

Galaxy

University of Georgia
Athens, GA
March 7, 2014

Carl Eberhard, Dannon Baker, Dave Clements
Johns Hopkins University

Raj Ayyampalayam
University of Georgia

<http://galaxyproject.org/>



The University of Georgia®

University of Georgia

QBCG

Quantitative Biology Consulting Group

iob  
institute of bioinformatics

 **GAACRC**



The Agenda

8:30 Introduction

Welcome, Logistics, Galaxy Platforms, Galaxy 101

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 RNA-Seq Example

14:45 Break

15:00 Galaxy @ UGA

17:00 Done

The Agenda

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

Not The Agenda

This workshop will *not* cover

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

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

What is Galaxy?

- A free (for everyone) web service
- Open source software
- These options result in several ways to use Galaxy

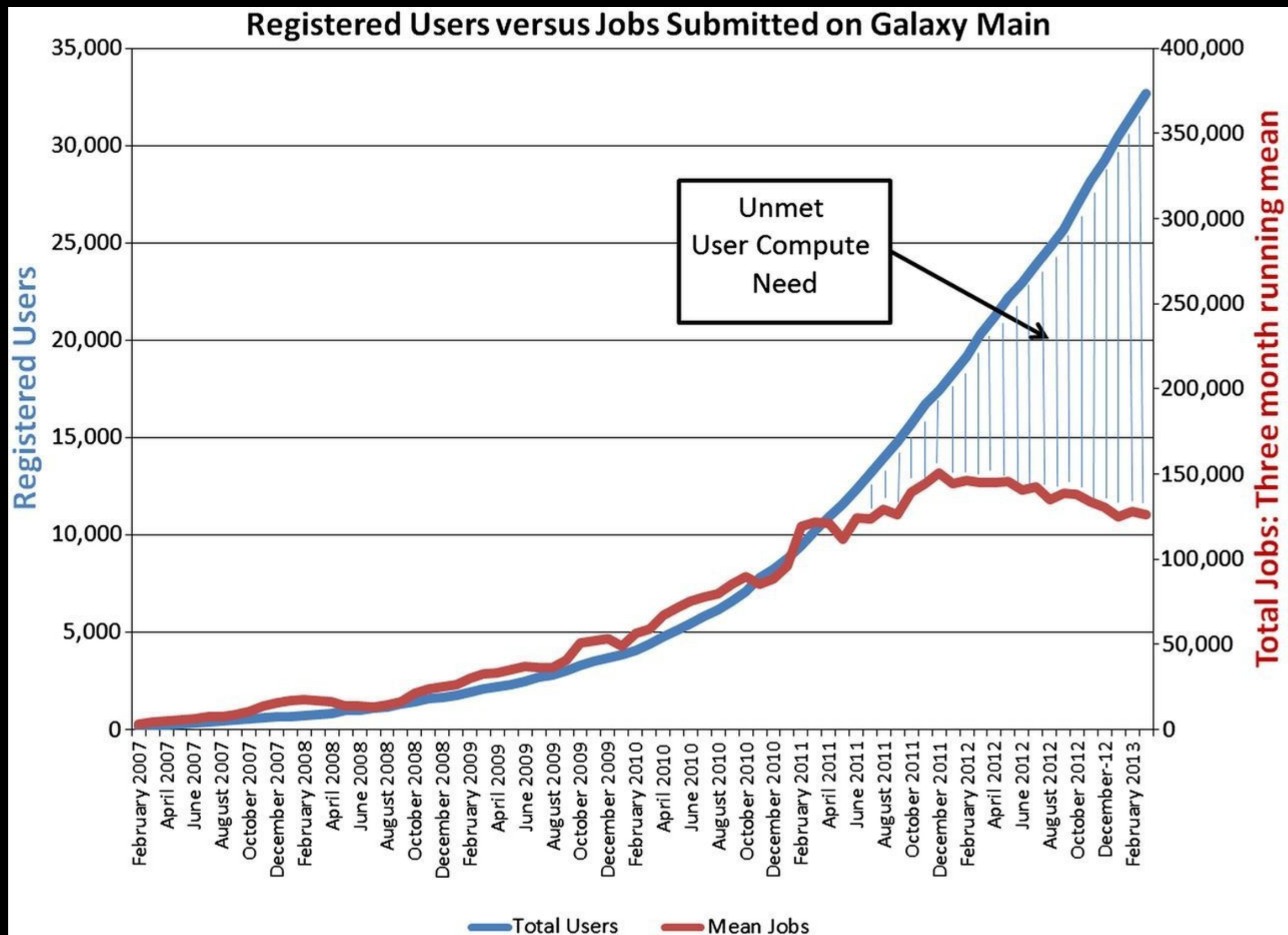
<http://galaxyproject.org>

Galaxy is available ...

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

<http://usegalaxy.org>

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



Leveraging the national cyberinfrastructure for biomedical research
 LeDuc, et al. *J Am Med Inform Assoc* doi:10.1136/amiajnl-2013-002059

Galaxy is available ...

- As a free (for everyone) web service

<http://usegalaxy.org>

- As open source software

<http://getgalaxy.org>

It is installed in locations around the world,
including:

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

Galaxy is available ...

- As a free (for everyone) web service

<http://usegalaxy.org>

- As open source software

<http://getgalaxy.org>

- ***On the Cloud***

We are using this today.

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

<http://globus.org/>

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



Galaxy is available ...

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



A ready-to-use appliance (BioTeam)

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

Consulting & Customization (Arctix, BioTeam, Deena Bioinformatics)

Galaxy Project: Further reading & Resources

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

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

<http://bit.ly/gxychoices>

Basic Analysis

Which genes have most overlapping
Repeats?

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

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

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

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

<http://cloud5.galaxyproject.org/>

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

Genes & Repeats: A General Plan

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

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

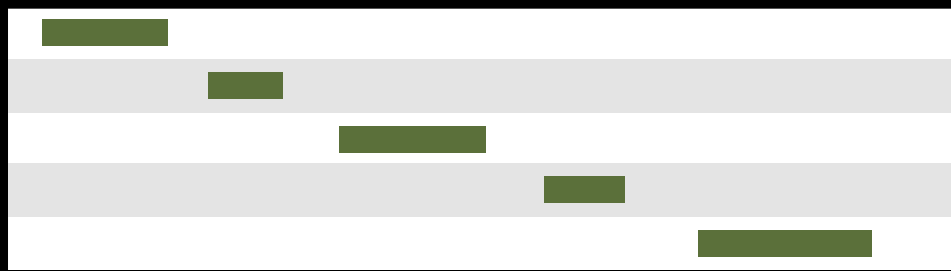
<http://cloud2.galaxyproject.org/>

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

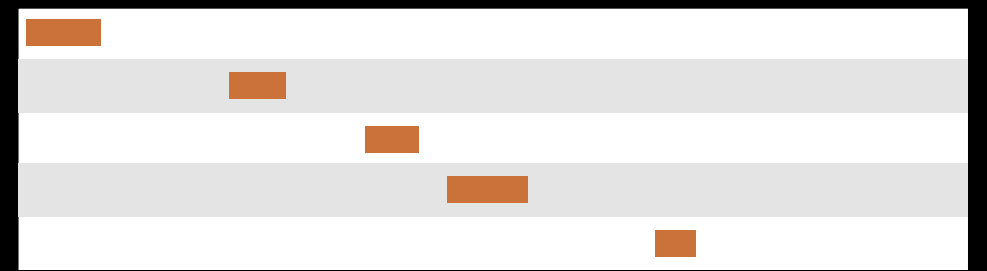
<http://cloud4.galaxyproject.org/>

<http://cloud5.galaxyproject.org/>

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

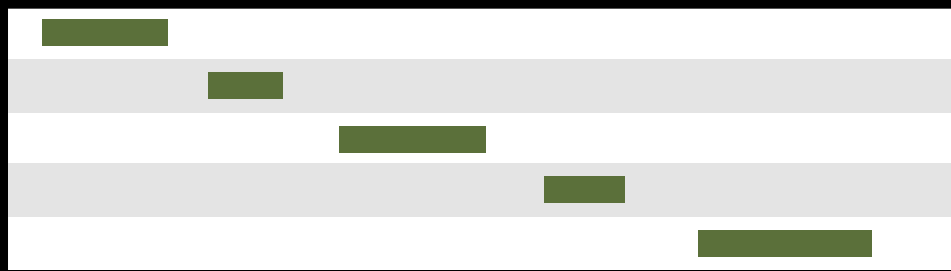


Exons

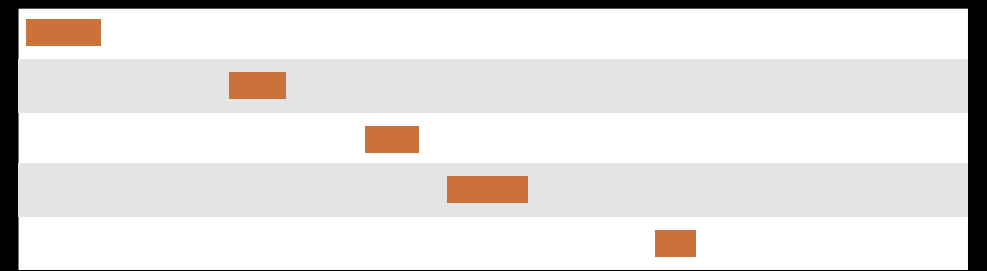


Repeats

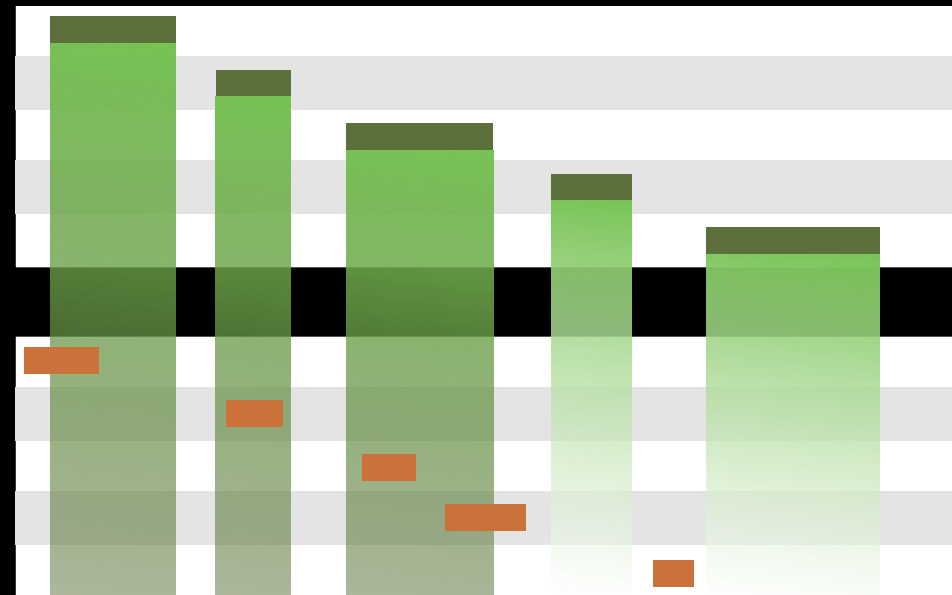
(Identify which genes/exons have Repeats)



