Introduction to **Galaxy** and **RNA-Seq** workshop

Enis Afgan
Johns Hopkins University

Slides available at bit.ly/Gxy-RNA-Seq-J15

The Agenda

Galaxy project overview: the core pillars

RNA-Seq exercise

What is Galaxy?

- **A data analysis and integration** tool
- **A (free for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple

Galaxy project: fundamental questions

When Biology (or any science) becomes dependent on computational methods:

- How can those methods best be made accessible to scientists?
- How best to ensure that analyses are reproducible?
- How best to facilitate transparent communication and reuse of analyses?

https://usegalaxy.org

Accessibility

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Tools

search tools

- Get Data
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- Genome Diversity

NGS TOOLBOX BETA

- Phenotype Association
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- NGS: Peak Calling
- NGS: RNA-seq

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.

Want help?
Get answers.

Biostars
GALAXY EXPLAINED

Tweets Follow

Guillaume Lobet @guillaumelobet 16 Jan
Great presentation of #usegalaxy by @jxtx at #NGS15 #openscience #reproducibility
Retweeted by Galaxy Project
Expand

Mark Veugelers @biotechbase 16 Jan
#ngs15: JT on future developments for Galaxy, e.g. analysis of 10.000+ samples
Retweeted by Galaxy Project
Expand

Mark Veugelers @biotechbase 16 Jan
#ngs15: James Taylor now speaking on reproducible genomics withy Galaxy. How to get reproducibility ? Easy. #use

Tweet to @galaxyproject

History

search datasets

Unnamed history

0 bytes

This history is empty. You can load your own data or get data from an external source

PENNSTATE 1855

JOHNS HOPKINS UNIVERSITY

TACC

iPlant Collaborative

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the iPlant Collaborative at the Texas Advanced Computing Center, with support from the

Accessibility

Need an analysis? There's a tool for that.

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- EBI SRA ENA SRA
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- Wormbase test server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- EpiGRAPH test server

NGS: Mapping

- Lastz map short reads against reference sequence
- Lastz paired reads map short paired reads against reference sequence
- Map with Bowtie for Illumina
- Map with Bowtie for SOLiD
- Map with BWA for Illumina
- Map with BWA for SOLiD
- Map with BFAST
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output
- Map with PerM for SOLiD and Illumina
- Re-align with SRMA
- Map with Mosaik

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat for SOLiD Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- Depth of Coverage on BAM files

REALIGNMENT

- Realigner Target Creator for use in local realignment
- Indel Realigner – perform local realignment

BASE RECALIBRATION

- Count Covariates on BAM files
- Table Recalibration on BAM files
- Analyze Covariates – draw plots

GENOTYPING

- Unified Genotyper SNP and indel caller

ANNOTATION

- Variant Annotator

FILTRATION

- Variant Filtration on VCF files

VARIANT QUALITY SCORE RECALIBRATION

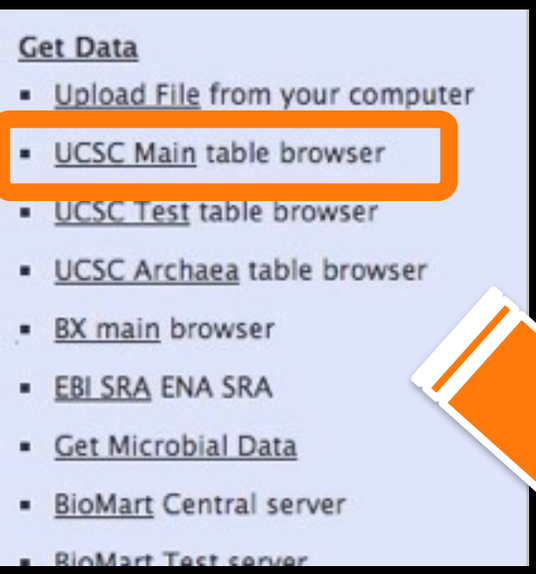
- Variant Recalibrator
- Apply Variant Recalibration

