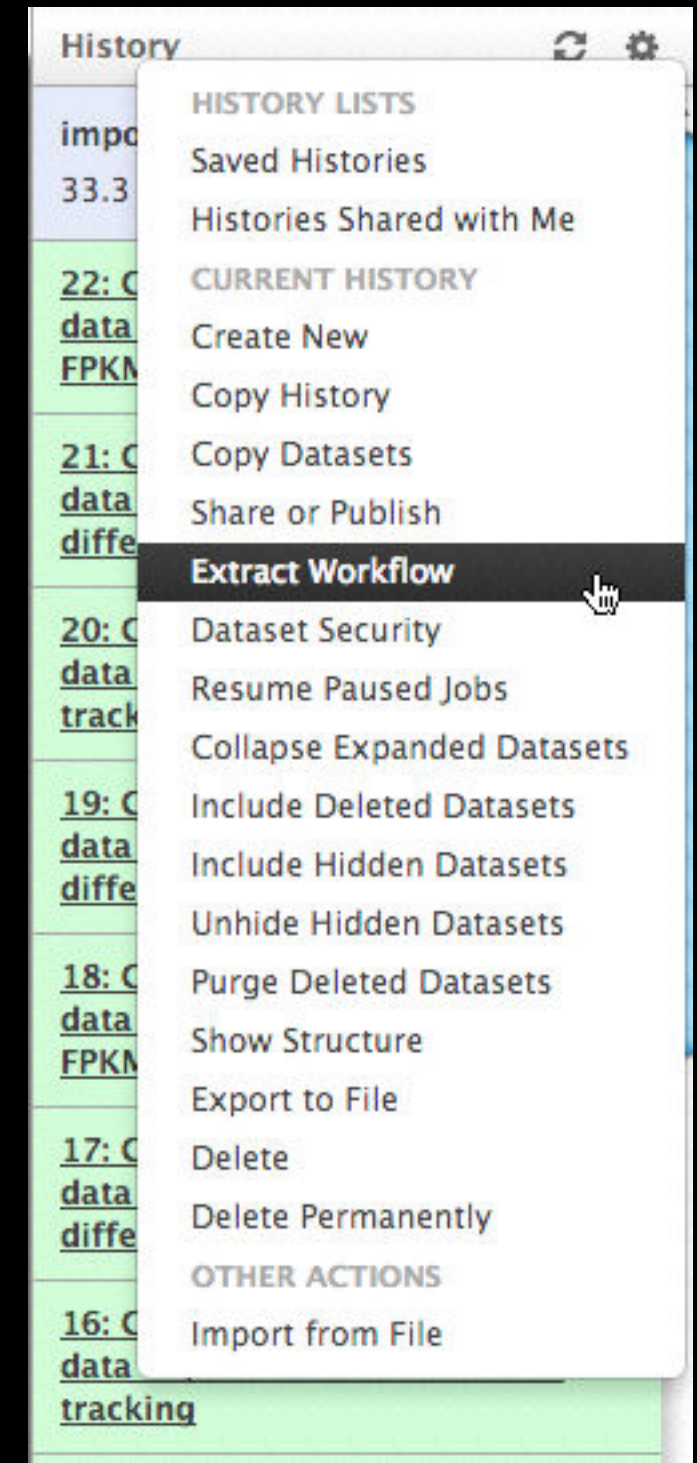
Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions




Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done

RNA-Seq Analysis: Get the Data

Create new history

 (cog) → Create New

Import:

Shared Data → Data Libraries

→ RNA-Seq UC Davis 2013 Example Data*

→ Unfiltered Reads

→ MeOH_REP1_R1.fastq and
MeOH_REP1_R2.fastq



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

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, **but 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]

S - Sanger	Phred+33,	93 values	(0, 93)	(0 to 60 expected in raw reads)
I - Illumina 1.3	Phred+64,	62 values	(0, 62)	(0 to 40 expected in raw reads)
X - Solexa	Solexa+64,	67 values	(-5, 62)	(-5 to 40 expected in raw reads)

http://en.wikipedia.org/wiki/FASTQ_format

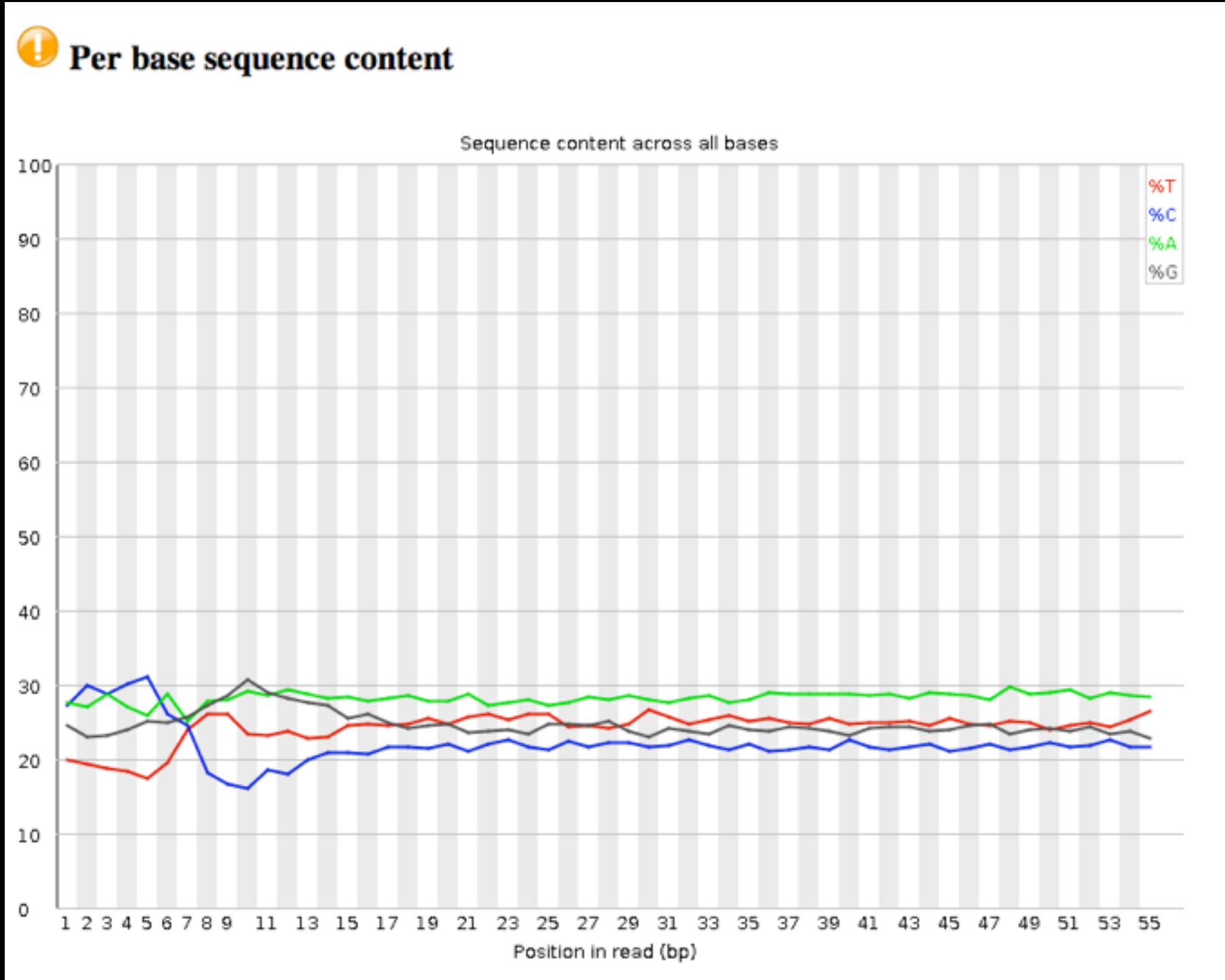
NGS Data Quality: Assessment tools

NGS QC and Manipulation → **FastQC**

Gives you 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?