Exons



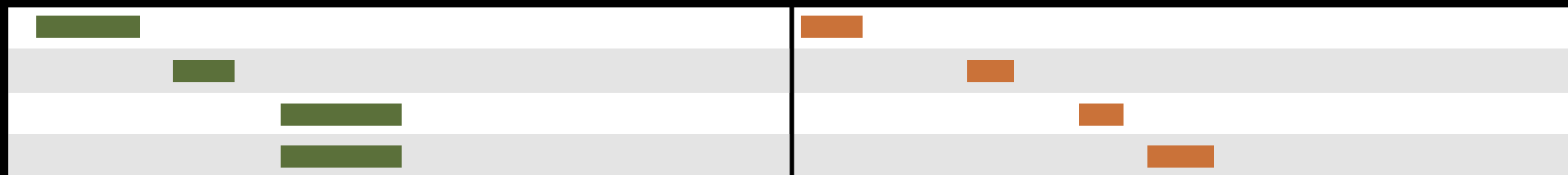
Repeats



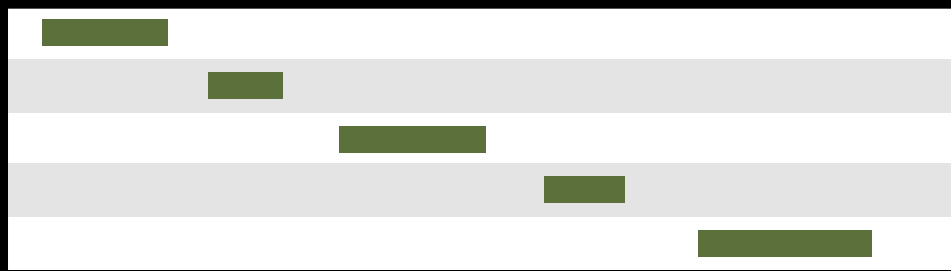
Exons

Repeats

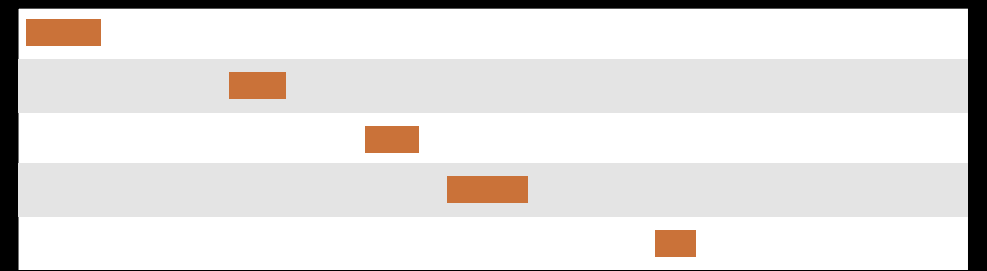
Overlap pairings



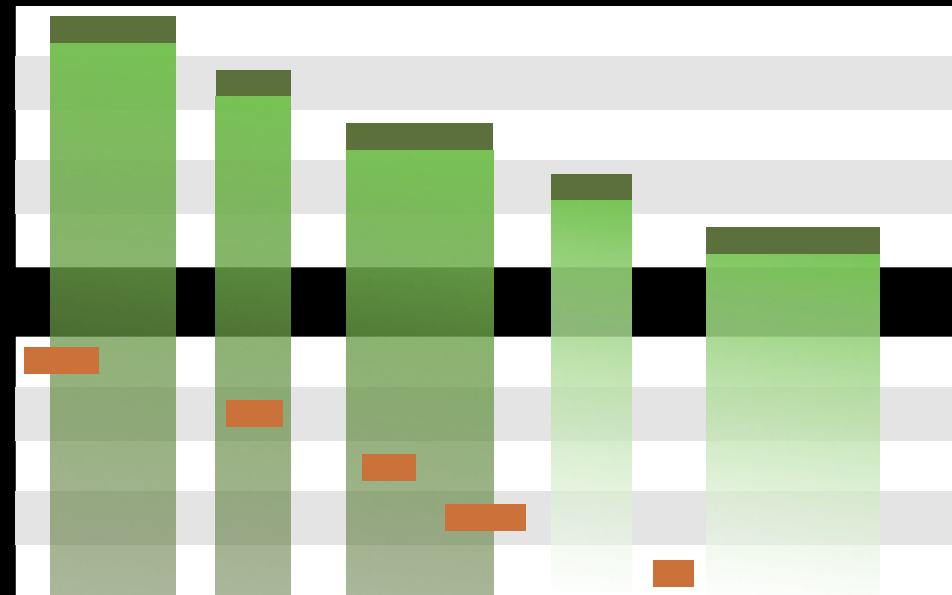
Operate on Genomic Intervals → Join
(Identify which genes/exons have Repeats)



Exons



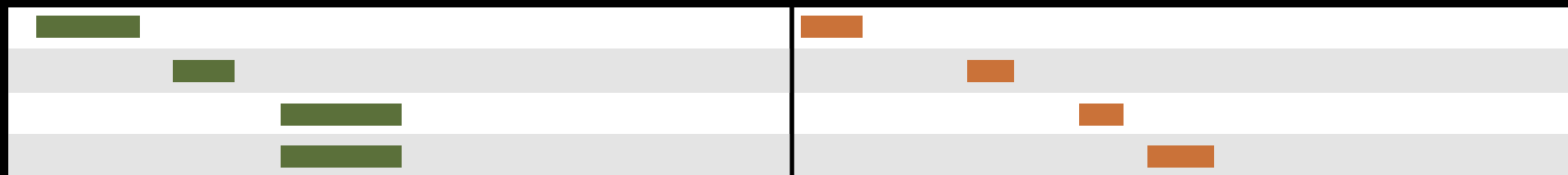
Repeats



Exons

Repeats

Overlap pairings

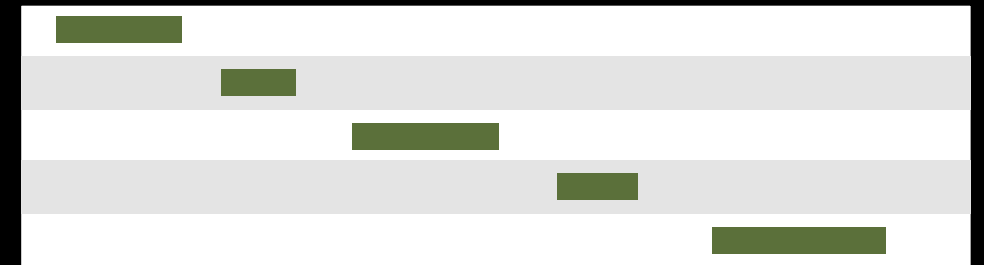


Exon overlap counts

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



Exon overlap counts

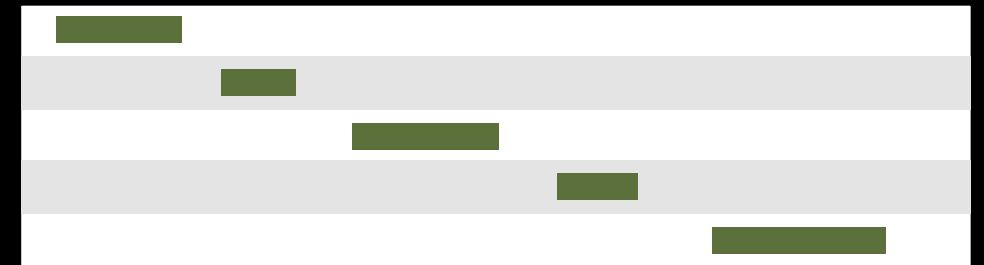


Exons

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

	1
	1
	2

Exon overlap counts



Exons

	1		0
	1		0
	2		0

Join on exon name

Join, Subtract, and Group → Join

(Incorporate the overlap count with rest of Exon information)