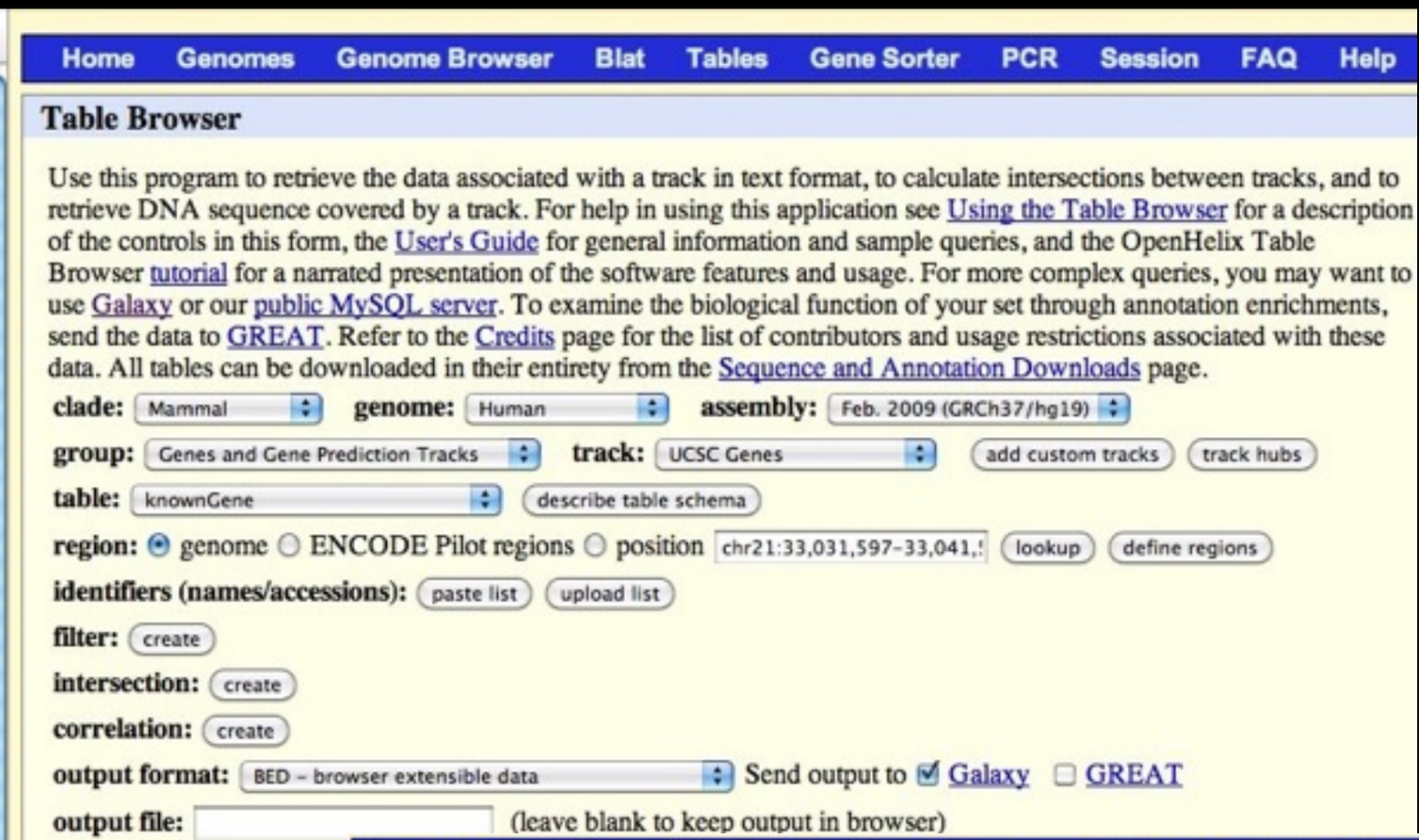
VARIANT UTILITIES

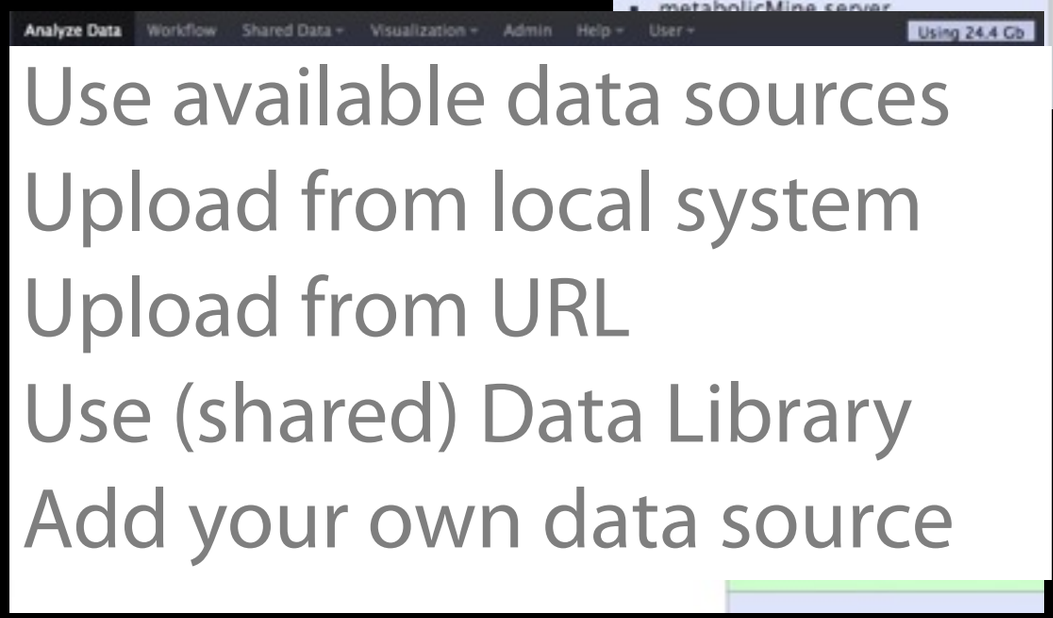
- Validate Variants
- Eval Variants
- Combine Variants