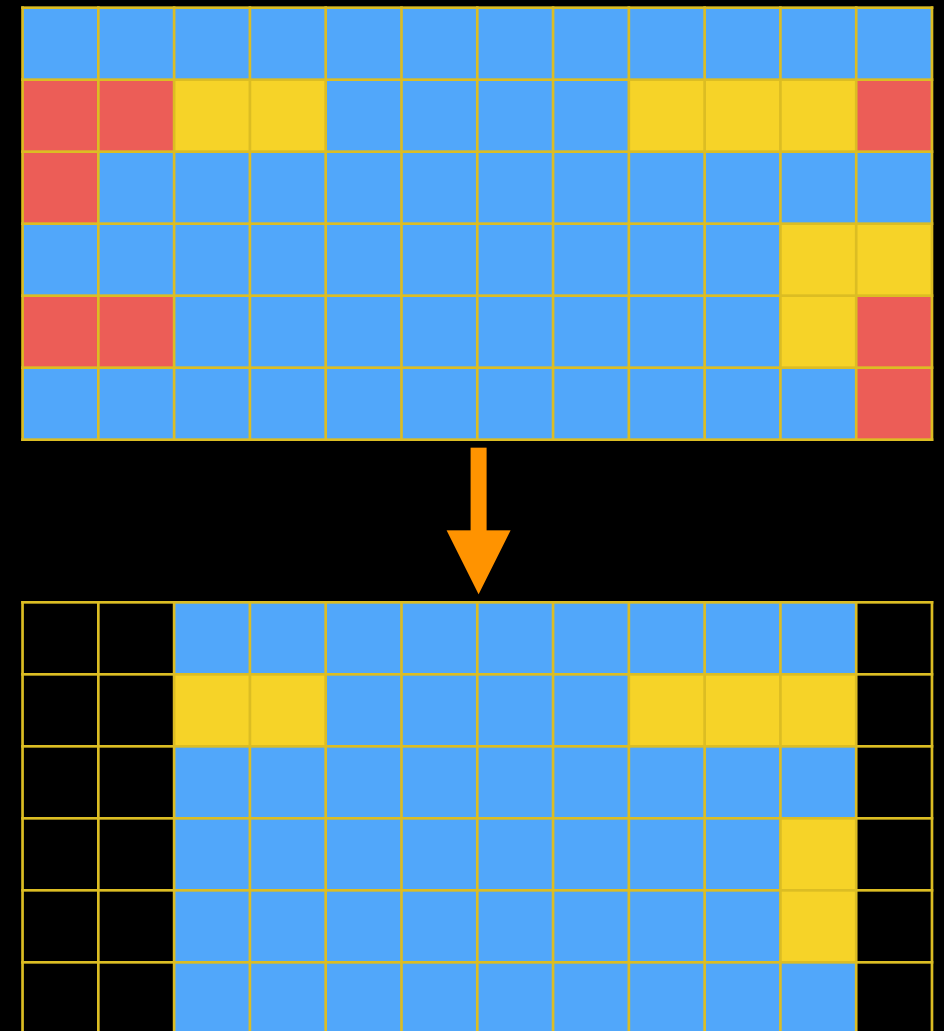


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

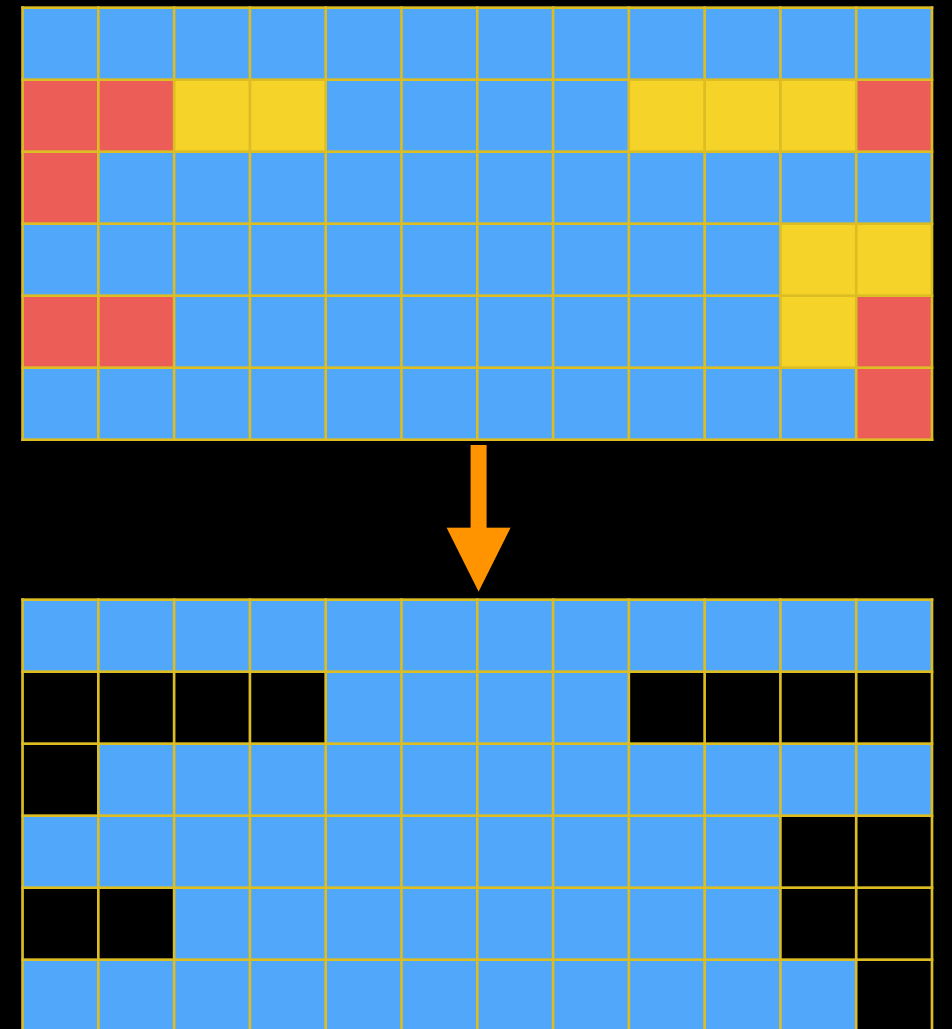
NGS Data Quality: Trim as we see fit

- 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

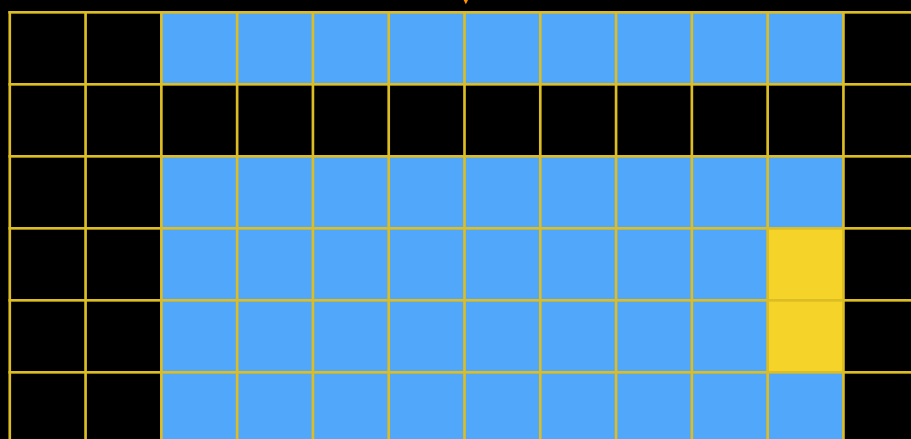
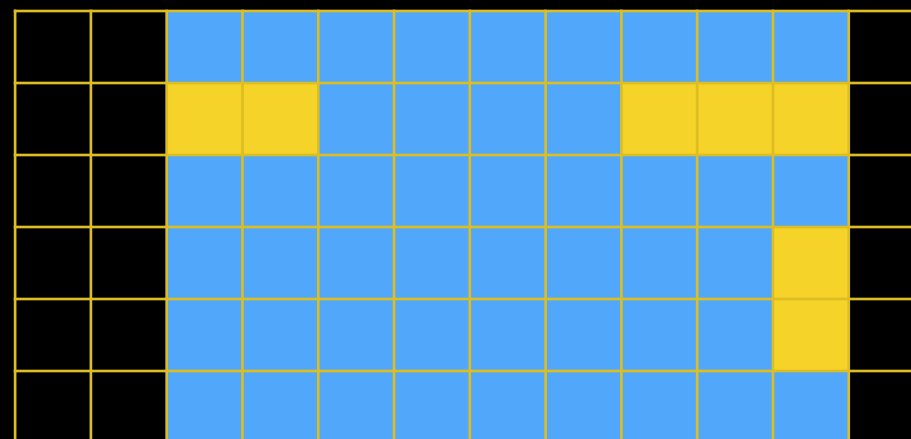
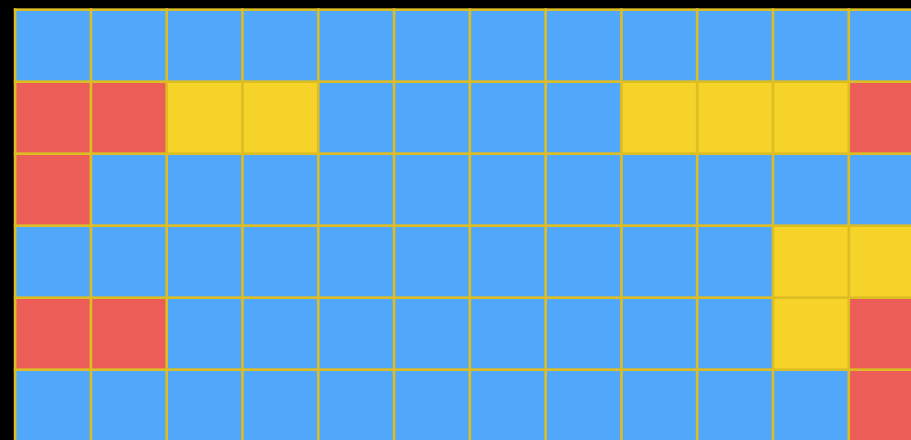


NGS Data Quality: Base Quality Trimming

- 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>



TopHat Overview

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie(2), and then analyzes the mapping results to identify splice junctions between exons. Please cite: Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36, 2013.

“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.”

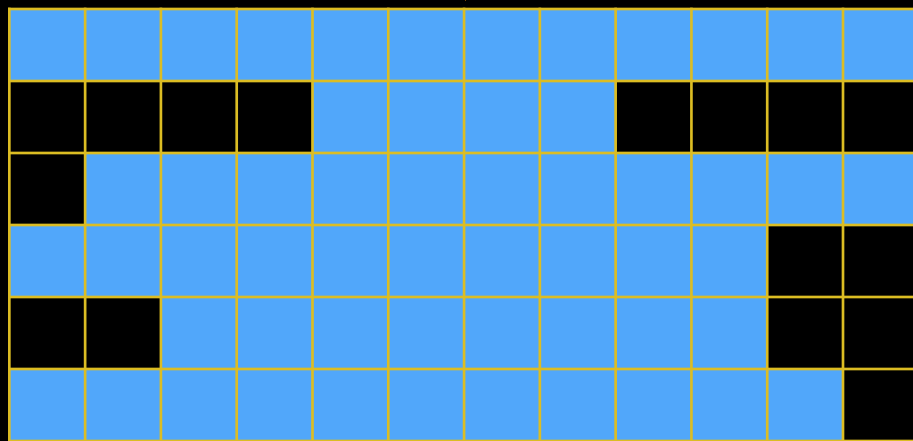
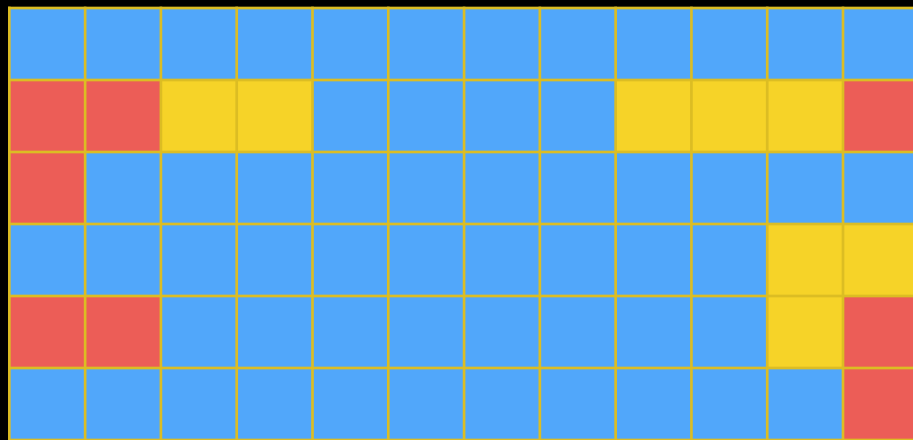
Most of us

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: Things to Try

- Don't bother.
- Run a workflow (try the "Re-Pair Paired ends after QC may have broken them" 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. (This does not work with Tophat.)