1	1
1	1
2	2

Exon overlap counts

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

Exons

Real cut

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>

The Agenda

8:30 Introduction

10:00 **Break**
With an exercise: Galaxy 101.5

10:15 Introduction continued

11:00 Lunch

12:45 RNA-Seq Example

14:45 Break

15:00 Galaxy @ UGA

17:00 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/>

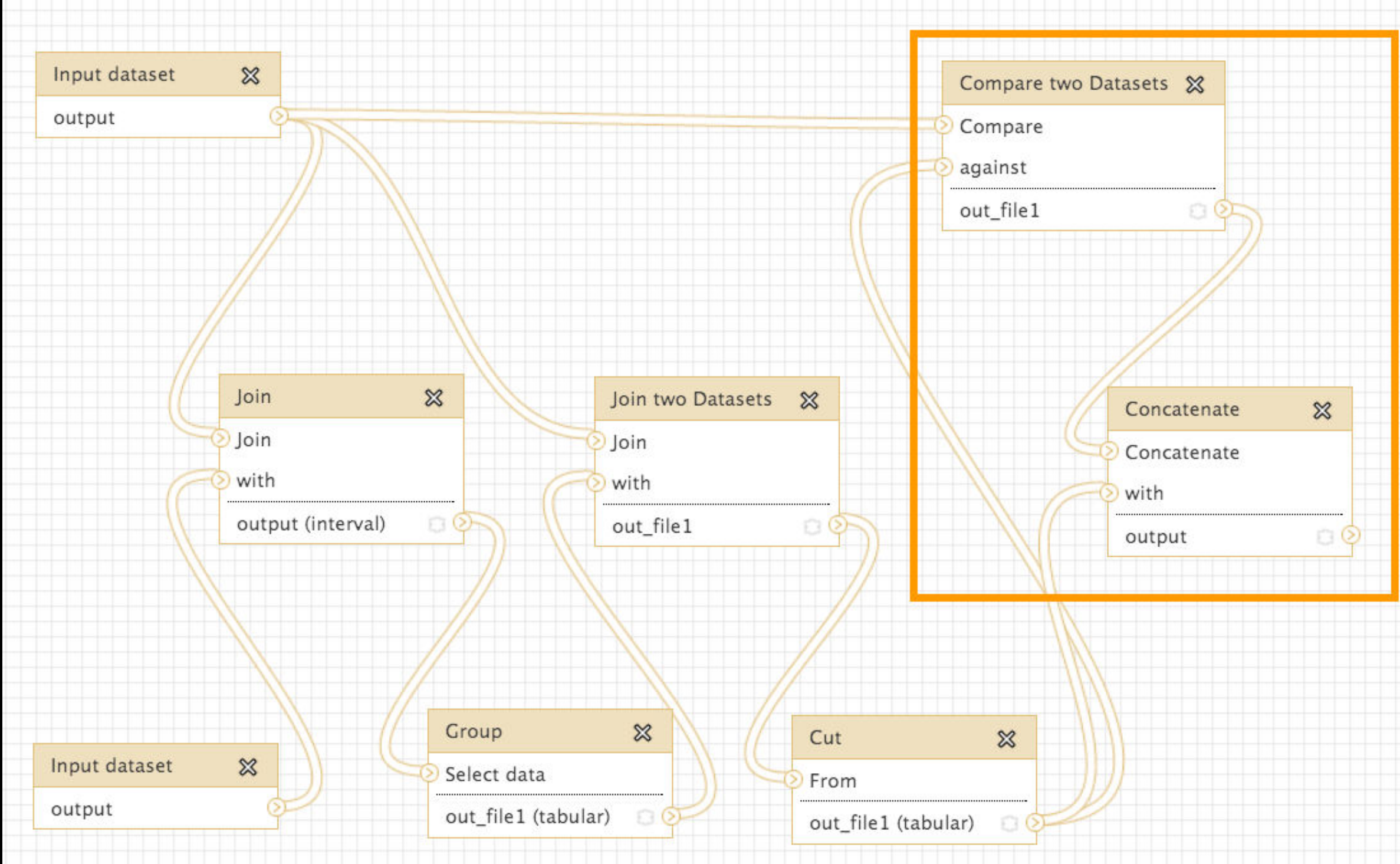
<http://cloud2.galaxyproject.org/>

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

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

<http://cloud5.galaxyproject.org/>

One Possible Solution



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

The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued
Repeatable workflows, maybe some QC

11:00 Lunch

12:45 RNA-Seq Example

14:45 Break

15:00 Galaxy @ UGA

17:00 Done

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

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

Workflow:

A series of analysis steps

Can be repeated with different data

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

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

Create a Workflow from a History

Extract Workflow from history

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



(cog) → Extract Workflow

Run / test it

Guided: rerun with same inputs

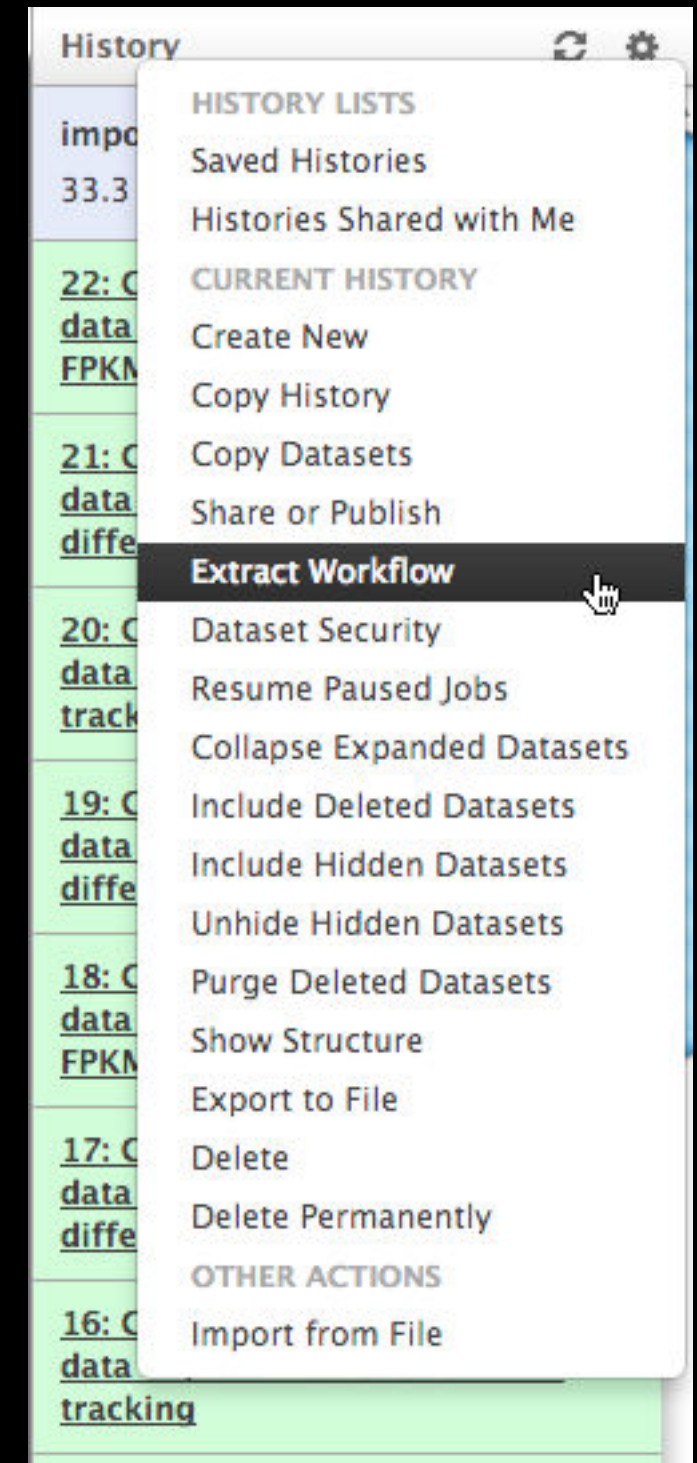
Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 RNA-Seq Example

14:45 Break

15:00 Galaxy @ UGA

17:00 Done

The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 **RNA-Seq Example**

QC, mapping, differential expression; Galaxy Community

14:45 Break

15:00 Galaxy @ UGA

17:00 Done

NGS Data Quality Control

- FASTQ format
- Examine quality in an RNA-Seq dataset
- Trim/filter as we see fit, hopefully without breaking anything.

Quality Control is not sexy.

It is vital.

What is FASTQ?

- Specifies sequence (FASTA) and quality scores (PHRED)
- Text format, 4 lines per entry