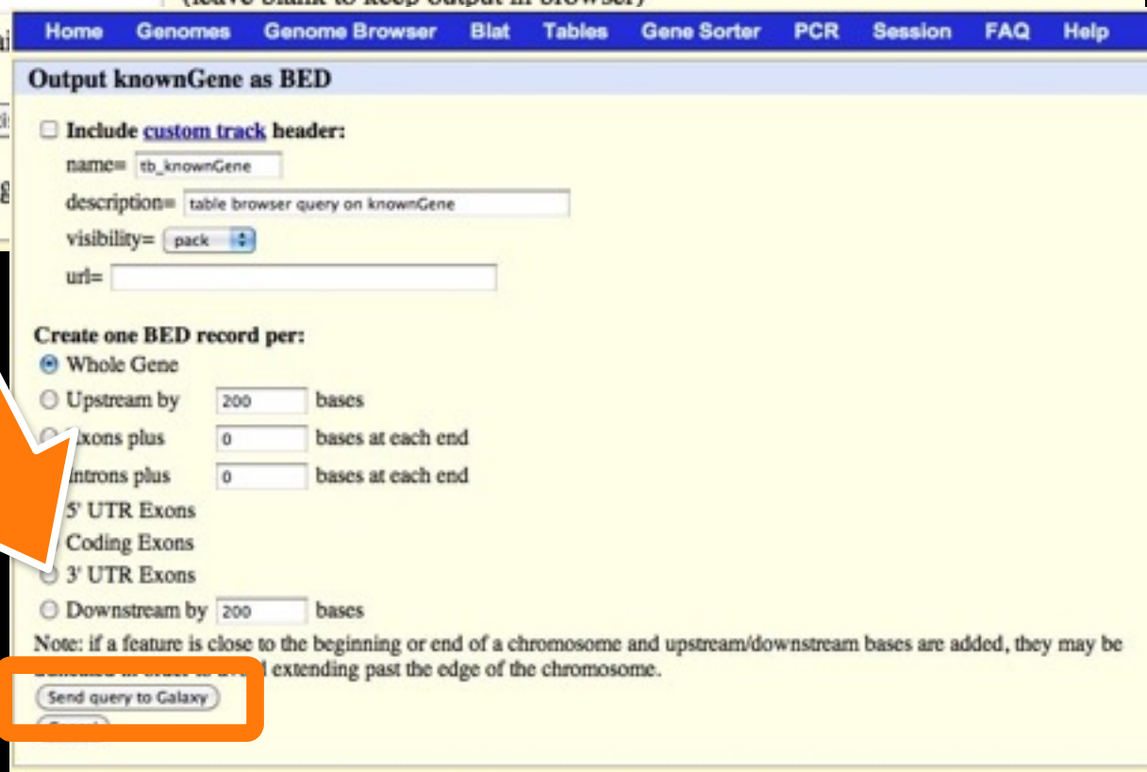
Getting the data

Accessibility









Accessibility

Concatenate datasets tail-to-head (Galaxy Tool Version 1.0.0) — Options

Concatenate Dataset

Single dataset | Multiple datasets | Dataset collection

84: Filter on data 33

Dataset

1: Dataset

Select

Single dataset | Multiple datas

36: Cuffdiff on data 13, data 9, and c
35: Cuffdiff on data 13, data 9, and c
34: Cuffdiff on data 13, data 9, and c
33: Cuffdiff on data 13, data 9, and c

This is a batch mode input field. A

Insert Dataset

Execute

WARNING: Be careful not to concatenate da
datasets being concatenated are in the sam

What it does

Concatenates datasets

Example

Concatenating Dataset:

```
chrX  151087187  151087355  A  0  -
chrX  151572400  151572481  B  0  +
```

catWrapper.xml

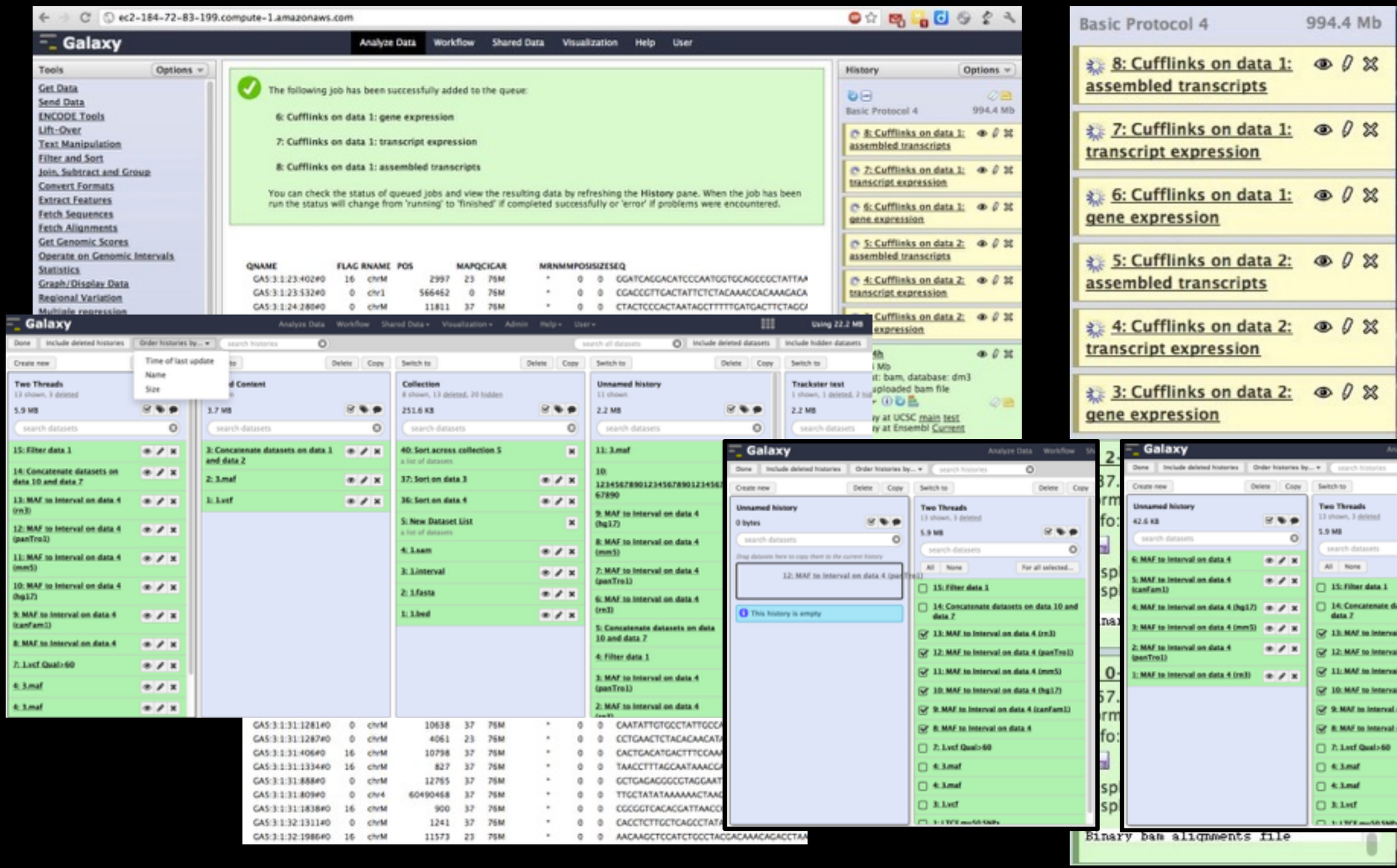
```
<tool id="cat1" name="Concatenate datasets">
    <description>tail-to-head</description>
    <command interpreter="python">
        catWrapper.py
        $out_file1
        $input1
        #for $q in $queries
            ${q.input2}
        #end for
    </command>
    <inputs>
        <param name="input1" type="data" label="Concatenate
        <repeat name="queries" title="Dataset">
            <param name="input2" type="data" label="Select"
        </repeat>
    </inputs>
    <outputs>
        <data name="out_file1" format="input" metadata_sour
    </outputs>
    <tests>
        <test>
            <param name="input1" value="1.bed"/>
            <param name="input2" value="2.bed"/>
            <output name="out_file1" file="cat_wrapper_out1.bed"/>
        </test>
        <!--TODO: if possible, enhance the underlying test code to handle this test
            the problem is multiple params with the same name "input2"
        <test>
            <param name="input1" value="1.bed"/>
            <param name="input2" value="2.bed"/>
            <param name="input2" value="3.bed"/>
            <output name="out_file1" file="cat_wrapper_out2.bed"/>
        </test>
        -->
    </tests>
    <help>

.. class:: warningmark

**WARNING:** Be careful not to concatenate datasets of different kinds (e.g., sequences with
intervals). This tool does not check if the datasets being concatenated are in the same format.

-----

**What it does**

Concatenates datasets

-----

**Example**

Concatenating Dataset::
```

Adding a tool?