NGS Data Quality: Base Quality Trimming



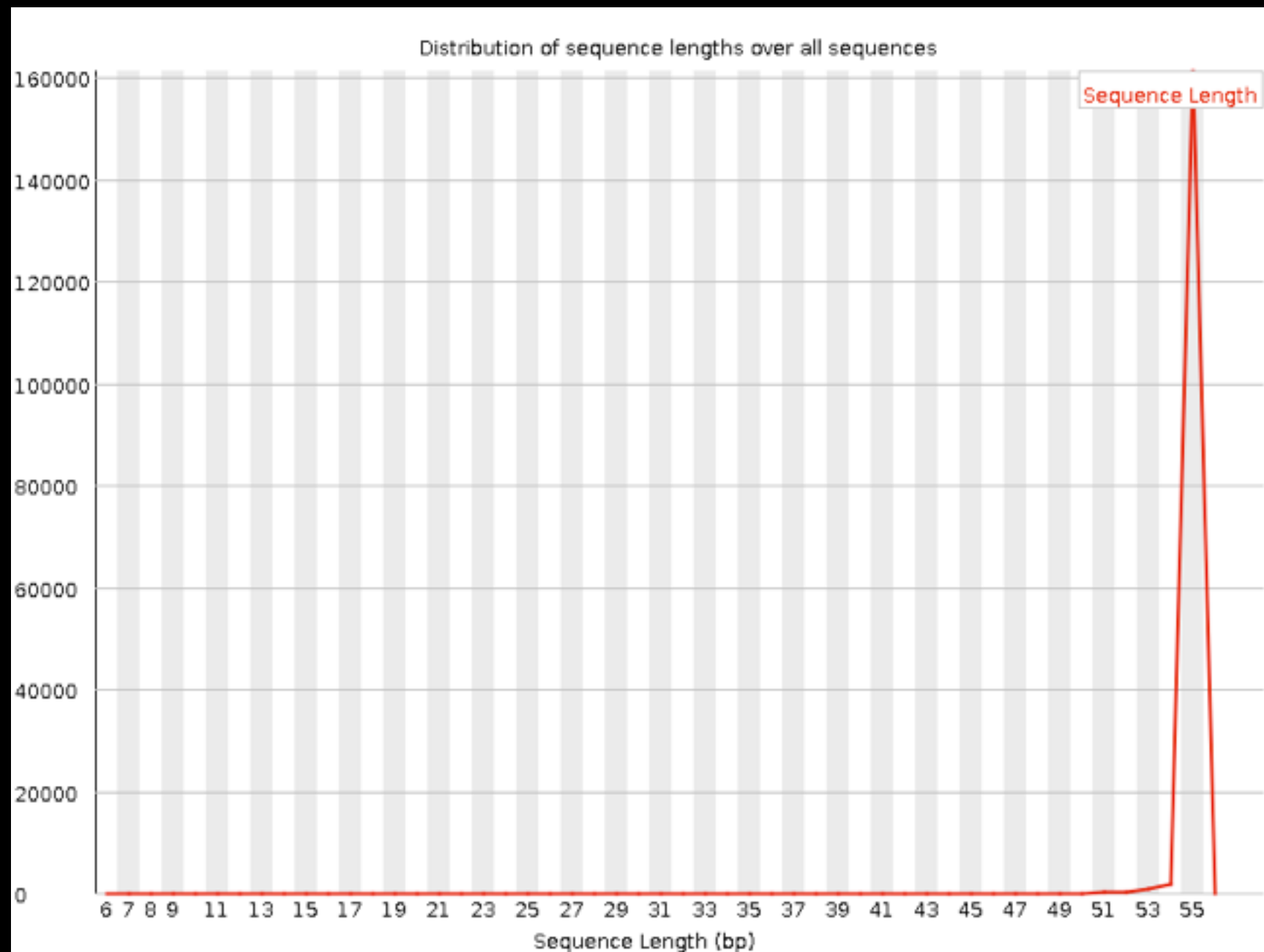
I'll use Option3, sliding windows, and run a workflow afterward to patch up pairings

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

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. Have potential to bog down mapping.

Option 2 can fix this, but breaks pairings (if you still have them).

Or, your **mapper may have an option to ignore shorter reads.**

RNA-Seq Analysis

I'll use option 2, since my pairings are already broken.

NGS QC and Manipulation


→ **Filter FASTQ reads by quality score and length**

Pick a minimum length. I used 32.

NGS Data Quality: Sequencing **Artifacts**

Repeat this process with MeOH Rep1 R2 (the reverse reads)

... and now we notice 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 (if we still have them).

Or, can rely on mapper to just not map them.

RNA-Seq Analysis: Restore Pairings

If your QC filters might have broken pairings, then you may want to restore them.

Shared Data → Published Workflows

- Re-Pair Paired ends after QC may have broken them
- Import

Workflows

- Re-Pair Paired ends after QC may have broken them
- Run

Re-Pair Paired ends after QC may have broken them

Workflow takes 4 inputs

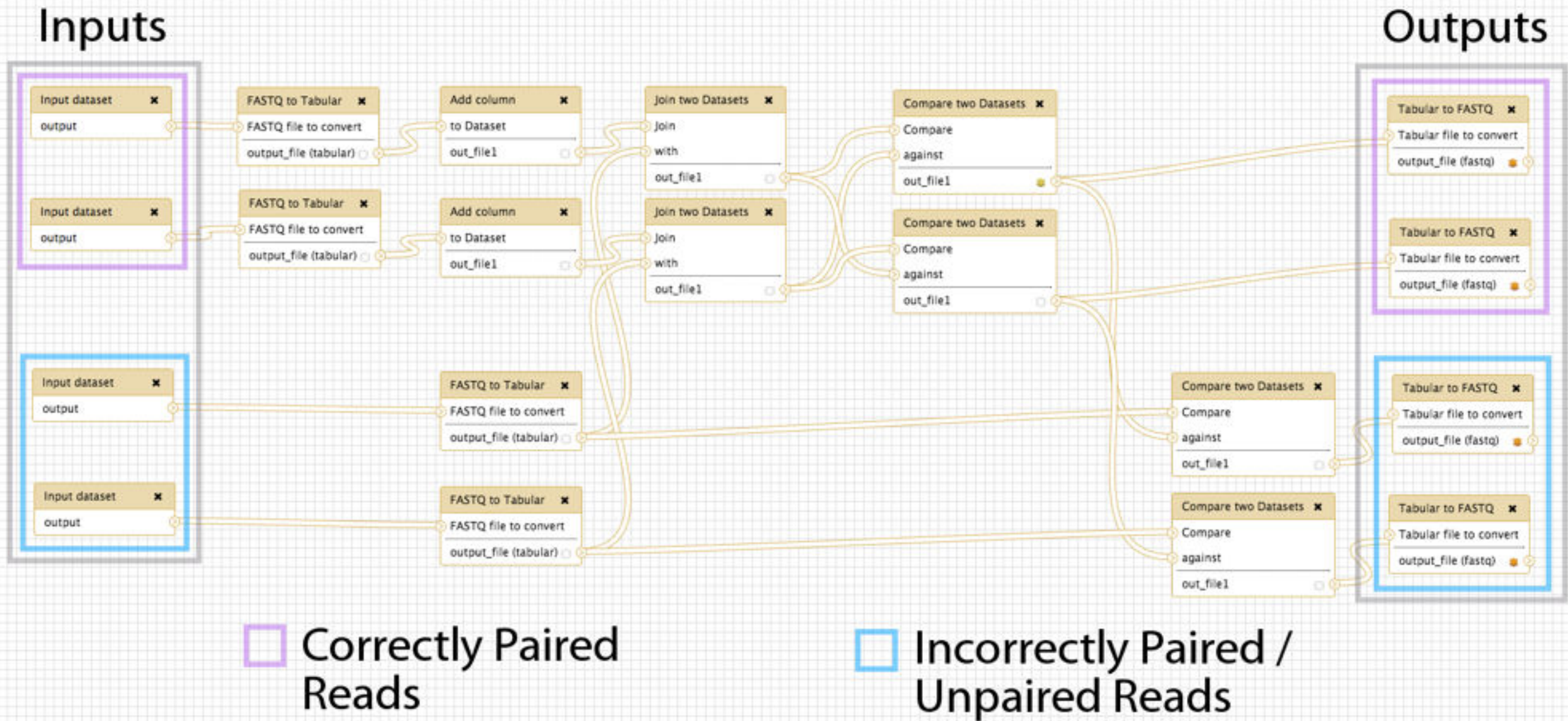
- Forward Reads, before QC
- Reverse Reads, before QC
- Forward Reads, after QC
- Reverse Reads, after QC

And produces 4 outputs

- Forward reads, re-paired
- Reverse reads, re-paired
- Forward reads, singletons
- Reverse reads, singletons

Workflow assumes pre-QC reads are correctly paired

Re-Pair Paired ends after QC may have broken them



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 trimmed+paired datasets from the

RNA-Seq UC Davis 2013 Example Data →

Reads, Post-QC, Re-Paired

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.*

Mapping with Tophat

RNA-Seq: Mapping with Tophat

Create new history

 (cog) → Create New

Get filtered reads

Shared Data → Data Libraries

→ RNA-Seq UC Davis 2013 Example Data*

→ Reads, Post-QC

→ Select MeOH_REP1_R1, MeOH_REP1_R2

Also select genes_chr12.gtf

And then Import to current history



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

RNA-seq Exercise: Mapping with Tophat

- Tophat looks for best place(s) to map reads, and best places to insert introns
- *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq mapping here.*

Mapping with Tophat: **mean inner distance**

Expected distance between paired 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 **breaks ties randomly**.

Tophat assigns equal fractional credit to all n mappings

Instructs TopHat to allow up to this many alignments to the reference for a given read, and choose the alignments based on their alignment scores if there are more than this number. The default is 20 for read mapping. Unless you use `--report-secondary-alignments`, TopHat will report the alignments with the best alignment score. **If there are more alignments with the same score than this number, TopHat will randomly report only this many alignments.** In case of using `--report-secondary-alignments`, TopHat will try to report alignments up to this option value, and TopHat may randomly output some of the alignments with the same score to meet this number.

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

Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done

Galaxy Community Resources

Galaxy Community Resources: Galaxy **Biostar**

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

Absolutely have to **encourage community support**.