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
! ' ' * ( ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) ) * * 5 5 C C F > > > > > C C C C C C C 6 5
```

- **FASTQ is such a cool standard, there are 3 (or 5) of them!**

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|          |         |        |              |                                  |               |
33          59       64       73             104                                126

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

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

NGS Data Quality Exercise

Create new history



(cog) → Create New

Get some data

Shared Data → Data Libraries

→ RNA-Seq Example*

→ Untrimmed FASTQ

→ Select MeOH_REP1_R1, MeOH_REP1_R2
and then Import to current history



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

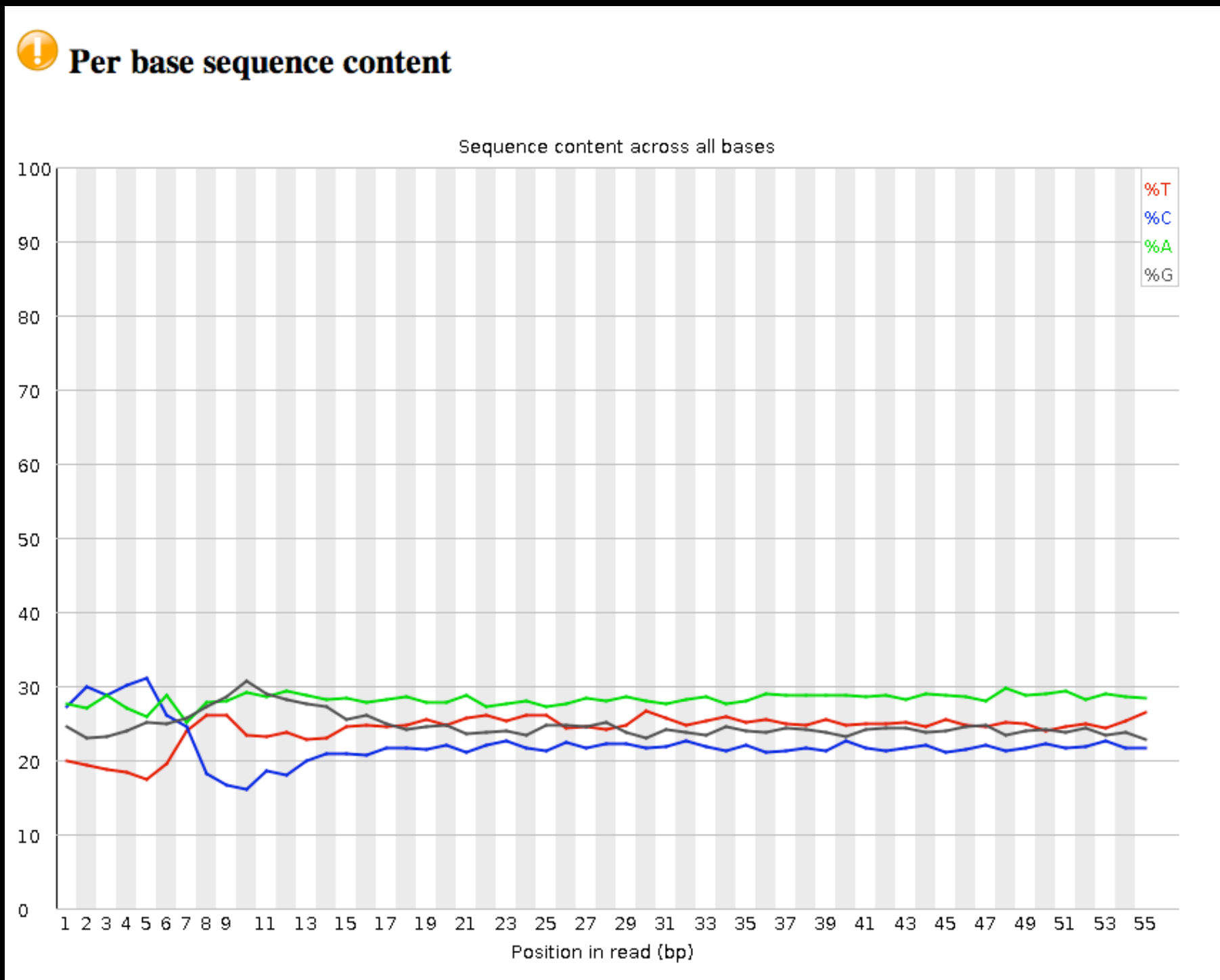
NGS Data Quality: Assessment tools

NGS QC and Manipulation → **FastQC**

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

<http://bit.ly/FastQCBoxPlot>

NGS Data Quality: Sequence bias at front of reads?

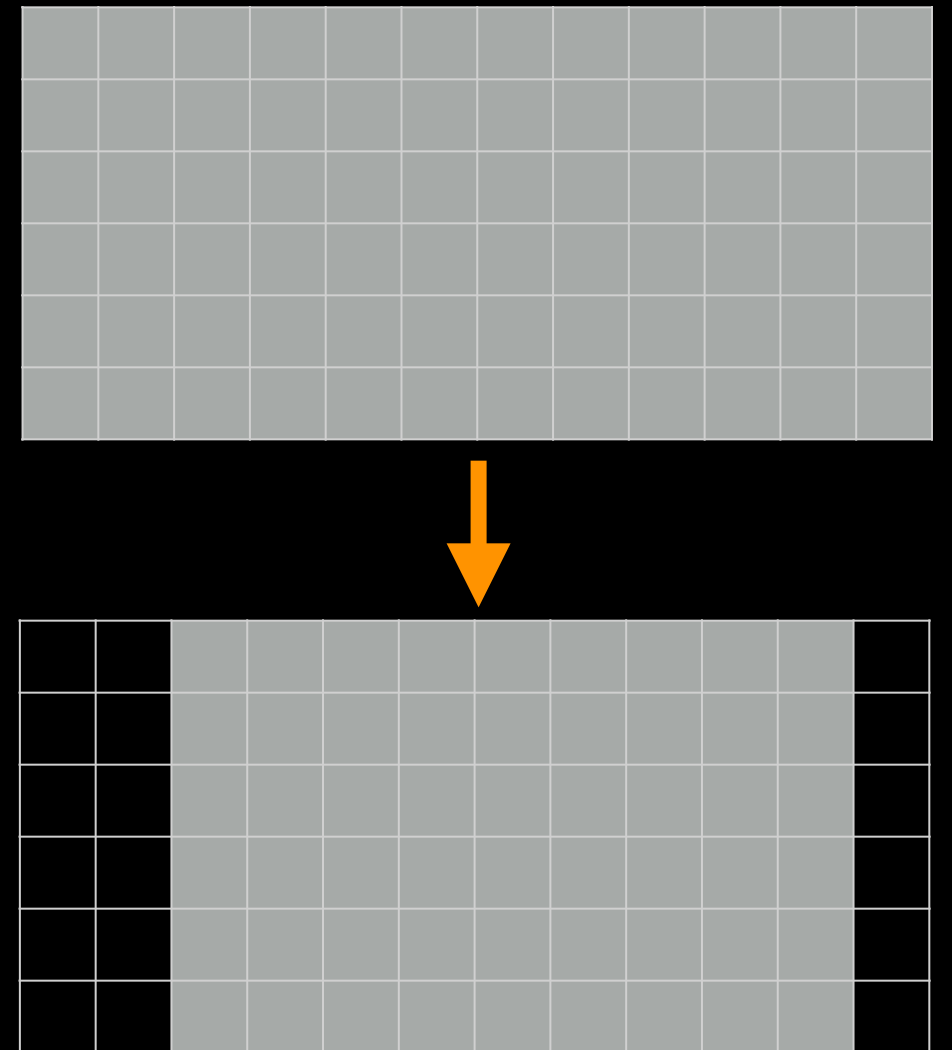


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

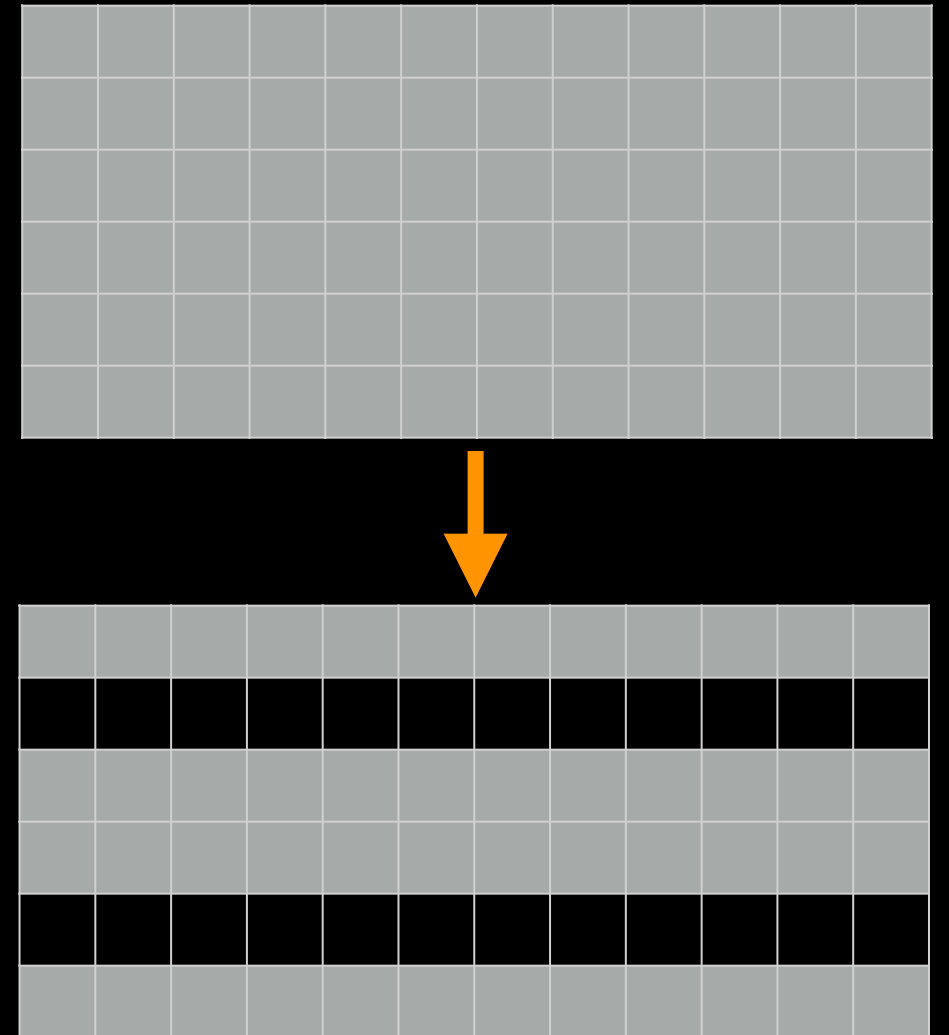
NGS Data Quality: Trim as we see fit

- Trim as we see fit: Option 1
 - NGS QC and Manipulation → **FASTQ Trimmer by column**
 - Trim same number of columns from every record
 - Can specify different trim for 5' and 3' ends



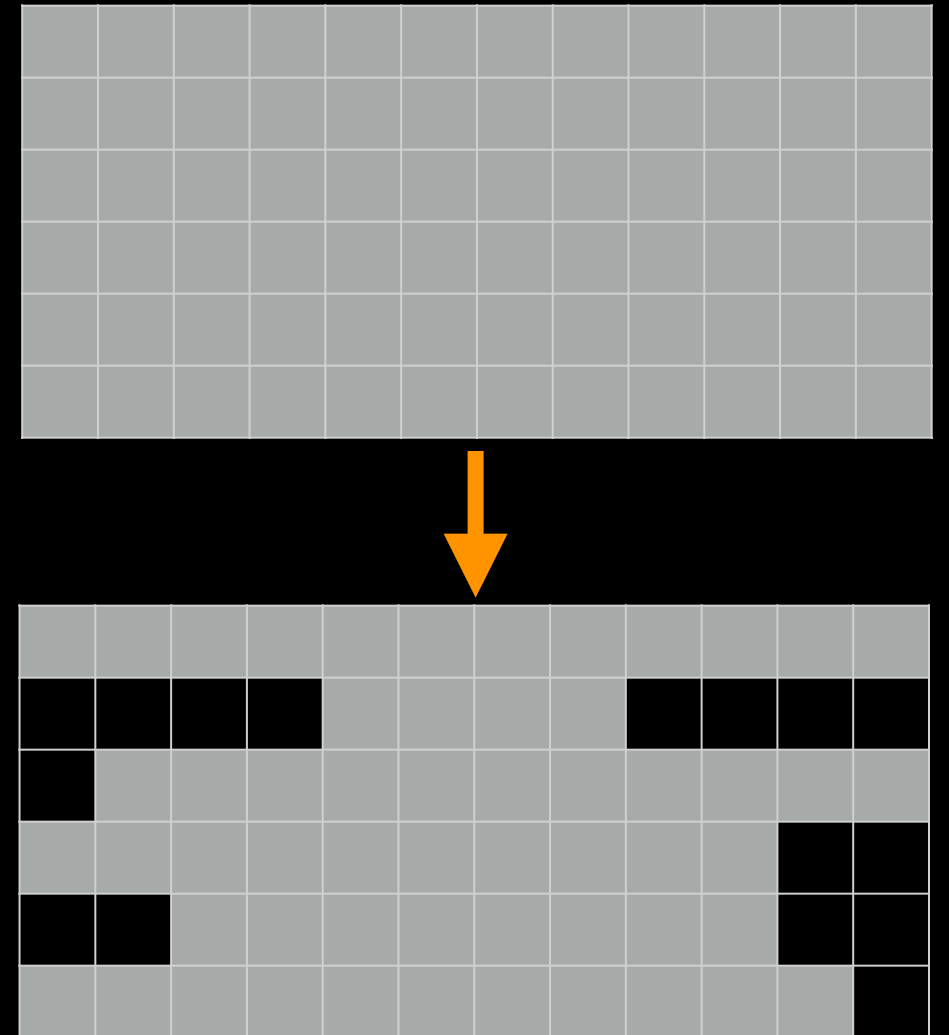
NGS Data Quality: Base Quality Trimming

- Trim Filter as we see fit: Option 2
- NGS QC and Manipulation →
Filter FASTQ reads by quality score and length
- **Keep or discard whole reads**
- Can have different thresholds for different regions of the reads.
- **Keeps original read length.**

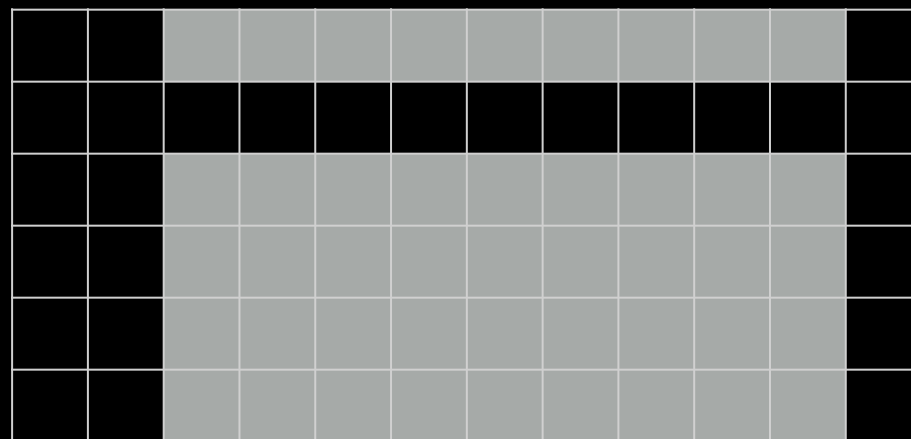
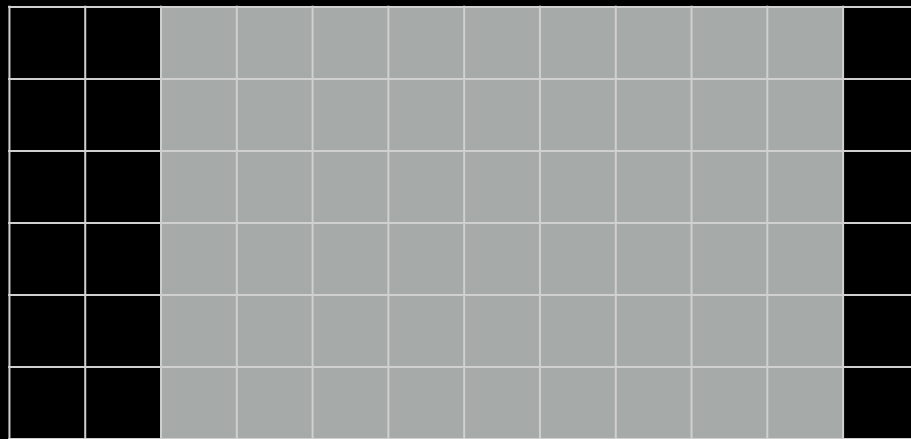
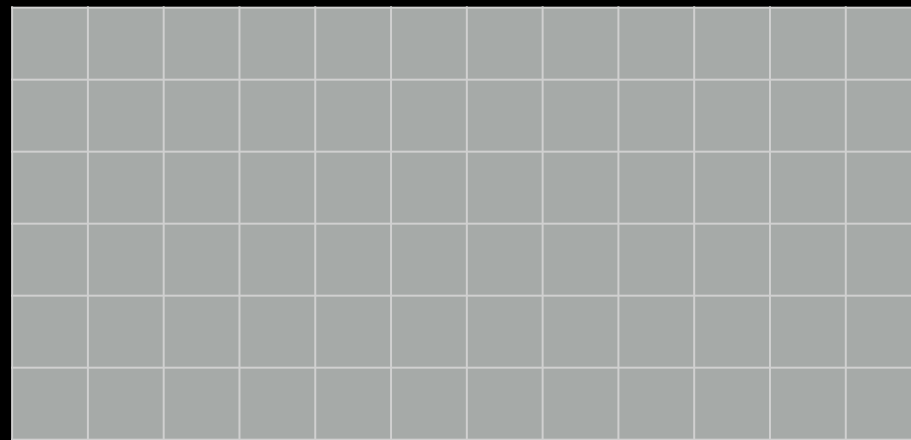


NGS Data Quality: Base Quality Trimming

- Trim as we see fit: Option 3
 - NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**
 - Trim from both ends, using sliding windows, until you hit a high-quality section.
 - **Produces variable length reads**



**Options are
not mutually
exclusive**



**Option 1
(by column)**

+

**Option 2
(by entire row)**

Trim? *As we see fit?*

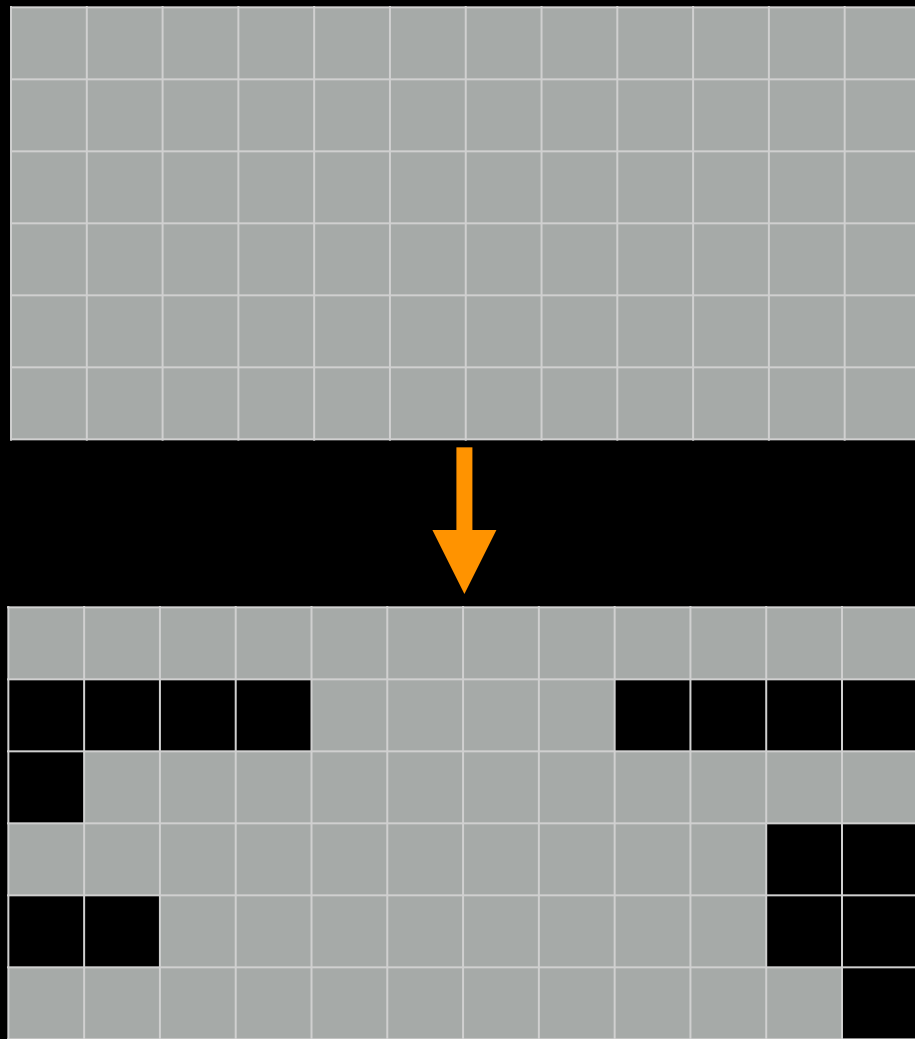
- Introduced 3 options
 - One preserves original read length, two don't
 - One preserves number of reads, two don't
 - Two keep/make every read the same length, one does not
- One preserves pairings, two don't

Trim? *As we see fit?*

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



NGS Data Quality: Base Quality Trimming



I really want to use Option 3:

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

but ...

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

Tophat Manual

“If you are performing RNA-seq analysis, there is no need to filter the data to ensure exact pairs before running Tophat.”

Jen Jackson

Galaxy User Support Person Extraordinaire

“Dang.”

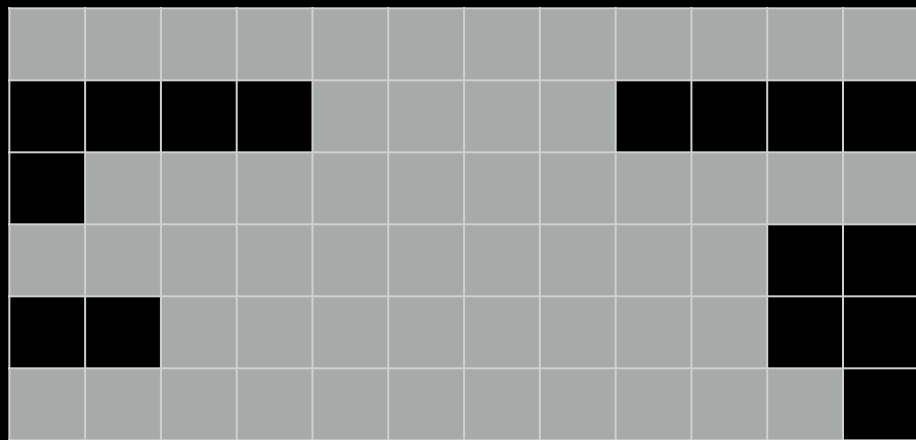
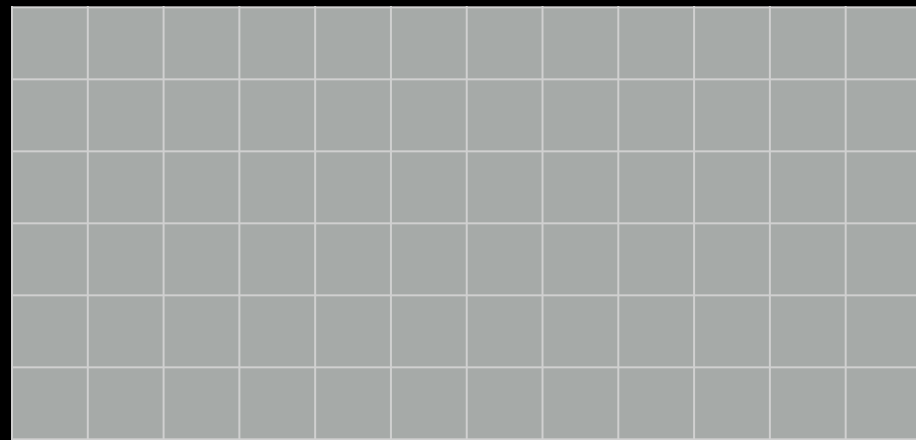
Most of us

Running Tophat on *no-longer-cleanly-paired* data *does map the reads*, but, it no longer keeps track of read pairs in the SAM/BAM file.

Keeping paired ends paired: Options

- Don't bother.
- Run a workflow that removes any unpaired reads before mapping.
- Run the Picard **Paired Read Mate Fixer** after mapping reads.
- Use sliding windows for QC, **but keep empty reads.**

NGS Data Quality: Base Quality Trimming



I'll use Option 3 (*but ...*):

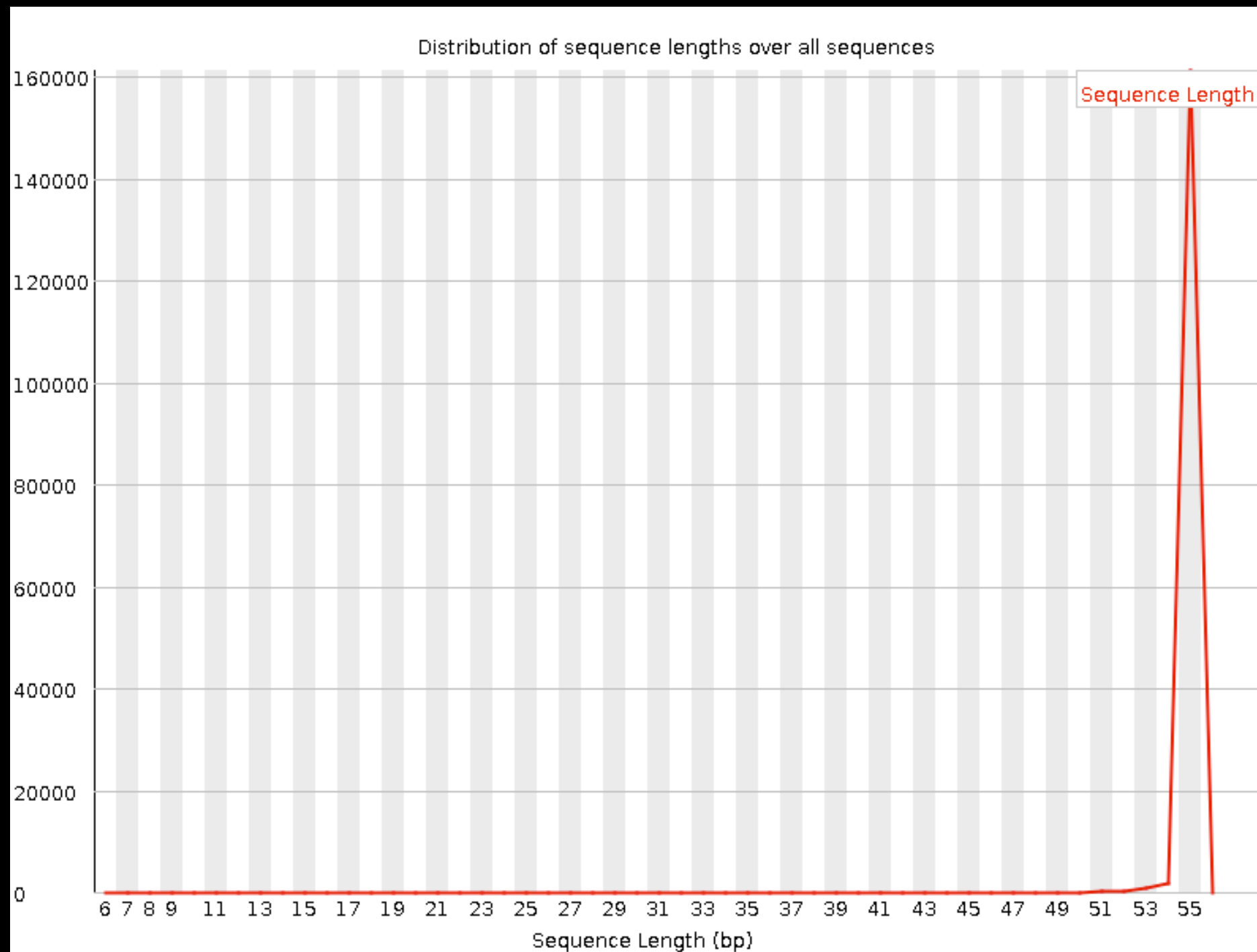
- NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**

Check "Keep reads with zero length"

Run again:

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

NGS Data Quality: Base Quality Trimming



New Problem?

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

Option 2 can fix this (but break pairings).

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

NGS Data Quality: Sequencing **Artifacts**

Repeat this process with MeOH Rep1 R2 (the reverse reads)
... and there's a problem in Overrepresented sequences:



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGTGTATTTGTCAATTTTCTTCTCCACGTTCTTCTCGGCCTGTTTCCGTAGCCT	590	0.3541692929220167	No Hit
TT	342	0.2052981325073385	No Hit
CGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	325	0.19509325457568719	No Hit
CTGCATTATAAAAAGGACAGCCAGATATCAACTGTTACAGAAATGAAATAAGACG	230	0.13806599554587093	No Hit
CGGCCGCAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	199	0.11945710049403614	No Hit
GTCAGCTCAACTTGTAGGCCCCAAAAGAAAACAGCGTCTTACTGGGGAGGGATAT	197	0.11825652661972422	No Hit

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