-Automatically generated web UI from a tool wrapper (any tool can be integrated)
-Integrated with other tools

Data analysis history

Accessibility



Reproducibility in Genomics

18 *Nat. Genetics* experiments in microarray gene expression

<50% of reproducible

Problems

- missing data (38%)
- missing software, hardware details (50%)
- missing methods, processing details (66%)

Ioannidis, J.P.A. et al. "Repeatability of published microarray gene expression analyses." Nat Genet 41, 149-155 (2009)

14 re-sequencing experiments in *Nat. Genetics*, *Nature*, *Science*

0% reproducible?

Problems

- missing primary data (50%)
- tools unavailable (50%)
- missing parameter setting, tool versions (100%)

"Devil in the details," Nature, vol. 470, 305-306 (2011).

Metadata = Reproducibility

Reproducibility

Automatic metadata

95: Filter on data 94

202 lines

format: **tabular**, database: **hg19**

Filtering with c14=='yes',
kept 44.40% of 455 valid lines (455 total lines).

View details

3	4
test_id	gene_id	gene	locus
AAAS	AAAS	AAAS	chr12:53701239-
ABCC9	ABCC9	ABCC9	chr12:21950323-
ABCD2	ABCD2	ABCD2	chr12:39945021-
ALKBH2	ALKBH2	ALKBH2	chr12:10952599
ANAPC5	ANAPC5	ANAPC5	chr12:12174604

Tool: Filter

Name:	Filter on data 94
Created:	Sat Jan 17 23:34:56 2015 (UTC)
Filesize:	22.0 KB
Dbkey:	hg19
Format:	tabular
Galaxy Tool Version:	1.1.0
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	fcc622f6b9afc8d2
Full Path:	/mnt/galaxy/files/001/dataset_1107.dat
Job Command-Line:	python /mnt/galaxy/galaxy-app/tools/stats/filtering.py /mnt/galaxy/files/001/dataset_1107.dat "c14==__sq__ye 1
Job Runtime (Wall Clock)	2 seconds
Cores Allocated	1
Job Start Time	2015-01-17 23:35:04
Job End Time	2015-01-17 23:35:06

Input Parameter	Value
Filter	94: Filt
With following condition	c14==
Number of header lines to skip	1

Reproducibility

User metadata

History Options

Variant Analysis for Sample E18

Tags:

snp × pileup × bowtie ×

demo × sample:e18 ×

Annotation / Notes:

Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.

10: Variants from sample E18

26,742 regions, format: interval, database: mm9

Info:

Tags:

pileup × sample:e18 ×

snps ×

Annotation:

Find variants with coverage >= 30 and quality score >= 20.

| display at UCSC main | view in GeneTrack | display at Ensembl Current

1.Chrom	2.Start	3.End	4	5	6	7
chr10	6882036	6882037	A	A	107	
chr10	14243075	14243076	G	G	96	
chr10	14243079	14243080	C	C	106	
chr10	14465082	14465083	T	K	173	
chr10	14465083	14465084	G	K	144	
chr10	14465084	14465085	T	T	117	

Data provenance

Reproducibility

9: Map with BWA on data 1

1,447,842 lines

format: **sam**, *database:* **hg19**

BWA run on single-end data

Run this job again

```
2.FLAG 3.RNAME 4.
GA5:3:1:23:402#0        16      chrM
        XA:Z:chr17,-22023750,76M,2;
GA5:3:1:23:532#0  0       chr1      566
:Z:chrM,+5913,76M,0;
GA5:3:1:24:280#0        0       chrM
GA5:3:1:24:1454#0       16      chrM
```

This job was initially run with tool id "bwa_wrapper", version "1.0.3", w
which is a derivation of the original tool.

Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use a built-

Use a built-in index

Select a reference genome:

Arabidopsis lyrata: Araly1

An invalid option was selected, please verify

Is this library mate-paired?:

Single-end

FASTQ file:

1: p1-c-b-1.fastq

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina

BWA settings to use:

Commonly Used

For most mapping needs use Commonly Used settings. If you want full c

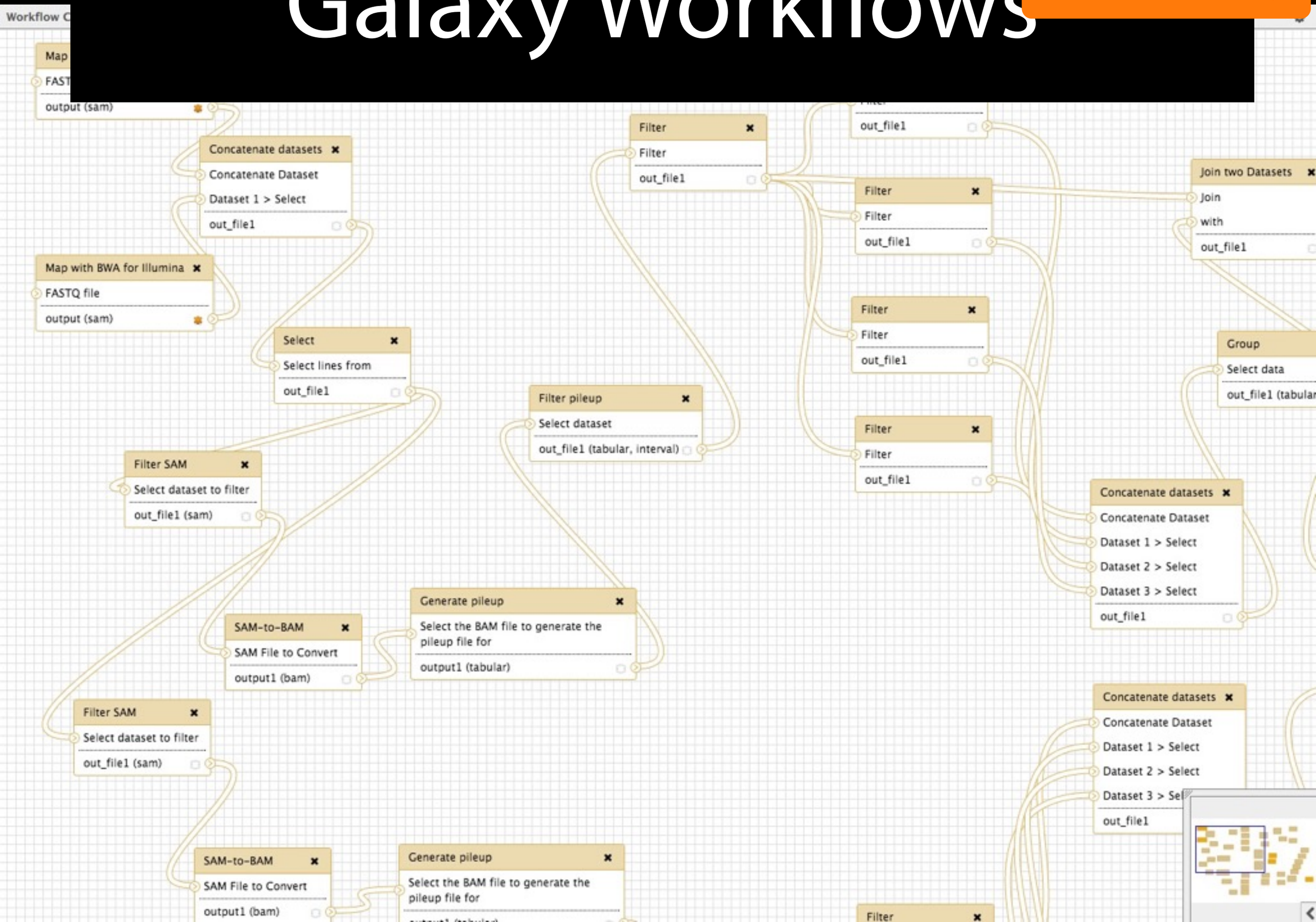
Suppress the header in the output SAM file:

BWA produces SAM with several lines of header information

Execute

Galaxy Workflows

Reproducibility



Transparency

Sharing and Publishing

Share or Publish History 'Mother child - sample'

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can access it. You can:

Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and import the history.

Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Share History with Individual Users

You have not shared this history with any users.

Share with a user

Transparency

Published Histories

search name, annotation, owner, and tag

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
Infravec		dan-lawson	★★★★★		Oct 01, 2014
SM_1186088	Datasets correspond to our paper published in Science by Peleg et al. entitled : Altered histone acetylation is associated with age-dependent memory...	publicdata	★★★★★		Apr 19, 2010
MOL470_Pset3_All		jbgreisman	★★★★★		Nov 18, 2012
SNP Calling		jallen	★★★★★		Feb 06, 2014
RNA-seq shared data		rna-seq-helin-group	★★★★★	illumina rnaseq	Jul 17, 2013
SRR534285		jallen	★★★★★		Feb 04, 2014
ChIP-seq shared data		chip-seq-helin-group	★★★★★	chip illumina	Jun 03, 2013
metagenomic analysis		aun1	★★★★★	ngs megablast galaxy 454 metagenomics	Mar 19, 2010
Galaxy vs MEGAN	Comparison of Galaxy vs. MEGAN pipeline.	aun1	★★★★★	megan galaxy metagenomics	Mar 19, 2010
Variant Analysis for Sample E18	Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.		★★★★☆	pileup bowtie demo snp sample	Sep 24, 2010
RNAseq_DGE_BASIC_Prep		eric-the-red	★★★☆☆		Aug 26, 2013

Transparency

Import history

Galaxy History ' week10 RNAseq 241213'

Dataset	Annotation
1: normal_1.fastq	
2: normal_2.fastq	
3: cancerous_1.fastq	
4: cancerous_2.fastq	
5: Tophat for Illumina on data 2 and data 1: insertions	
6: Tophat for Illumina on data 2 and data 1: deletions	
7: Tophat for Illumina on data 2 and data 1: splice junctions	
8: Tophat for Illumina on data 2 and data 1: accepted_hits	
9: Tophat for Illumina on data 4 and data 3: insertions	
10: Tophat for Illumina on data 4 and data 3: deletions	
11: Tophat for Illumina on data 4 and data 3: splice junctions	
12: Tophat for Illumina on data 4 and data 3: accepted_hits	
13: Cufflinks on data 8: gene expression	
14: Cufflinks on data 8: transcript expression	
15: Cufflinks on data 8: assembled transcripts	
17: Cufflinks on data 12: gene expression	
18: Cufflinks on data 12: transcript expression	

Inspect or import and use

Author

arik

Related Histories

All published histories
Published histories by arik

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community: none

Yours:

Galaxy Pages

Transparency

Galaxy 101: The first thing you should try

In this very simple example we will introduce you to bare basics of Galaxy:

Getting data from UCSC
Performing simple data manipulation
Understanding Galaxy's History system
Creating and editing workflows
Applying workflows to your data
You can watch a step-by-step explanation of this entire tutorial here.

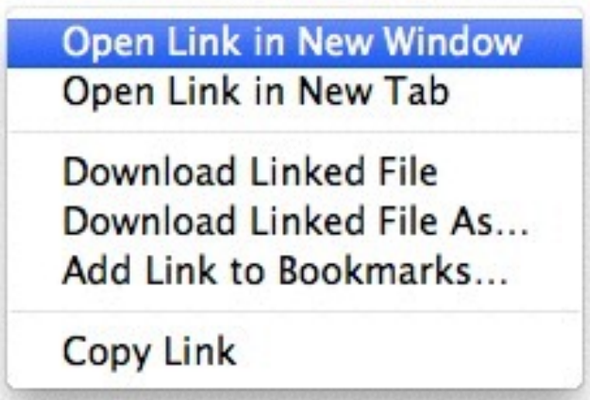
What are we trying to do?

Suppose you get the following question: "Mom (or Dad) ... Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?". You think to yourself "Wow! This is a simple question ... I know exactly where the data is (at UCSC) but how do I actually compute this?" The truth is, there is really no straightforward way of answering this question in a time frame comparable to the attention span of a 7-year-old. Well ... actually there is and it is called Galaxy. So let's try it...