Project traditionally used mailing list

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)

(3246 posts in 2014, 1000+ members)

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (47 posts in 2013, 3400+ members)

(34 posts in 2014, 4400+ members)

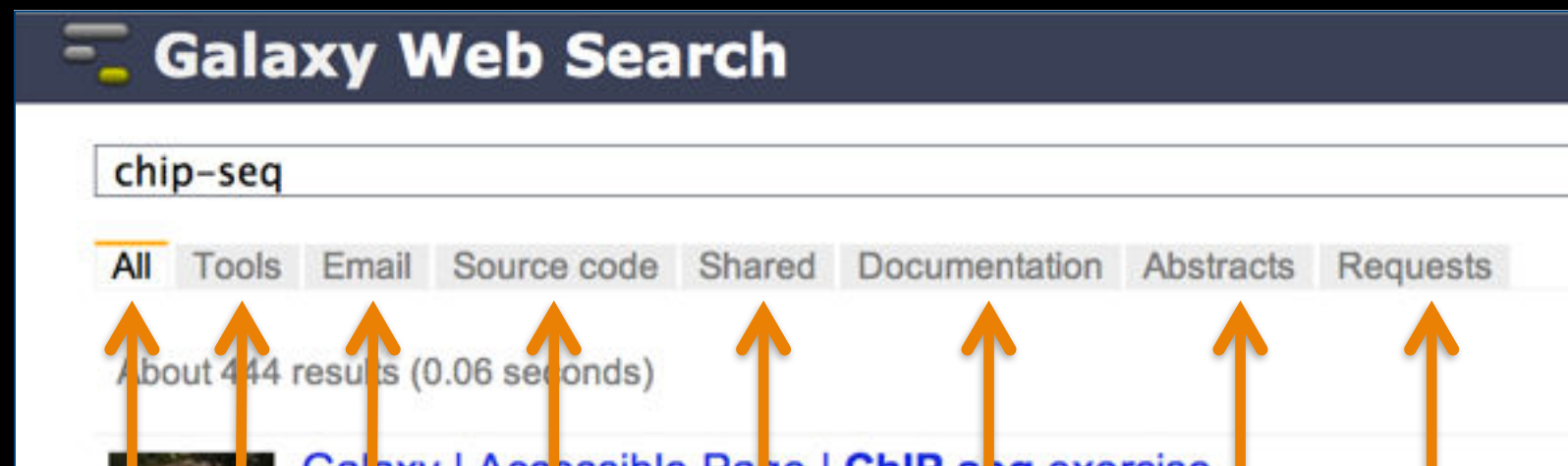
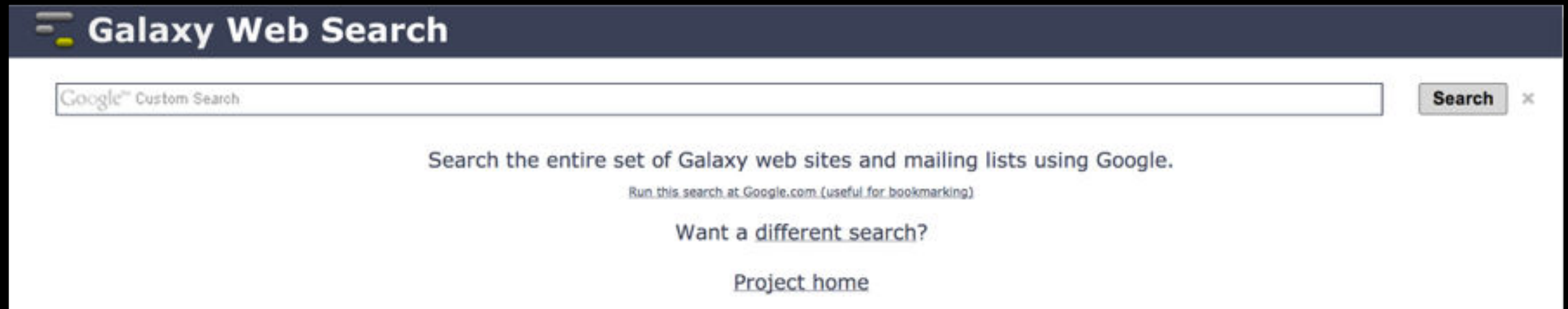
Galaxy-User (discontinued 2014/05)

Questions about using Galaxy and usegalaxy.org

High volume (1328 posts in 2013, 2600+ members)

(358 posts in 2014, 2600+ members)

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



Find

- Everything on ...
- Tools for ...
- Email about ...
- Source code for ...
- Published Histories, Pages, Workflows, about ...
- Documentation on ...
- Papers using Galaxy for ...
- Related feature requests



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 web server 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.



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](#)



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](#) • [Choices](#)
[Share](#) • [Search](#)

Communicate

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

Deploy Galaxy

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

Contribute

[Develop](#) • [Tools](#)
[Issues & Requests](#)
[Logs](#) • [Deployments](#)
[Teach](#)

Galaxy Project

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

Events

News

[DaveClements](#)
[Settings](#)
[Logout](#)
|
 Search:
[Titles](#)
[Text](#)

Events

Galaxy Event Horizon

Events with Galaxy-related content are listed here.

Also see the [Galaxy Events Google Calendar](#) for a listing of events and deadlines that are in 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, send it to outreach@galaxyproject.org.

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

Upcoming Events

Date	Topic/Event	Venue/Location
December 12	Introduction to Galaxy Workshop	Virginia State University, Petersburg, Virginia
December 16-19	RNA-Seq and ChIP-Seq Analysis with Galaxy	UC Davis, California, United States
2015		
January 10-14	Galaxy for SNP and Variant Data Analysis	Plant and Animal Genome XXIII (PAG2014), States
January 19-20	NGS pipelines with Galaxy	e-Infrastructures for Massively Parallel Sequencing, Sweden
February 9-13	Analyse bioinformatique de séquences sous Galaxy	Montpellier, France
February 16-18	Accessible and Reproducible Large-Scale Analysis with Galaxy	Genome and Transcriptome Analysis, Pacific Conference, San Francisco, California
	Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic	Genomics & Sequencing Data Integration, of Molecular Medicine Tri-Conference, San Francisco, California

Opening at McMaster University

The [McArthur Lab](#) in the [McMaster University Department of Biochemistry & Biomedical Sciences](#) is seeking a Systems Administrator / Information Technologist to help establish a new bioinformatics laboratory at McMaster, plus develop the next generation of the [Comprehensive Antibiotic Resistance Database \(CARD\)](#).

From the [job announcement on Evoldir](#):

The candidate will configure BLADE and other hardware for general bioinformatics analysis, development of a GIT version control system, **construction of an in house Galaxy server (usegalaxy.org)**, and development of a new interface, stand-alone tools, APIs, and algorithms for the CARD (based on [Chado](#)).

See the [full announcement](#) for details.

Posted to the [Galaxy News](#) on 2014-12-05

December 2014 Galaxy Newsletter

As always there's a lot going on in the Galaxy this month. "Like what?" you say. Well, read the dang [December Galaxy Newsletter](#) we say! Highlights include:

- [Galaxy Day! In Paris! This Wednesday!](#)
- Near Richmond, Virginia? There's a [Galaxy Workshop at Virginia State U on December 12](#).
- [GCC2015 needs sponsors!](#)
- Other [upcoming events](#) on two continents
- **96 new papers**, including 6 highlighted papers, referencing, using, extending, and implementing Galaxy.
- [Job openings at 7+ organizations](#)
- A new mailing list: [Galaxy-Training](#)
- [15 new ToolShed repositories](#) from 10 contributors
- And, 10 other juicy (well maybe not *juicy*, but certainly not *crunchy*) [bits of news](#)

Dave Clements and the *crisp* Galaxy Team

Posted to the [Galaxy News](#) on 2014-12-01

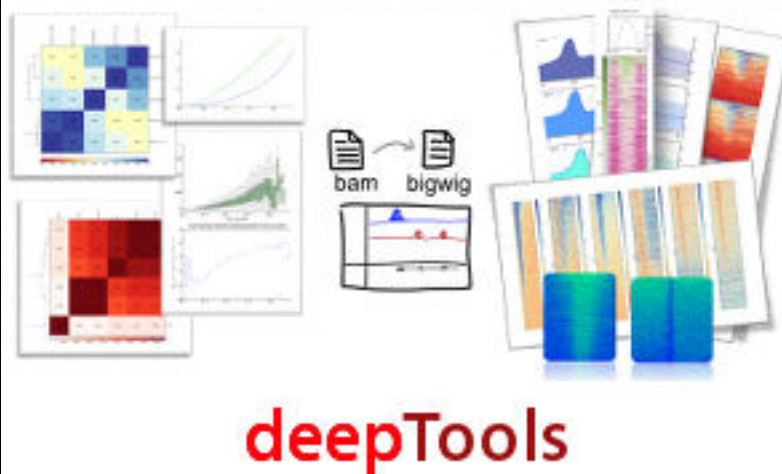
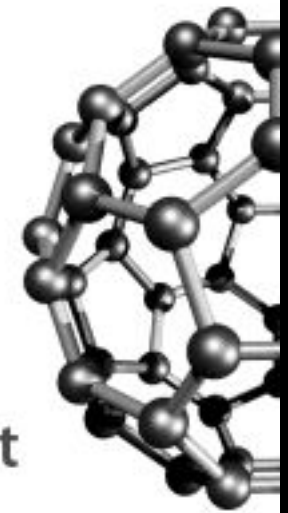
Bioinformaticians, Freiburg

[Max Planck Institute of Immunobiology and Epigenetics](#) in Freiburg, Germany has an opening for a Bioinformatician for an initial period of two years. The successful candidate will work at the interface between an in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the institute. Main responsibilities include

primary analysis of deep-sequencing data and quality control

[illegible]

Powered by the
**Biochemical
Algorithms
Library
Project**



CoSSci
Galaxy for Complex Social Sciences

galaxy.berkeleybop.org

QGIS (Quantum GIS) is a free tool for viewing, editing, and analyzing geographic information system (GIS) data. It is a cross-platform, open-source GIS application that runs on Windows, Mac OS X, and Linux. QGIS is a powerful tool for working with vector and raster data, and it can be used to create maps, perform spatial analysis, and manage GIS data. QGIS is a free and open-source GIS application that runs on Windows, Mac OS X, and Linux. It is a powerful tool for working with vector and raster data, and it can be used to create maps, perform spatial analysis, and manage GIS data. QGIS is a free and open-source GIS application that runs on Windows, Mac OS X, and Linux. It is a powerful tool for working with vector and raster data, and it can be used to create maps, perform spatial analysis, and manage GIS data.

The server is free for public use as a research tool for any institution, but not for commercial or non-commercial purposes (e.g., for sale) and can be compared to 12 million and a different record structure to the WHOIS database.

For your benefit, the server requires a separate set of the most important data points along with the name corrected. It is 7 million and a different record structure.

Food Item	Vegetarian	Vegan	Flexitarian	Omnivore
Coffee	~10%	~5%	~15%	~25%
Eggs	~5%	~0%	~10%	~20%
Lamb	~0%	~0%	~5%	~15%
Fish	~0%	~0%	~10%	~25%

Processing Pipeline



CloudBased Image Analysis & Processing Toolbox

More information can be found on the [NeCTAR website](#), and the project [blog](#).

This project is supported in part by NeCTAR, and CSIRO.



The Microbiome Analysis Center
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

We have developed the MIAZ Workplace Portal, a flexible and customizable web-based, with the aim of simplifying control, usage, access, and analysis of microclimate and air quality data. The Portal uses a relational database management system and distributed statistical measures and includes several tools such as: measure charts, maps,

bit.ly/gxyServers

Community can create, vote and comment on issues

The screenshot displays the Trello interface for the 'Galaxy: Development' board, which is set to 'Public'. The top navigation bar includes links for HOME, TOUR, GOLD, BUSINESS CLASS, and BLOG, along with the Trello logo and 'Sign Up' / 'Log In' buttons. A banner at the top encourages users to 'Sign up for free' to subscribe, vote, or comment on cards.