But this will break pairings.

NGS Data Quality: Done with 1st Replicate!

Now, only 3 (or 5) more to go!

Workflows:

Create a QC workflow that does all these steps

Or, cheat and import the shared workflow.

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

NGS Data Quality: Further reading & Resources

FastQC 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.*

The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 **RNA-Seq Example**

QC, **mapping**, differential expression; Galaxy Community

14:45 Break

15:00 Galaxy @ UGA

17:00 Done

RNA-seq Exercise: Mapping with Tophat

Create a new history

Import all datasets from library:

RNA-Seq Example → **Trimmed FASTQ**

all trimmed FASTQ and **genes_chr12.gtf**

NGS: RNA Analysis → **TopHat for Illumina**

RNA-seq Exercise: Mapping with Tophat

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

Mapping with Tophat: **mean inner distance**

Expected distance between paired end reads

- Determined by sample prep
- We'll use **90*** for **mean inner distance**
- We'll use **50** for **standard deviation**

* The library was constructed with the typical Illumina TruSeq protocol, which is supposed to have an average insert size of 200 bases. Our reads are 55 bases (R1) plus 55 bases (R2). So, the Inner Distance is estimated to be $200 - 55 - 55 = 90$

From the 2013 UC Davis Bioinformatics Short Course

Mapping with Tophat: Use Existing Annotations?

You can bias Tophat towards known annotations

- Use Own Junctions → Yes
 - Use Gene Annotation → Yes
 - Gene Model Annotation → genes_chr12.gtf
- Use Raw Junctions → Yes (tab delimited file)
- Only look for supplied junctions → Yes

Mapping with Tophat: **Make it quicker?**

Warning: Here be dragons!

- **Allow indel search** → **No**
- **Use Coverage Search** → **No** (wee dragons)

TopHat generates its database of possible splice junctions from two sources of evidence. The first and strongest source of evidence for a splice junction is when two segments from the same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found *ab initio*. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. **We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million).** This latter option will only report alignments across "GT-AG" introns

Mapping with Tophat: **Max # of Alignments Allowed**

Some reads align to more than one place equally well.

For such reads, how many should Tophat include?

If more than the specified number, Tophat will pick those with the best mapping score.

Tophat **break ties randomly**.

Tophat assigns equal fractional credit to all n

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

Mapping with Tophat: Lets do it some more!

NGS: RNA Analysis → Tophat

for the remaining replicates

Or not.

RNA-Seq Mapping With Tophat: Resources

RNA-Seq Concepts, Terminology, and Work Flows

by Monica Britton

Aligning PE RNA-Seq Reads to a Genome

by Monica Britton

both from the UC Davis 2013 Bioinformatics Short Course

RNA-Seq Analysis with Galaxy

by Jeroen F.J. Laros, Wibowo Arindrarto, Leon Mei

from the GCC2013 Training Day

RNA-Seq Analysis with Galaxy

by Curtis Hendrickson, David Crossman, Jeremy Goecks

from the GCC2012 Training Day

The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 **RNA-Seq Example**

QC, mapping, differential expression; **Galaxy Community**

14:45 Break

15:00 Galaxy @ UGA

17:00 Done

Galaxy Resources and Community: Mailing Lists

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

Galaxy-Announce

Project announcements, low volume, moderated

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

Galaxy-User

Questions about using Galaxy and usegalaxy.org


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

Galaxy-Dev


Questions about developing for and deploying Galaxy

High volume (5200 posts in 2013, 900+ members)

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

 **Galaxy Web Search**

Google™ Custom Search

Search 

Search the entire set of Galaxy web sites and mailing lists using Google.

[Run this search at Google.com \(useful for bookmarking\)](#)

Want a [different search](#)?

[Project home](#)

Find

Everything on ...

Tools for ...

Email about ...


Source code for ...

Published Histories, Pages, Workflows, about ...

Documentation on ...

Papers using Galaxy for ...

Related feature requests

 **Galaxy Web Search**

chip-seq

All Tools Email Source code Shared Documentation Abstracts Requests

About 444 results (0.06 seconds)

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

Community: Public Galaxy Instances

<http://bit.ly/gxyServers>

Interested in:

ChIP-chip and ChIP-seq?

✓ Cistrome, Nebula

Statistical Analysis?

✓ Genomic Hyperbrowser

Protein synthesis?

✓ GWIPS-viz

de novo assembly?

✓ CBIIT Galaxy