0. Organizing your windows and setting up Galaxy account

0.0. Getting your display sorted out

To get the most of this tutorial open two browser windows. One you already have (it is this page). To open the other, right click this link and choose "Open in a New Window" (or something similar depending on your operating system and browser):



Author

aun1

Related Pages

All published pages
Published pages by aun1

Rating

Community
(46 ratings, 4.9 average)

Yours

Tags

Community:

tutorial snps exons

Yours:

Visualization within Galaxy

Galaxy

- Tool integration framework
- Heavy focus on usability
- Sharing, publication framework

Genome Browser

- Physical depiction of data
- Visually identify correlations
- Find interesting regions, features

Trackster

Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Google

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 1,290 - 4,168,475

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,000

UCSC Main on Human: knownGene (chr19) — Auto (Squish)

UCSC Main on Human: all_est (chr19) — Auto (coverage histogram)

UCSC Main on Human: phyloP46wayPrimates (chr19) — Histogram

h1-hESC Tophat Mapped Reads — Auto (coverage histogram)

h1-hESC Cufflinks assembled transcripts — Auto (Squish)

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,000

Display a menu

Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Google

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 663,032 - 663,110

30 663,040 663,050 663,060 663,070 663,080 663,090 663,100

g g c c c g g g c c T C A C C G G C A G G C G C G G G A C G A T C T C C A C G G A G C A G C A G T G G C A G A A G T A C C G T C C G G G A T G C G G C G A C

UCSC Main on Human: knownGene (chr19) Auto (Pack)

UCSC Main on Human: all_est (chr19) Dense

UCSC Main on Human: phyloP46wayPrimates (chr19) Histogram

1

-1

h1-hESC Tophat Mapped Reads Auto (Pack)

HWI-EAS229_1:2:7:78:211

HWI-EAS229_1:2:39:1023:1868

HWI-EAS229_1:2:83:31:1502

HWI-EAS229_1:2:36:329:1847

h1-hESC Cufflinks assembled transcripts Auto (Pack)

30 663,040 663,050 663,060 663,070 663,080 663,090 663,100

Canceled opening the page

HTS Analysis

- Set parameters
- Run tools / workflow
- Wait...
- Visualize output

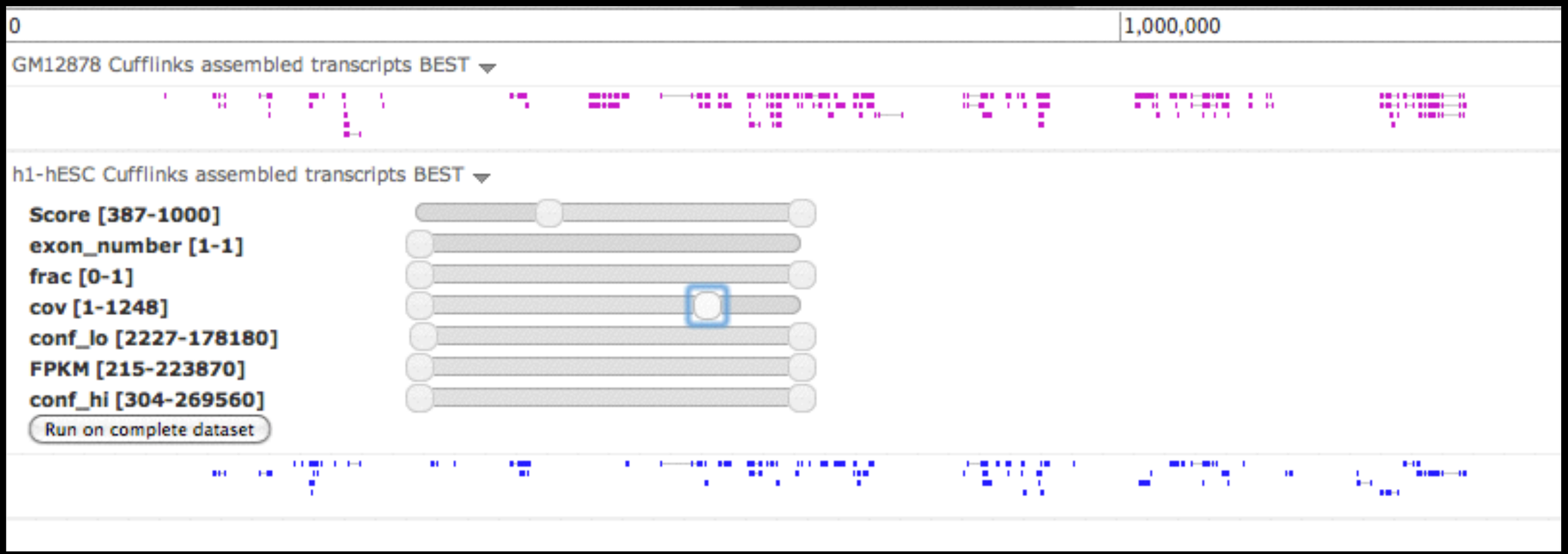
Experimentation

- Understand, debug, and “tune” analyses
- Easy with Galaxy
  - Parameter values saved
  - Can rerun tools and analyses
  - All outputs stored