The board is organized into several columns, each representing a different category of issues:

- Inbox:** Contains cards such as 'To add cards, use http://galaxyproject.org/trello' (4 votes, 2 comments), 'To request reference genome, comment on this card.' (1 vote, 5 comments), 'Toolshed installation fails silently' (3 votes, 1 comment), 'Handle cluster job preemption' (2 votes, 1 comment), 'Return code 271 causes traceback for PBS torque' (1 vote, 2 comments), 'BUG: Tool shed repository export to capsule does not always capture all dependencies' (1 vote, 1 comment), and 'Remove manual_builds.txt from source control and replace with a .sample version' (1 vote, 1 comment).
- Tool Requests:** Includes cards like '595: Add SAMTools "Sort"' (4 votes, 13 comments), '601: SAM-to-BAM tool enhancements' (2 votes, 1 comment), 'Tools: Add tool to generate simulated reads to Main' (3 votes, 1 comment), 'default max insert size of Bowtie2 should be increased' (2 votes, 5 comments), '307: A tool to produce a set of random intervals.' (2 votes, 2 comments), 'Converter Tool: SAM to BAM enhancements' (2 votes), and 'New Tool: convert IUPAC chars to N' (1 vote, 7 comments, 1 attachment).
- Bug Reports:** Features cards such as 'Usability: expanding datasets near the bottom of panel' (marked as 'CE'), 'Bug: SICER on Main dependency issue' (2 votes, 20 comments, 3/5 progress), 'Profile Annotations bad values when "select all"' (1 vote, 5 comments), 'Filter pileup tool doesn't recognize pileup output data' (1 vote, 2 comments), 'Bug: Odd Fetch Taxonomy tool behavior' (1 vote, 1 comment), and 'Strip message after pause jobs resumed' (1 vote, 1 comment).
- Ideas:** Contains cards like '697: Workflow job control functions' (10 votes, 9 comments), 'User Metrics and Analytics' (3 votes, 3 comments, 1/2 progress, marked as 'CE'), 'Tuxedo RNA-seq tools: report command-line' (2 votes, 3 comments), 'Tools: Incorporate key Cuffdiff output files for Cummerbund' (2 votes, 1 comment, 0/3 progress), 'Moving objects between Galaxy instances, data federation, distributed storage, and data locality' (2 votes), and 'Workflow Editor: Provide explicit access to implicit datatype converter tools' (1 vote).

On the right side, a 'Menu' sidebar is visible, showing 'Members' (a grid of user avatars), 'Activity' (a list of recent actions, including 'Lance Parsons on Add or update wrappers for SamTools 1.0'), and a 'Pull Request' section.

<http://bit.ly/gxytrello>



GALAXY

COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Slides, posters & videos now online
<http://bit.ly/gcc2014>





GCC 2015

Galaxy Community Conference

6-8th July 2015

The Sainsbury Laboratory
Norwich, UK

gcc2015.tsl.ac.uk

Galaxy Australasia Workshop

2
0
1
4

We also support
community
organized efforts
and events.

swiss
german

galaxy
tour

Bern
30 Sep - 1 Oct

Freiburg
2 Oct

Galaxy Resources & Community: Videos

The screenshot shows the Vimeo channel for the Galaxy Project. The header includes the Vimeo logo and navigation links: Me, Videos, Create, Watch, Tools, Upload. A search bar is located on the right. The channel name "Galaxy Project" is displayed with a "PLUS" badge and a note "Joined 1 month ago". Below this, there are statistics: 54 Videos, 0 Likes, 0 Following, 1 Group, 6 Channels, and 0 Albums. A "Recently Uploaded" section features four video thumbnails. The first two are titled "Using Galaxy protocol 3" and "Using Galaxy protocol 2", both by "CPB Using Galaxy" and uploaded 5 days ago. The third is "Using Galaxy protocol 1" by "CPB Using Galaxy 1", also uploaded 5 days ago. The fourth is "FASTQ Prep Illumina" by "FASTQ Prep - Illumina", uploaded 1 week ago. A "Settings" button is visible on the left side of the channel page.

Galaxy Project PLUS
Joined 1 month ago

54 Videos 0 Likes 0 Following 1 Group 6 Channels 0 Albums

Recently Uploaded + See all 54 videos

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

FASTQ Prep Illumina
FASTQ Prep
Illumina
FASTQ Prep - Illumina
1 week ago

Settings

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>

Galaxy Resources & Community: CiteULike Group



[CiteULike](#) [Group: Galaxy](#) [Search](#) [Register](#) [Log in](#)

Group: Galaxy - library 2099 articles

[Search](#) [Copy](#) [Export](#) [Sort](#) [Hide Details](#)

✓ **Stress-induced endogenous siRNAs targeting regulatory intron sequence**
RNA, Vol. 21, No. 2. (01 February 2015), pp. 145-163, [doi:10.1261/rna.047662.114](#)
by [Hsiao-Lin V. Wang](#), [Brandon L. Dinwiddie](#), [Herman Lee](#), [Julia A. Chekanova](#)
posted to [methods](#) by [galaxyproject](#) to the group [Galaxy](#) on 2015-02-12 22:35:13 ★★
■ [Abstract](#)

✓ **Technical Perspectives on Knowledge Management in Bioinformatics W**
International Journal of Advanced Computer Science and Applications, Vol. 6, No. 1. (2015), pp. 1-6, [doi:10.14569/ijacsa.2015.060126](#)
by [Walaa](#), [Aksoy](#)
posted to [workbench](#) by [galaxyproject](#) to the group [Galaxy](#) on 2015-02-12 22:33:37 ★
■ [Abstract](#)

✓ **Analysis of the complete chloroplast genome of *Castanea pumila* var. *pumila***
In Tree Genetics & Genomes, Vol. 11, No. 1. (2015), pp. 1-6, [doi:10.1007/s11295-015-0151-1](#)
by [Fenny Dane](#), [Zhuoyu Wang](#), [Leslie Goertzen](#)
posted to [methods](#) [usemain](#) by [galaxyproject](#) to the group [Galaxy](#) on 2015-02-12 22:33:37 ★
■ [Abstract](#)

✓ **Multi-omic data analysis using Galaxy**
Nat Biotech, Vol. 33, No. 2. (6 February 2015), pp. 137-139, [doi:10.1038/nbt.3134](#)
by [Jorrit Boekel](#), [John M. Chilton](#), [Ira R. Cooke](#), et al.
posted to [other](#) [project](#) [refpublic](#) [tools](#) by [galaxyproject](#) to the group [Galaxy](#) on 2015-02-12 22:29:35 ★★★★★ [along](#)
with 8 people

Group Tags

All tags in the group Galaxy

Filter:

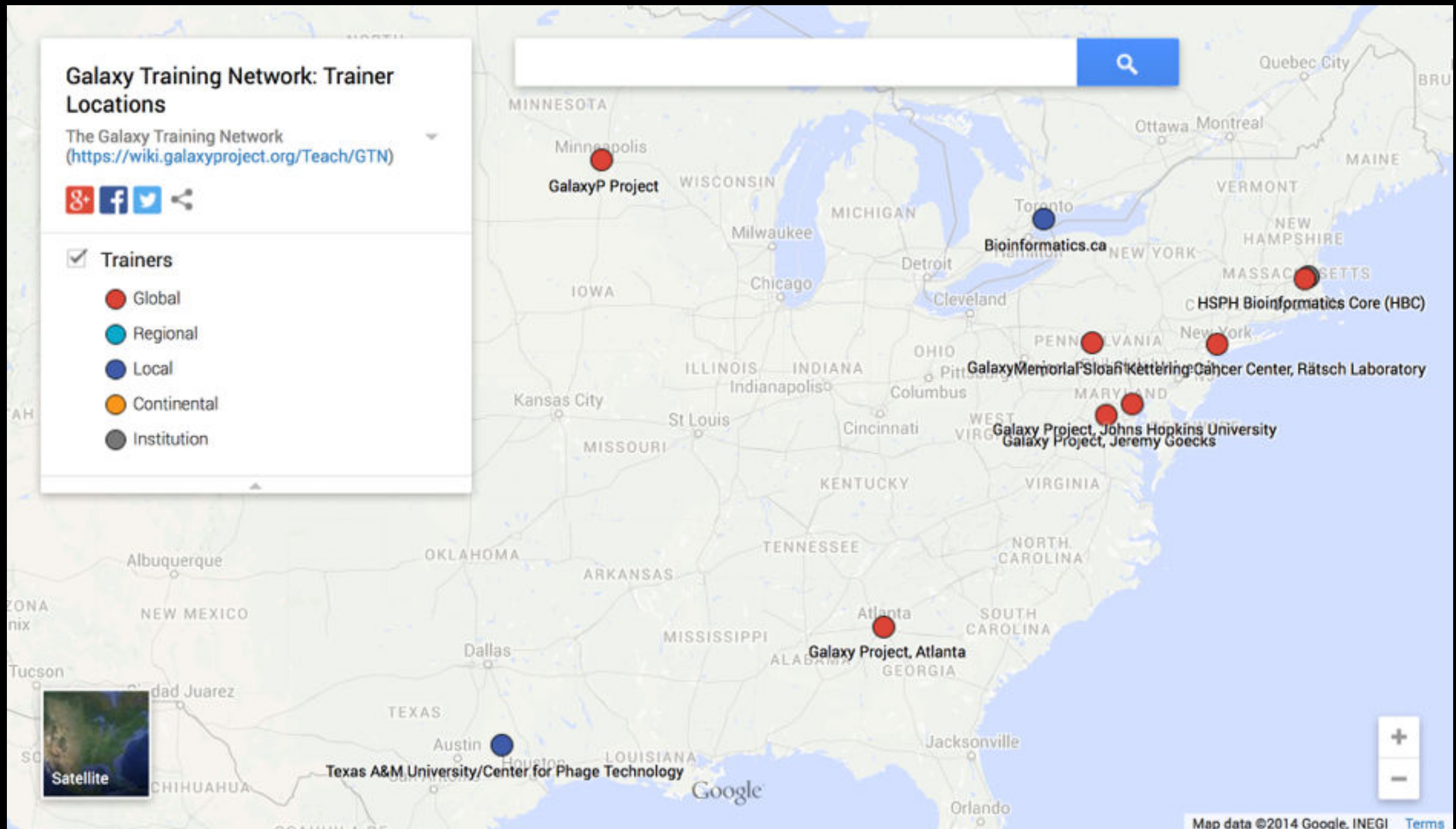
[\[Display as Cloud\]](#)

methods	1023
workbench	649
usemain	200
tools	152
isgalaxy	116
usepublic	93
cloud	78
uselocal	73
shared	73
other	57
unknown	50
reproducibility	44
howto	43
project	41
refpublic	41
visualization	12
usecloud	4

Over
2100
papers

<http://bit.ly/gxycul>

Scaling Training



Galaxy Training Network launched In October.
bit.ly/gxygtn

Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done

Tuxedo Suite: Some parts of the ensemble


Bowtie	Short read mapper. Bowtie2 can do gapped alignments and emphasizes reads > 50 bases
Tophat	Intron-aware mapper for RNA-Seq data. Works with Bowtie to find best mapping locations
Cufflinks	Construct transcript predictions from mapped reads (from Tophat output)
Cuffmerge	Merges multiple sets of transcript predictions into a unified set with one coherent set of IDs.
Cuffdiff	Differential expression analysis; Can work with Tophat output directly or Cufflinks/merge, if looking for novel genes/transcripts

Used already	Will not use today	Will use next
--------------	--------------------	---------------

RNA-Seq: Differential Expression with Cuffdiff

RNA-Seq Differential Expression: Get the Data

Create new history

 (cog) → Create New

Import:

Shared Data → Data Libraries

→ RNA-Seq UC Davis 2013 Example Data*

→ Tophat Outputs

→ Select all **accepted_hits** datasets

Also select **genes_chr12.gtf**

And then **Import to current history**



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

Cuffdiff

- Part of the Tuxedo RNA-Seq Suite (as are Tophat and Bowtie)
- Identifies differential expression between multiple datasets
- Widely used and widely installed on Galaxy instances

NGS: RNA Analysis → Cuffdiff

Cuffdiff

Cuffdiff previously used FPKM/RPKM as central statistic.