Reasoning with ontologies?

✓ GO Galaxy

Repeats!

✓ RepeatExplorer

Over 50 public Galaxy servers

Community can create, vote and comment on issues

The screenshot displays the Trello interface for the 'Galaxy: Development' board. The board is organized into several columns, each representing a different category of issues or tasks. The columns are: Inbox, Tool Requests, Bug Reports, Ideas, Pull Requests / Patches, and Project in Planning. Each card in the board contains a title, a description, and a progress bar. The cards are sorted by their position in the column, with the most recent or important cards at the top. The right sidebar shows the 'Menu' section, which includes 'Members' (a list of users with their profile pictures), 'Activity' (a list of recent actions), and 'Tools' (a list of tools used in the board). The top navigation bar includes links to 'HOME', 'TOUR', 'GOLD', 'BUSINESS CLASS', and 'BLOG'. The top right corner has 'Sign Up' and 'Log In' buttons. The bottom of the board has a 'Sign up for free' button and a link to 'learn more about Trello'.

Galaxy: Development • Public

Want to subscribe, vote or comment on these cards? [Sign up for free](#) or [learn more about Trello](#)

Inbox

- To add cards, use <http://galaxyproject.org/trello> (4 votes, 1 comment)
- add ma seq metrics and downsample sam to picard tools (3 votes, 1 comment)
- Reference genomes (2 votes, 2 comments)
- Please merge patch to bowtie2 wrapper (add support for mapping fasta files) (1 vote, 1 comment)
- R 3.0.2 woes on test.g2.bx.psu.edu - libgomp.so not found when job runs: [Rscript error while loading shared libraries: libgomp.so.1: cannot open shared object file: No such file or directory] - possibly execution node missing gfortran? (1 vote, 1 comment)
- unhandled exception when installing metaphlan from source repo (1 vote, 1 comment)

Tool Requests

- 595: Add SAMTools "Sort" (5 votes, 1 comment)
- 601: SAM-to-BAM tool enhancements (2 votes, 1 comment)
- Bug: some characters not permitted in 'add column' tool (2 votes, 5 comments)
- 307: A tool to produce a set of random intervals. (2 votes, 2 comments)
- Tool: Add tool to generate simulated reads to Main (2 votes, 1 comment)
- default max insert size of Bowtie2 should be increased (1 vote, 4 comments)
- Wrapper for bigWigToWig (1 vote, 1 comment)
- Converter Tool: SAM to BAM enhancements (1 vote, 1 comment)
- 607: Create new tool to "trim" coordinates to ref chrom lengths (1 vote, 1 comment)
- New Tool: convert IUPAC chars to N (5 votes, 1 comment)
- Optimize FASTQ tools. (1 vote, 1 comment)
- Tool 'Extract Genomic DNA' should parse GFF/GTF better so to include gene_id or transcript_id attributes (1 vote, 1 comment)
- Enh: tabular-to-fasta should let you choose how to concatenate the id string (1 vote, 1 comment)

Bug Reports

- Impersonate a user admin option broken when using external authentication (13 votes, 1 comment)
- Bug: SIGER on Main dependency issue (2 votes, 18 comments, 3/5)
- Toolshed: Installing multiple versions of the same tool results in separate entries in the tool panel. (1 vote, 14 comments)
- Profile Annotations bad values when "select all" (1 vote, 2 comments)
- The option from_file="internal.log" is broken. (1 vote, 1 comment)
- 68: Apparent bug in Intersect intervals, overlapping pieces (5 votes, 1 comment)
- 106979439 108792355 1: problem: SAMTOOLS is using "BAM" for the ID of this file (536670812) (1 vote, 1 comment)
- Bug: Returning Bitset error 536670812 (4 votes, 1 comment)
- EMBOSS: several tools fail with default options (4 votes, 3 comments)
- Tools: Cloudmap reference files not found (4 votes, 2 comments)
- Run: Patch taxonomi (1 vote, 1 comment)

Ideas

- Implement JavaScript build process (1 vote, 6 comments, 0/13)
- Tools: Incorporate key Cuffdiff output files for Cumberbund (1 vote, 1 comment, 0/3)
- Workflow Editor: Provide explicit access to implicit datatype converter tools (1 vote, 1 comment, 0/3)
- Google Drive / Dropbox / Box / ... integration (6 votes, 3 comments)
- 720: Capture and report time taken to run each job (8 votes, 0/2)
- Allow administrators to "trust" certain HTML outputs based on tool producing them. (4 votes, 4 comments)
- Workflows: highlight the noodles in the workflow editor upon hovering (4 votes, 2 comments, CE)
- 5: Option to disable automatic history creation (4 votes, 2 comments, CE)
- Allowing workflow step dependencies when no input/output files exist (4 votes, 1 comment)
- Assistive UI (4 votes, 1 comment, 0/4, CE)
- For sensible output. Add input name to Son_string (4 votes, 0/3)
- RFC: Implement sophisticated user behavior analysis tool (1 vote, 1 comment)

Pull Requests / Patches

- 685: Patch for FASTQ paired-end issue (1 vote, 1 comment)