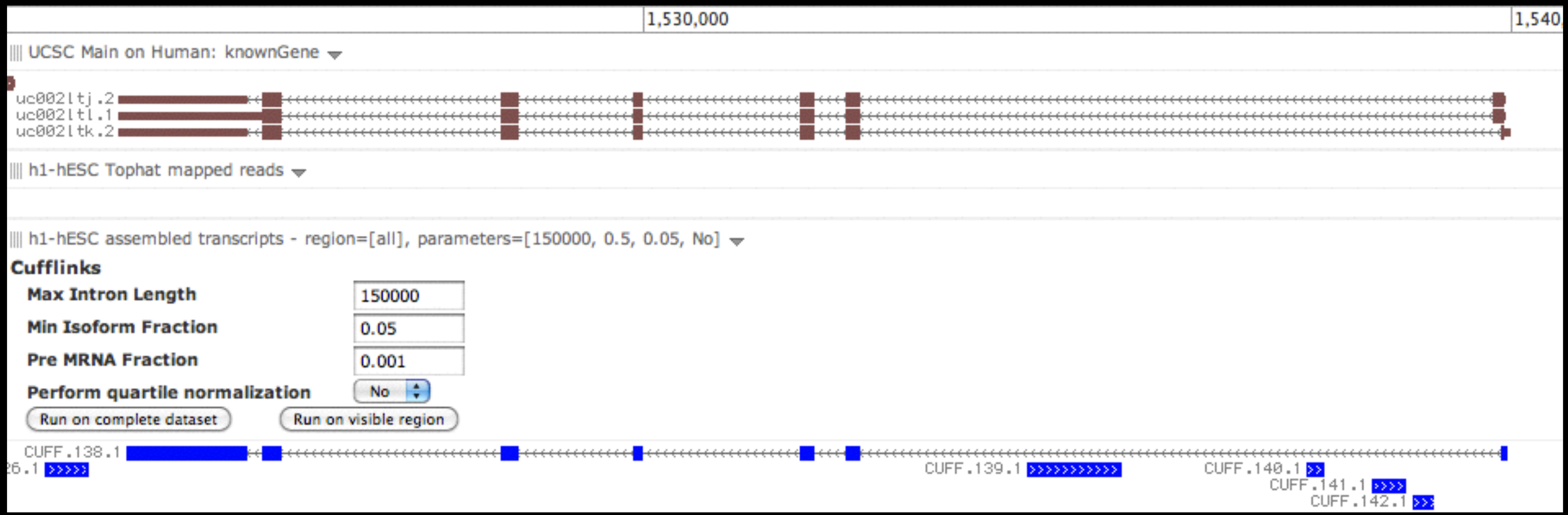
Visualization Then and Now

- Then
  - Visualization endpoint of analysis
  - Difficult to change analyses based on visualization
- Now
  - Integrate analysis and visualization
  - Make it simple to move between analysis and visualization