Total # mapped reads heavily influences FPKM/RPKM.

Can lead to challenges when you have very highly expressed genes in the mix.

Cuffdiff

- Running with 2 Groups: MeOH and R3G
- Each group has 3 replicates each

Cuffdiff

- Which Transcript definitions to use?
 - Official (**genes_chr12.gtf** in our case)
 - MeOH or R3G **Cufflinks** transcripts
 - Results of **Cuffmerge** on MeOH & R3G Cufflinks transcripts
- Depends on what you care about

NGS: RNA Analysis → Cuffdiff

Cuffdiff

Produces many output files, all explained in doc

We'll focus on gene differential expression testing

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
A2M	A2M	A2M	chr12:9217772-9268558	MeOH	R3G	NOTEST	3.32147	3.13694	-0.0824644	0	1	1	no
A2M-AS1	A2M-AS1	A2M-AS1	chr12:9217772-9268558	MeOH	R3G	NOTEST	7.45797	13.9413	0.902515	0	1	1	no
A2ML1	A2ML1	A2ML1	chr12:8975149-9029381	MeOH	R3G	NOTEST	4.83055	7.79884	0.691072	0	1	1	no
A2MP1	A2MP1	A2MP1	chr12:9381128-9386803	MeOH	R3G	NOTEST	2.49656	0	-inf	0	1	1	no
AAAS	AAAS	AAAS	chr12:53701239-53715412	MeOH	R3G	OK	269.035	159.23	-0.756683	-2.22857	0.0005	0.00194017	yes
AACS	AACS	AACS	chr12:125549924-125627871	MeOH	R3G	NOTEST	29.2933	35.0339	0.258178	0	1	1	no
ABCB9	ABCB9	ABCB9	chr12:123405497-123451056	MeOH	R3G	NOTEST	4.68869	1.7732	-1.40283	0	1	1	no
ABCC9	ABCC9	ABCC9	chr12:21950323-22089628	MeOH	R3G	OK	553.247	487.261	-0.18323	-2.02806	0.0004	0.00162143	yes
ABCD2	ABCD2	ABCD2	chr12:39945021-40013843	MeOH	R3G	OK	86.1377	172.795	1.00435	4.3436	5e-05	0.000246739	yes
ACACB	ACACB	ACACB	chr12:109577201-109706030	MeOH	R3G	NOTEST	8.45306	15.5772	0.881885	0	1	1	no
ACAD10	ACAD10	ACAD10	chr12:112123856-112194911	MeOH	R3G	NOTEST	21.8237	27.8326	0.350882	0	1	1	no
ACADS	ACADS	ACADS	chr12:121163570-121177811	MeOH	R3G	NOTEST	38.644	16.1739	-1.25658	0	1	1	no
ACRBP	ACRBP	ACRBP	chr12:6747241-6756580	MeOH	R3G	NOTEST	2.96987	3.26939	0.138621	0	1	1	no
ACSM4	ACSM4	ACSM4	chr12:7456927-7480969	MeOH	R3G	NOTEST	0	0	0	0	1	1	no
ACSS3	ACSS3	ACSS3	chr12:81471808-81649582	MeOH	R3G	NOTEST	0	0	0	0	1	1	no
ACTR6	ACTR6	ACTR6	chr12:100593864-100618202	MeOH	R3G	OK	475.594	421.324	-0.174799	-0.797581	0.1588	0.258406	no
ACVR1B	ACVR1B	ACVR1B	chr12:52345450-52390863	MeOH	R3G	NOTEST	32.5737	38.3075	0.233922	0	1	1	no
ACVRL1	ACVRL1	ACVRL1	chr12:52301201-52317145	MeOH	R3G	NOTEST	1.27713	2.16161	0.759201	0	1	1	no
ADAM1A	ADAM1A	ADAM1A	chr12:112336866-112339706	MeOH	R3G	NOTEST	30.0162	55.2154	0.879331	0	1	1	no
ADAMTS20	ADAMTS20	ADAMTS20	chr12:43748011-43945724	MeOH	R3G	NOTEST	0.453322	0.502067	0.147346	0	1	1	no
ADCY6	ADCY6	ADCY6	chr12:49159974-49182820	MeOH	R3G	NOTEST	9.32722	17.6743	0.922135	0	1	1	no
ADIPOR2	ADIPOR2	ADIPOR2	chr12:1800246-1897845	MeOH	R3G	OK	207.468	179.333	-0.210248	-1.02392	0.09	0.158988	no
AEBP2	AEBP2	AEBP2	chr12:19592607-19675173	MeOH	R3G	OK	143.039	128.293	-0.156957	-0.688267	0.2254	0.344537	no
AGAP2	AGAP2	AGAP2	chr12:58118075-58135944	MeOH	R3G	OK	98.2385	116.302	0.243511	0.935119	0.11475	0.198086	no
AICDA	AICDA	AICDA	chr12:8754761-8765442	MeOH	R3G	NOTEST	78.1514	63.4313	-0.301077	0	1	1	no
AKAP3	AKAP3	AKAP3	chr12:4724675-4754343	MeOH	R3G	NOTEST	6.12385	7.89626	0.366731	0	1	1	no
ALDH1L2	ALDH1L2	ALDH1L2	chr12:105413561-105478341	MeOH	R3G	NOTEST	7.11374	8.11722	0.190377	0	1	1	no
ALDH2	ALDH2	ALDH2	chr12:112204690-112247789	MeOH	R3G	NOTEST	12.8033	8.05635	-0.668321	0	1	1	no
ALG10	ALG10	ALG10	chr12:34175215-34181236	MeOH	R3G	NOTEST	54.8575	59.3459	0.11346	0	1	1	no
ALG10B	ALG10B	ALG10B	chr12:38710556-38723528	MeOH	R3G	NOTEST	43.8157	63.0457	0.524952	0	1	1	no
ALKBH2	ALKBH2	ALKBH2	chr12:109525992-109531293	MeOH	R3G	OK	679.517	297.183	-1.19316	-3.34255	5e-05	0.000246739	yes
ALX1	ALX1	ALX1	chr12:85674035-85695561	MeOH	R3G	NOTEST	0	0	0	0	1	1	no

Cuffdiff: differentially expressed genes

Column	Contents
test_stat	value of the test statistic used to compute significance of the observed change in FPKM
p_value	Uncorrected P value for test statistic
q_value	FDR-adjusted p-value for the test statistic
status	Was there enough data to run the test?
significant	and, was the gene differentially expressed?

Cuffdiff

- Column 7 (“status”) can be FAIL, NOTEST, LOWDATA or OK
 - Filter and Sort → Filter
 - `c7 == 'OK'`
- Column 14 (“significant”) can be yes or no
 - Filter and Sort → Filter
 - `c14 == 'yes'`

Returns the list of genes with

- 1) enough data to make a call, and
- 2) that are called as differentially expressed.

Cuffdiff: Next Steps

Try running Cuffdiff with different **normalization** and **dispersion estimation** methods.

Compare the differentially expressed gene lists.
Which settings have what type of impacts on the results?

RNA-Seq Differential Expression with Cuffdiff: Resources

RNA-Seq Concepts, Terminology, and Work Flows

by Monica Britton

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

Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break**
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done

Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 **Sharing, Publishing, and Reproducibility**
- 3:25 **Setting up your own Galaxy Cluster on Amazon**
- 4:00 Done

More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Sharing & Publishing enables **Reproducibility**

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**):

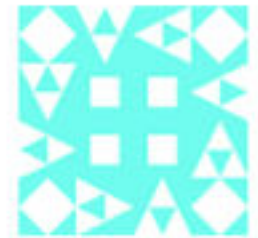
 **Galaxy History | metagenomic analysis**  

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 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

megam

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

Basic Analysis: Further reading & Resources

<http://usegalaxy.org/galaxy101>

<https://vimeo.com/76343659>

Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done

Galaxy is available ...



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

<http://globus.org/>

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

We are using the cloud today.

AWS in Education Grants Program



[**http://aws.amazon.com/education**](http://aws.amazon.com/education)

What is our path?

Today we will:

- Launch our own Galaxy server on AWS
- Make the server dynamically scalable in response to demand.
- Run some basic analysis on it.
- Make it go away.

Full Disclosure

To use AWS you must create an AWS account with a credit card associated with it.

You must also have created a key pair.

We will use the IAM account for this workshop.

IAM Accounts

Imagine, a link to a list of accounts, and
credentials, here.

CloudLaunch: From UseGalaxy.org

[Analyze Data](#) [Workflow](#) [Shared Data](#) [Visualization](#) [Cloud](#) [Help](#) [User](#)

[New Cloud Cluster](#)

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).

Identifying
sequence

Galaxy 1


Start sma

The very first tutorial

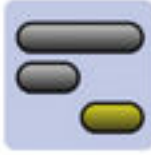
eer Lab

Tweets


[Follow](#)



Galaxy Project @galaxyproject 1h
Galaxy "evolves thanks to a collaborative community of users & developers as a shared effort"
biomedcentral.com/1471-2229/15/48
Amen! #usegalaxy
[Expand](#)



Galaxy Project @galaxyproject 2h
Selected Approaches & Frameworks to Carry out Genomic Data Analysis on the Cloud, by Church & Goscinski
bit.ly/1wngLD9 #usegalaxy



Sebastian Schönherr @seppinho 11h
great talk by @EnisAfgan about Galaxy CloudMan & the Genomics

CloudLaunch

[Analyze Data](#)[Workflow](#)[Shared Data ▾](#)[Visualization](#)[Cloud ▾](#)[Help ▾](#)[User ▾](#)

Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See [Amazon's pricing](#) for more information.

Key ID

This is the text string that uniquely identifies your account, found in the [Security Credentials section of the AWS Console](#).