- Tools: Bowtie Wrapper Pull Requests from Community (2 votes, 6 comments, 1 comment)
- Pull Request #343 - Need to traverse the other_value dict to find dependencies for ParamValueFilter in dynamic_options when the dependencies are scoped in a conditional. Error was noted attempting to run iuc SnpEff 3.4 in a workflow. (0/3)
- Pull Request #336 - Patch to expose the actual dataset id in the LDDA and HDA to_dict calls (in addition to the instance id). (0/3)
- Pull Request #336 - Traverse context for SelectToolParameter need_jate_validation. (0/3)
- Pull Request #334 - Trello Card #1437: Optional Input Datasets Not Compatible with Parallelism Tag (0/3)
- Pull Request #281 - tools/fastq/fastq_pair_end_joiner: added support for recent Illumina headers (0/3)

Project in Planning

- 308: Demystifying the first ever Galaxy login experience - make tools offer test data if empty history? (3 votes, 2 comments)
- Data Manager: Genome Builds / dbkeys: Make adding builds accessible by Data Manager tools (2 votes, 0/3)
- resetting the password deactivates the user (1 vote, 4 comments)
- Tools: Moving to BAM format as primary representation of sequence data (1 vote, 2 comments)
- Libraries: Role selection (1 vote, 1 comment)
- Data Manager: Rsync version (1 vote, 1 comment)
- UI enhancements (1 vote, 0/7)
- BWA aln -n param update (2 votes, 2 comments)
- Show placement in queue / throughput (1 vote, 1 comment)
- Core: Make the jobs admin interface not suck (6 votes, 1 comment)
- Deleting history using the API does not delete/stop jobs (6 votes, 7 comments, CE)
- Tool Shed (and Galaxy?) should have user profiles. (3 votes, 5 comments)

Members

- CE
- NS
- DF
- G

Activity

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

<http://bit.ly/gxytrello>


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!](#)



BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Early Registration & Abstract Submission are now open




24-25 March, Melbourne

Use Galaxy


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

Communicate

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

Deploy Galaxy

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



Galaxy made easy.

Contribute

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

Events

News

Galaxy Wiki
DaveClements Settings Logout | Search:

Events

Galaxy Event Horizon

Events with Galaxy-related content are listed here.

Also see the [Galaxy Events Google Calendar](#) for a listing of events and deadlines that affect 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](mailto:outreach@galaxyproject.org).

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

Upcoming Events

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

Galaxy Wiki

News

News

Announcements of interest to the Galaxy Community from the Galaxy Team or the Galaxy Community, 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 started with the RSS feed. Older news items are available in the [archive](#).

See also

- [Galaxy News Briefs](#)
- [Galaxy Updates](#)
- [Galaxy on Twitter](#)
- [Events](#)
- [Learn](#)
- [Support](#)
- [About the Galaxy Project](#)

News Items


January 2014 CloudMan Release

We just released an update to Galaxy CloudMan – an easy way to get a personal and completely private Galaxy instance in the cloud in just a few minutes, without needing to install anything locally.

This update brings a large number of updates, with the most prominent ones being:


- [Improved security](#)
- [New features](#)
- [Bug fixes](#)

[Adam Kraut](#), [Nate Coraor](#),
[Anushka Brownley](#), [Tristan Lubinski](#), [James Reaney](#)

 Galaxy Wiki
Login | Search:

News

Announcements of interest to the Galaxy Community. These can include items from the Galaxy Team or the Galaxy community and can address anything that is of wide interest to the community.

The Galaxy News is also available as an [RSS feed](#) .

See [Add a News Item](#) below for how to get an item on this page, and the RSS feed. Older news items are available in the Galaxy [News Archive](#).

See also

- [Galaxy News Briefs](#)
- [Galaxy Updates](#)
- [Galaxy on Twitter](#)
- [Events](#)
- [Learn](#)
- [Support](#)
- [About the Galaxy Project](#)

News Items

[January 2014 CloudMan Release](#)
[GCC2014 Training Day Topics: Vote!](#)
[January 2014 Galaxy Update](#)
[2013 Galaxy Day Report](#)
[Galaxy Community Log Board](#)
[Galaxy Deployment Catalog](#)
[Nominate 2014 Training Day Topics](#)
[December 2013 Galaxy Update](#)
[Nov 04, 2013 Galaxy Distribution](#)
[November 2013 Galaxy Update](#)
[December Bioinformatics Boot Camps](#)
[GCC2014: Save These Dates!](#)
[Galaxy Day, 4 décembre à Paris](#)


[News Archive](#)

News Items

January 2014 CloudMan Release

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