Secret Key

This is your AWS Secret Key, also found in the [Security Credentials section of the AWS Console](#).

CloudLaunch

Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See [Amazon's pricing](#) for more information.

Key ID

This is the text string that uniquely identifies your account, found in the [Security Credentials](#) section of the [AWS Console](#).

Secret Key

This is your AWS Secret Key, also found in the [Security Credentials](#) section of the [AWS Console](#).

Instances in your account

Cluster Name

This is the name for your cluster. You'll use this when you want to restart.

Cluster Password

Cluster Password - Confirmation

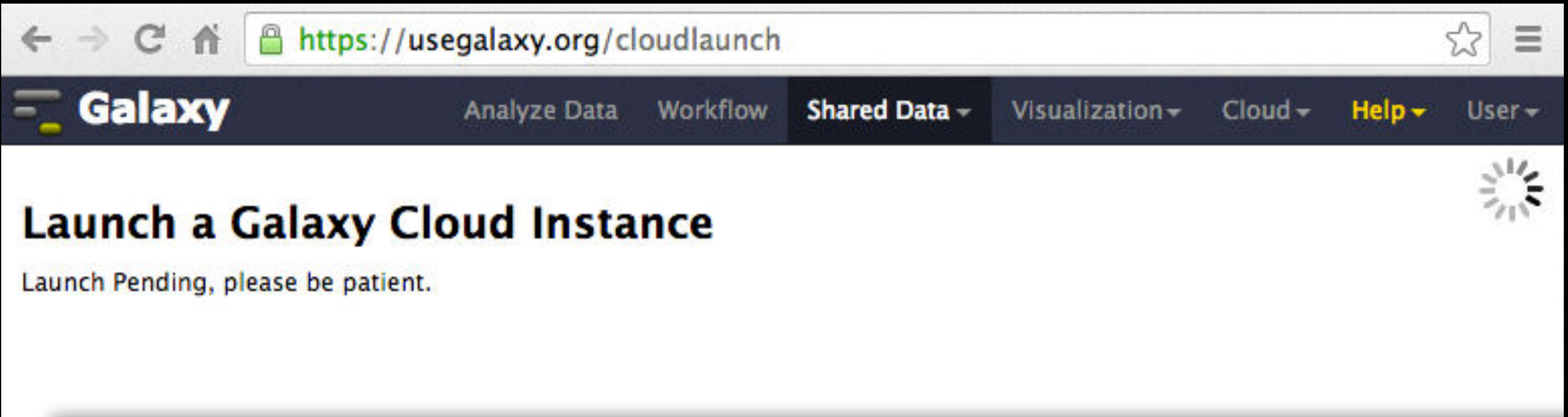
Key Pair


Instance Type

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

Submit

CloudLaunch



 **Galaxy**

Analyze Data Workflow **Shared Data** Visualization Cloud Help User

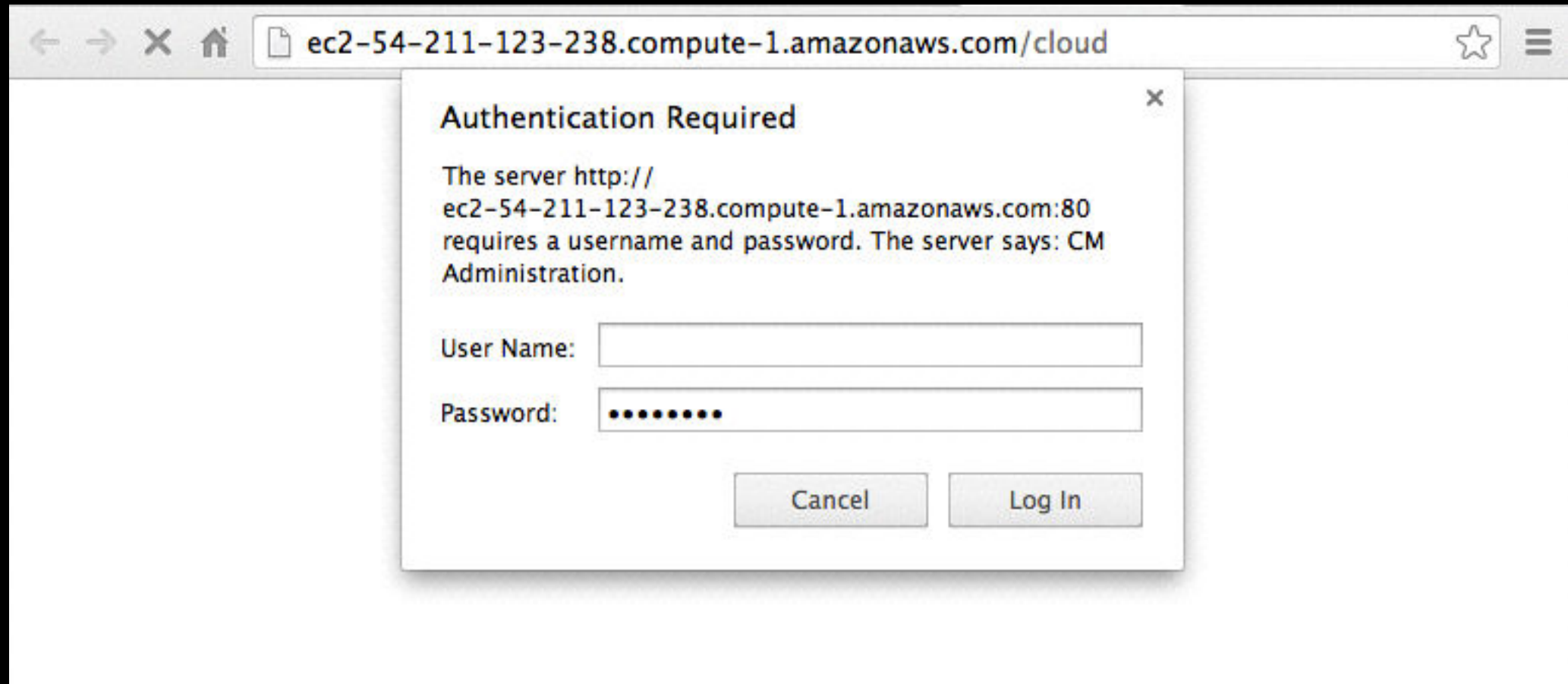
Launch a Galaxy Cloud Instance

Access Information

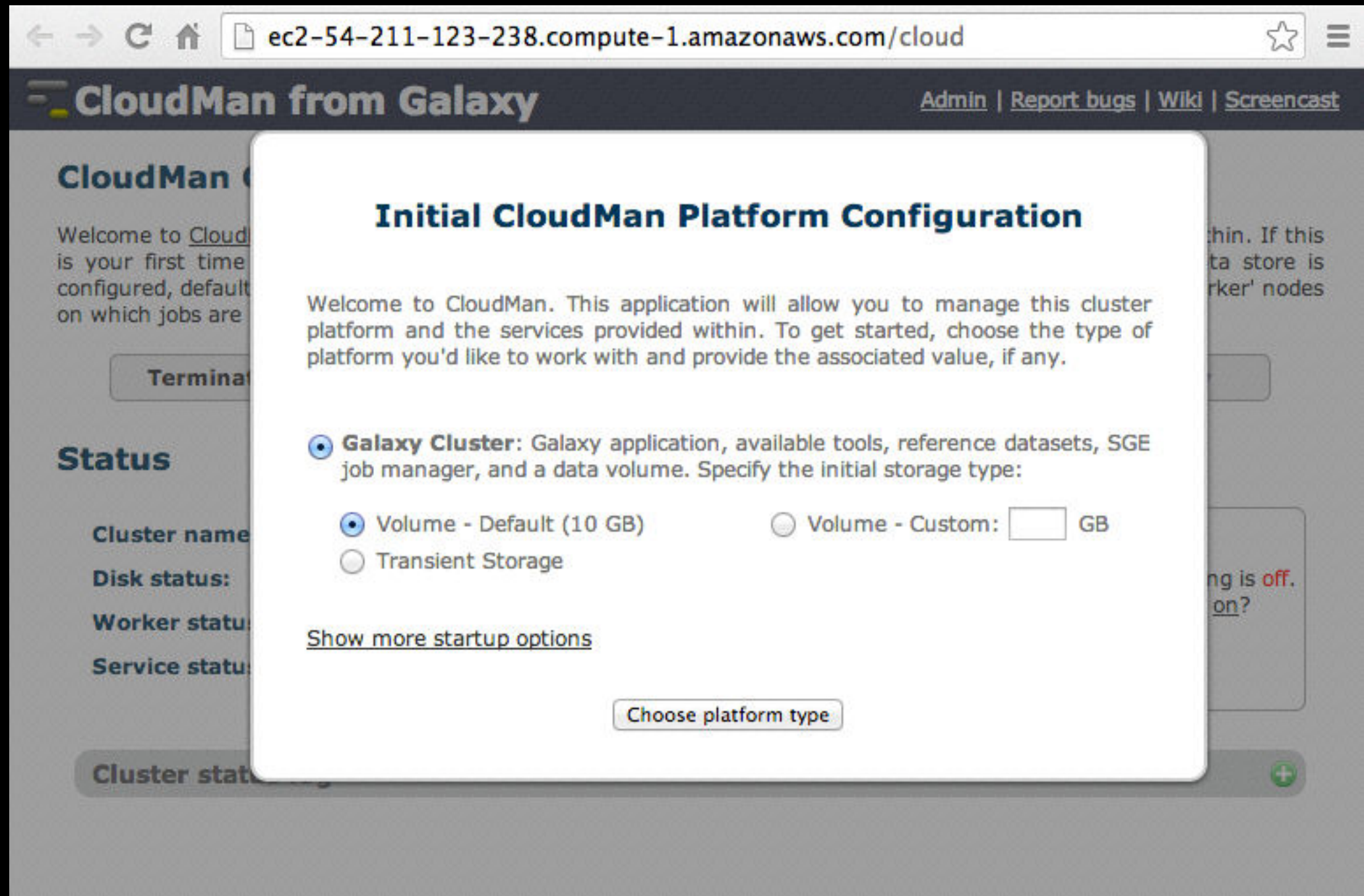
Your instance 'i-61503e9b' has been successfully launched using the 'ami-a7dbf6ce' AMI. While it may take a few moments to boot, you will be able to access the cloud control panel at ec2-54-196-164-110.compute-1.amazonaws.com/cloud. SSH access is also available using your private key. From the terminal, you would execute something like:

```
`ssh -i cloudman_key_pair.pem ubuntu@ec2-54-196-164-110.compute-1.amazonaws.com`
```

CloudLaunch



CloudLaunch



The screenshot shows a web browser window with the address bar displaying `ec2-54-211-123-238.compute-1.amazonaws.com/cloud`. The page title is "CloudMan from Galaxy". The main content area is a modal dialog titled "Initial CloudMan Platform Configuration". The dialog contains a welcome message, a list of radio buttons for selecting the platform type, and a "Choose platform type" button. The background of the page is dimmed, showing a sidebar with "CloudMan" and "Status" sections.

← → ↻ 🏠 `ec2-54-211-123-238.compute-1.amazonaws.com/cloud` ☆ ☰

CloudMan from Galaxy [Admin](#) | [Report bugs](#) | [Wiki](#) | [Screencast](#)

CloudMan

Welcome to CloudMan. This application will allow you to manage this cluster platform and the services provided within. To get started, choose the type of platform you'd like to work with and provide the associated value, if any.

☒ **Galaxy Cluster:** Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage type:

☒ Volume - Default (10 GB) ☐ Volume - Custom: GB

☐ Transient Storage

[Show more startup options](#)

Status

Cluster name:

Disk status:

Worker status:

Service status:

Cluster status:

← → ↺ 🏠 ec2-54-211-123-238.compute-1.amazonaws.com/cloud ⭐ ☰

✕

Messages

Initializing 'Galaxy' cluster type. Please wait... (2014-01-15 06:48:34)

Welcome to [CloudMan](#). This application allows you to manage this cloud cluster and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Terminate cluster

Add nodes ▼

Remove nodes

Access Galaxy

Status

Cluster name: PAG_CLOUD_2

Disk status: 0 / 0 (0%)

Worker status: Idle: 0 Available: 0 Requested: 0

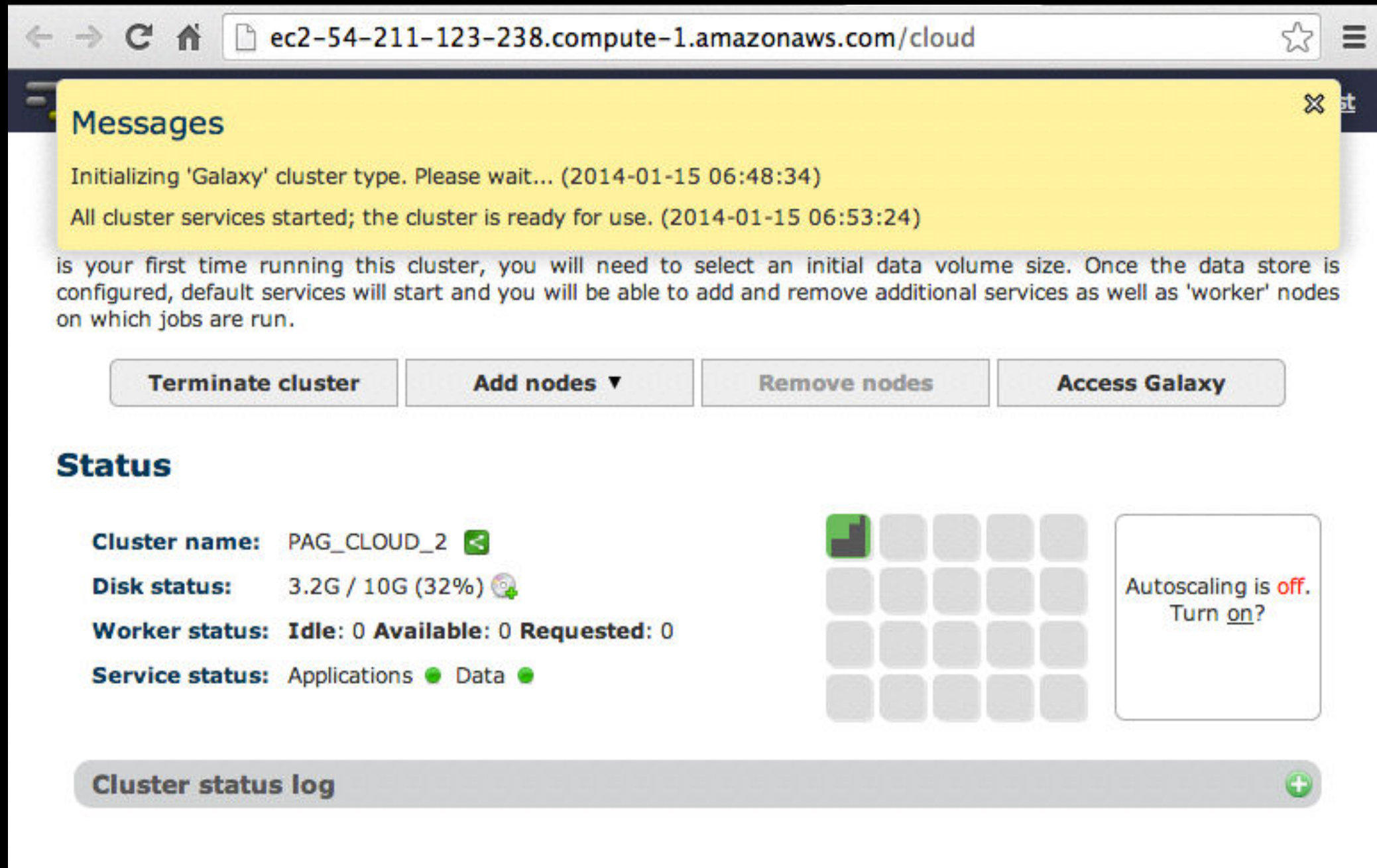
Service status: Applications ● Data ●

Autoscaling is off.
Turn on?

Cluster status log

+

Cloud Launched



Cool things to do

- Create a **login**
- Become an **admin**
- Set up **autoscaling**
- Run ~ **Galaxy 101**
 - <http://usegalaxy.org/galaxy101>
- **Shut it down**

Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done (almost)

We Need Your Feedback

bit.ly/NUgxy201502

The Galaxy Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Marten Cech



John Chilton



Dave Clements



Nate Coraor



Carl Eberhard



Jeremy Goecks



Sam Guerler



Jen Jackson



Ross Lazarus



Anton Nekrutenko



Nick Stoler



James Taylor

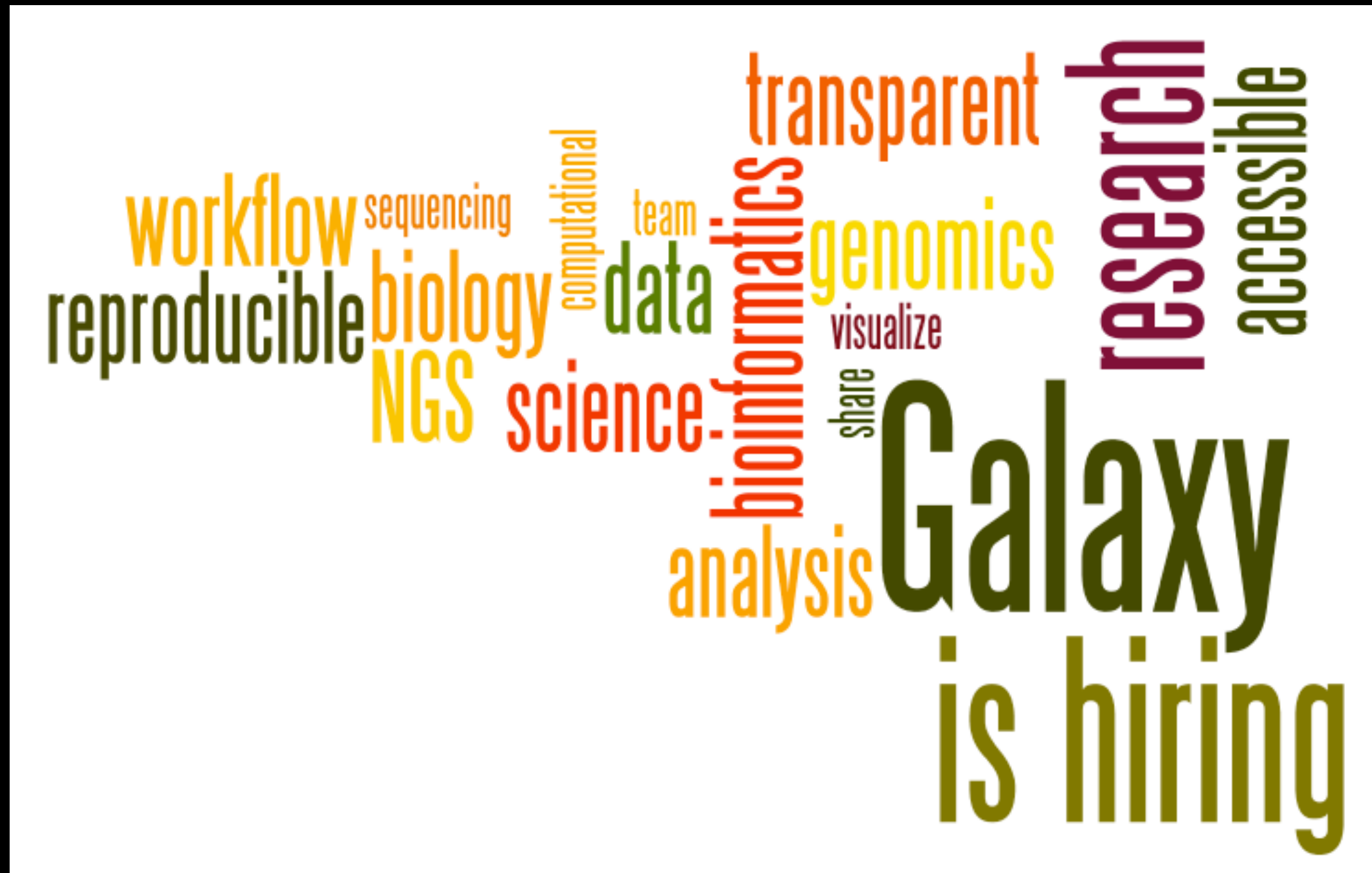


Nitesh Turaga

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

bit.ly/NUgxy201502

Galaxy is hiring post-docs and software engineers



Please help.

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

Also Thanks To



Pamela Shaw
Kristi Holmes

National Institutes of Health

bit.ly/NUgxy201502

Agenda

- 9:00 Welcome
- 9:30 Basic Analysis with Galaxy
- 10:45 Break
- 11:00 Basic Analysis into Reusable Workflows
- 11:30 RNA-Seq Example Part I
- 12:20 Lunch (on your own)
- 1:35 RNA-Seq Example Part II
- 2:45 Break
- 3:00 Sharing, Publishing, and Reproducibility
- 3:25 Setting up your own Galaxy Cluster on Amazon
- 4:00 Done

bit.ly/NUgxy201502

Thanks



Dave Clements

Galaxy Project

Johns Hopkins University

clements@galaxyproject.org

Matt Schipma

NGS Core Facility

Center for Genetic Medicine (CGM)

Northwestern University

m-schipma@northwestern.edu

bit.ly/NUgxy201502

Cuffmerge

- Each Cufflinks run creates a set of transcript predictions.
- **Cuffmerge** unifies all those predictions into a single set.
- Makes this incredibly tedious task easy.