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


CloudMan



GALAXY

COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

<http://bit.ly/gcc2014>



Galaxy Australasia Workshop

2
0
1
4

We also support
community
organized efforts
and events.



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 in the top right. The channel name "Galaxy Project" is displayed with a "PLUS" badge and a note "Joined 1 month ago". Below this, a statistics bar shows: 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. On the left side, there is a "Settings" button and a paragraph of text describing the Galaxy project as an open, web-based platform for data-intensive biomedical research, supported by various institutions.

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**
usegalaxy.org
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](#) [MyCiteULike](#) [Group: Galaxy](#) [Search](#) Logged in as [galaxyproject](#) [Log Out](#)

Group: Galaxy - library 1437 articles

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

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

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

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

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

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

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

☐ **✓ Web-based Workflow Planning Platform Supporting the Design and Execution of Complex Multiscale Cancer Models**
pp. 1-1, [doi:10.1109/jbhi.2013.2297167](#)

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

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

Over
1400
papers

17 tags

<http://bit.ly/gxycul>

The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 **RNA-Seq Example**

QC, mapping, **differential expression**; Galaxy Community

14:45 Break

15:00 Galaxy @ UGA

17:00 Done

Cuffdiff?

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

NGS: RNA Analysis → Cuffdiff

Cuffdiff?

Cuffdiff uses FPKM/RPKM as a central statistic.
Total # mapped reads heavily influences FPKM/RPKM.
Can lead to challenges when you have very highly
expressed genes in the mix.

Cuffdiff Alternatives

Rapaport, *et al.*, "Comprehensive **evaluation of differential gene expression analysis** methods for RNA-seq data."

Genome Biology 2013, 14:R95 doi:10.1186/gb-2013-14-9-r95

Reviews **7 packages**

Each tool has it's own strengths and weaknesses.

What's a biologist to do?

Alternatives: What's a biologist to do?

Learn the strengths and weaknesses of the tools you have ready access to. Are they a good match for the questions you are asking?

If not, then research alternatives, identify good options and then work with your bioinformatics/systems people to get access to those tools.

Cuffdiff Alternatives: DESeq

DESeq is an R based differential expression analysis package where expression analysis is much more effectively isolated between features.

Cuffdiff Alternatives: DESeq

Takes a simple, tab delimited list of features and read counts across different samples.

First, have to create that list.

htseq-count

Is a tool that walks BAM files producing these lists

Cuffdiff Alternatives: DESeq

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

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

Cut out the duplicate gene name columns

NGS: RNA Analysis → DE Seq

Cuffdiff Alternatives: DESeq

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

How many genes have an adjusted P value <
0.05 ?

Differential Expression: Reading & Resources

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data
by Rapaport, *et al.*

DESeq Reference Manual

DESeq Galaxy Wrapper
by Nikhil Joshi

htseq-count Galaxy Wrapper
by Lance Parsons

The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 RNA-Seq Example

14:45 **Break**
(Almost)

15:00 Galaxy @ UGA

17:00 Done

The Galaxy Team



Enis Afgan



Dannon Baker



Dan Blankenberg



Dave Bouvier



Marten Cech



John Chilton



Dave Clements



Nate Coraor



Carl Eberhard



Dorine Francheteau



Jeremy Goecks



Sam Guerler



Jen Jackson



Greg von Kuster



Ross Lazarus



Anton Nekrutenko



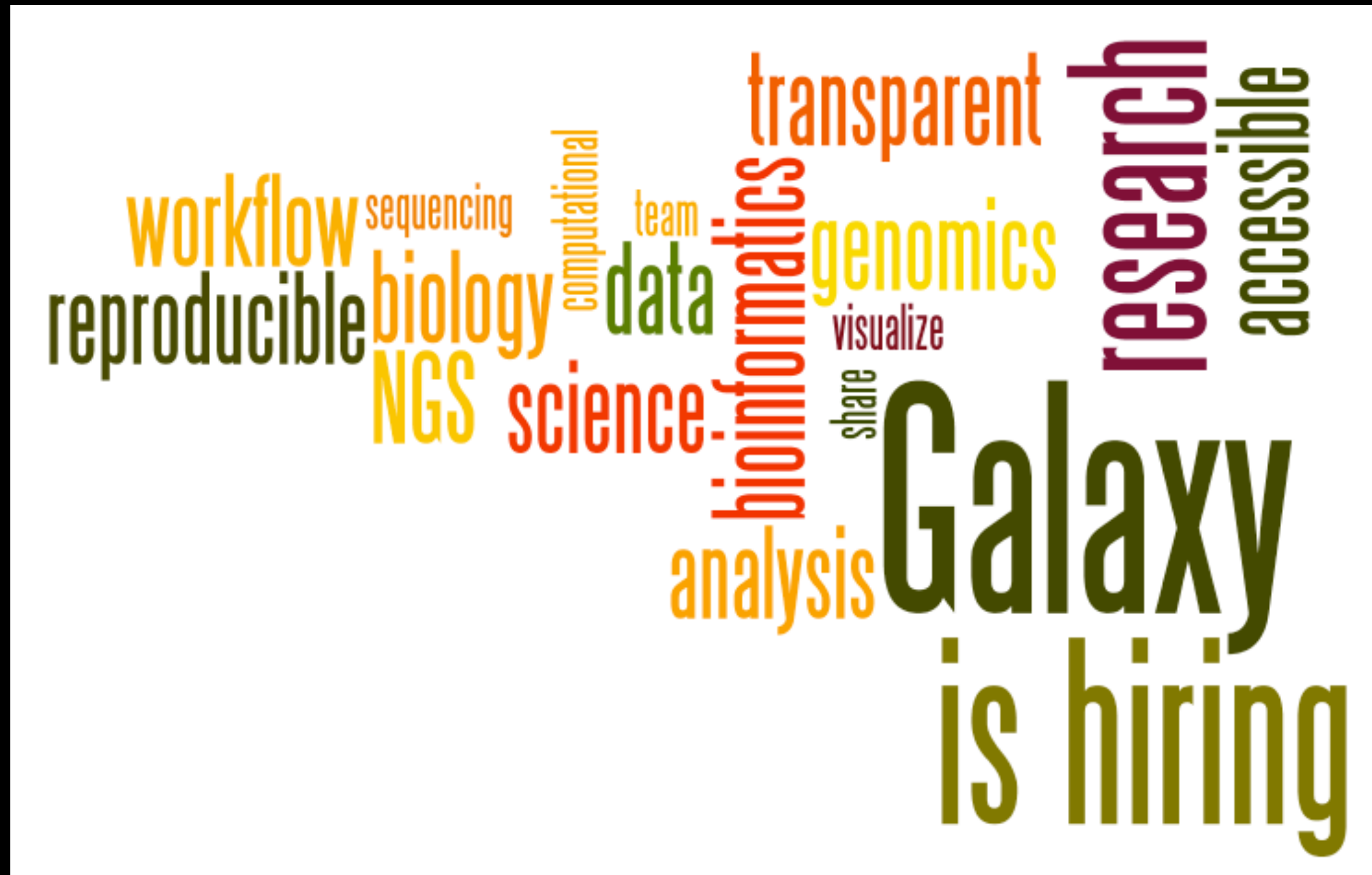
Nick Stoler



James Taylor

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

Galaxy is hiring post-docs and software engineers



Please help.

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

Also Thanks To



Jessie Kissinger
Raj Ayyampalayam
Carrie Jarrard

Thanks



Dave Clements

Dannon Baker

Carl Eberhard

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org

The Agenda

8:30 Introduction

10:00 Break

10:15 Introduction continued

11:00 Lunch

12:45 RNA-Seq Example

14:45 **Break**
(Really)

15:00 Galaxy @ UGA

17:00 Done