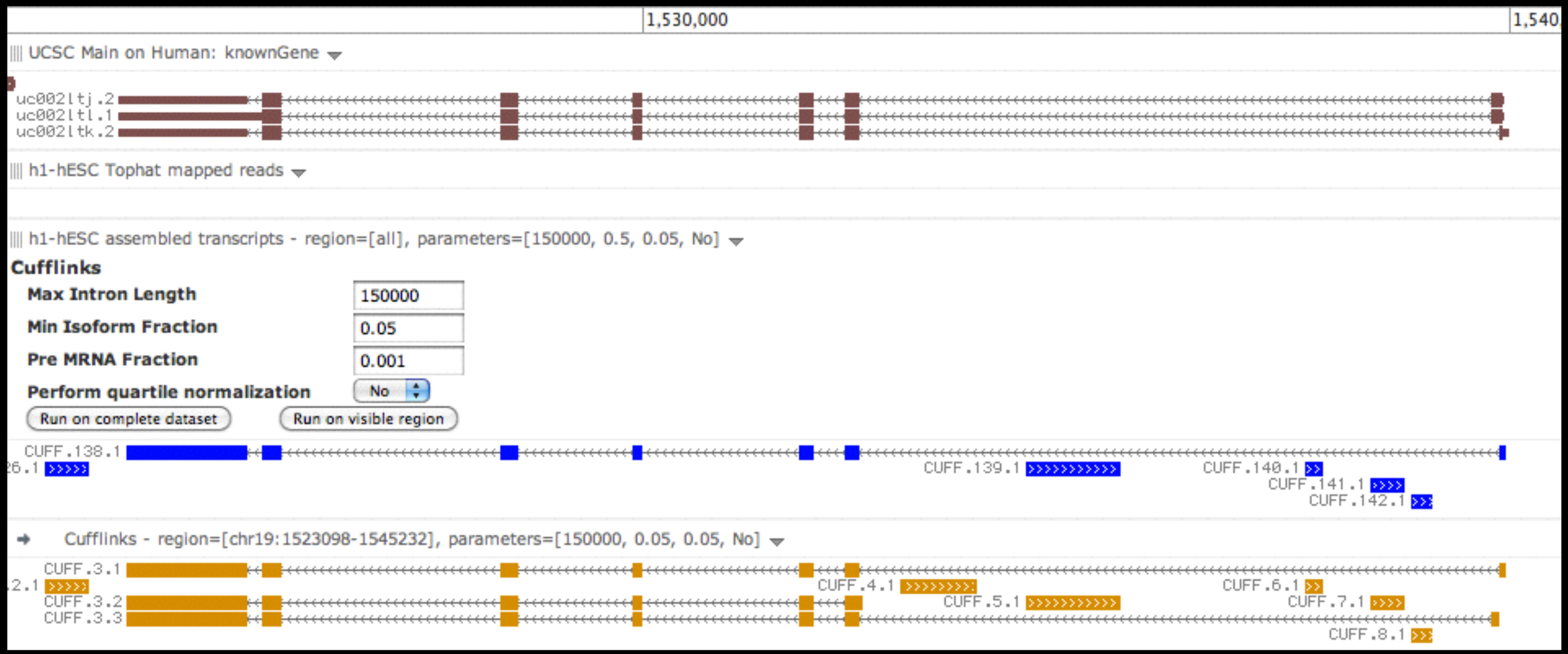
Dynamic Filtering



Integrating Tools and Visualization



Integrating Tools and Visualization



Three ways to use Galaxy

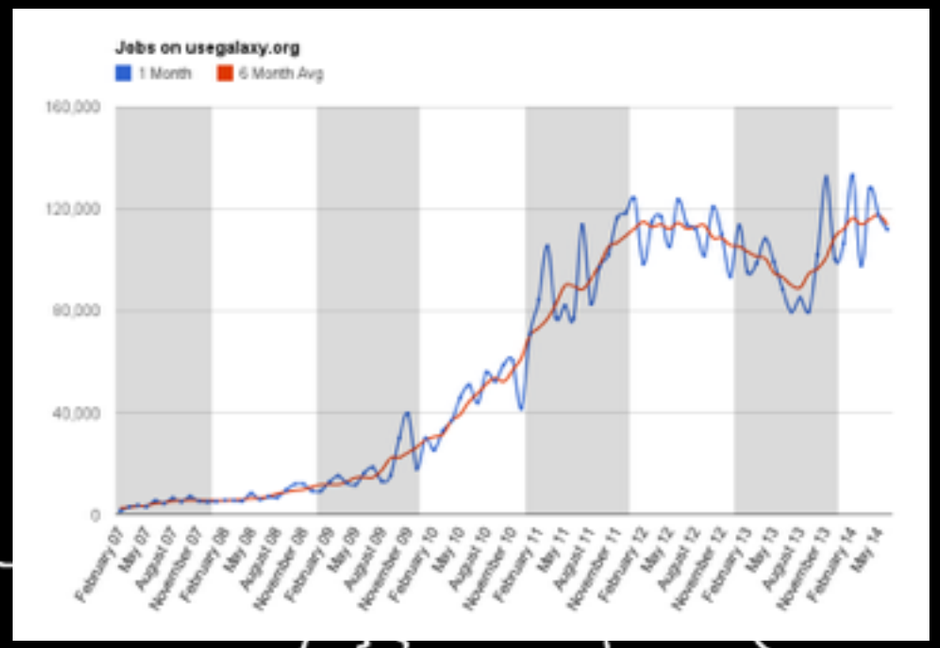
Public website

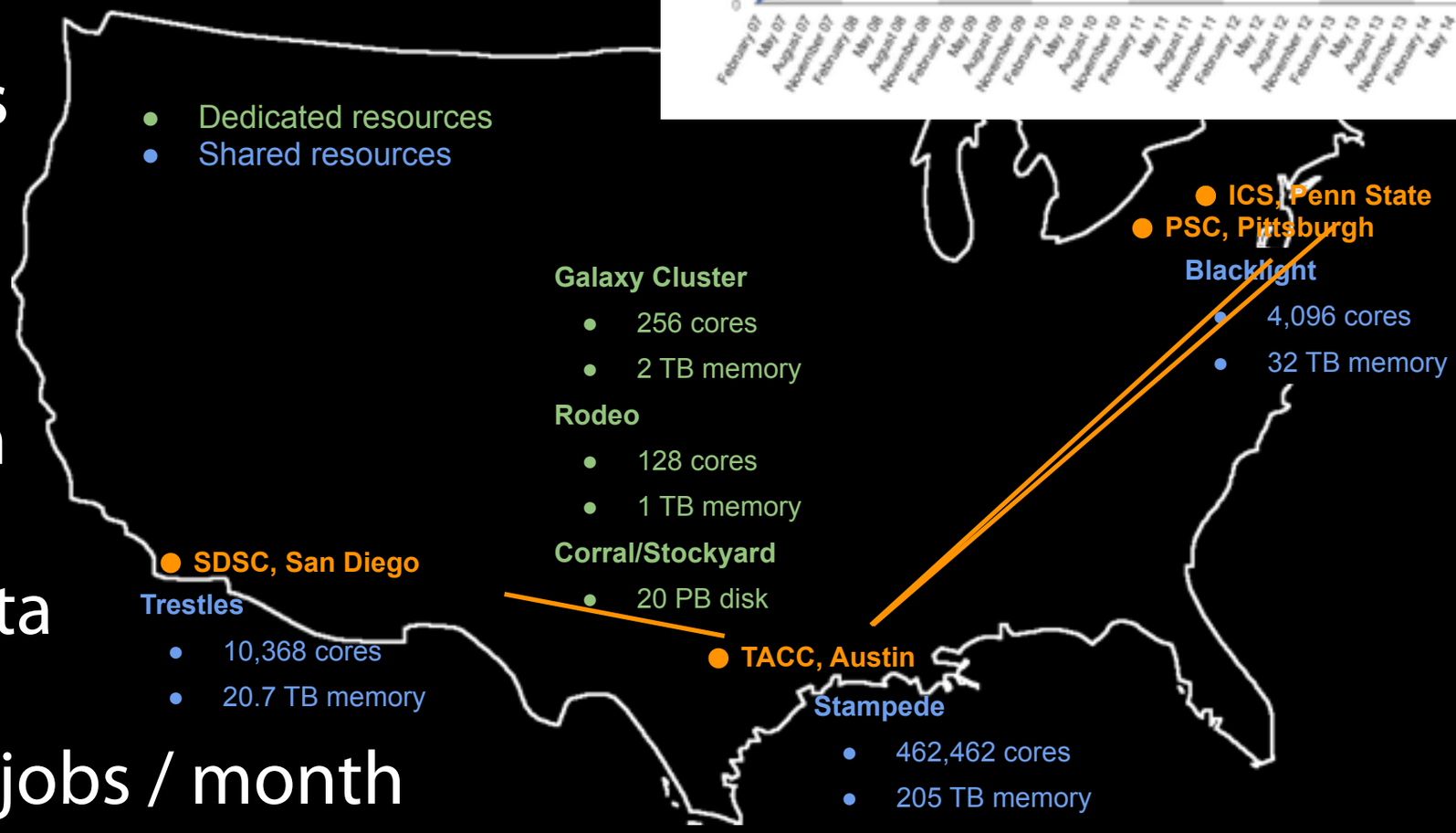
Download and Run Locally

Run on the Cloud

http://usegalaxy.org (a.k.a. Main)

- **Public web site**
- **Anybody can use it**
- **Hundreds of tools**
- Persistent
- +500 users/month
- ~200TB of user data
- ~140,000 analysis jobs / month





http://bit.ly/gxystats

Public Galaxy Servers

https://wiki.galaxyproject.org/PublicGalaxyServers

Interested in:

60+ Public Servers

ChIP-chip and ChIP-seq?

✓ Cistrome

Statistical Analysis?

✓ Genomic Hyperbrowser

Sequence and tiling arrays?

✓ Oqtans

Text Mining?

✓ DBCLS Galaxy

Reasoning with ontologies?

✓ GO Galaxy

Internally symmetric protein structures?

✓ SymD

Local Galaxy Instances

http://getgalaxy.org

Galaxy is designed for local installation and customization

- Easily integrate new tools
- Easy to deploy and manage on nearly any (Unix) system

Got your cluster?

- Move tool execution to other systems
- Galaxy works with any DRMAA compliant cluster job scheduler (which is most of them).

- Galaxy is just another client to your scheduler.

Need a cluster? Galaxy CloudMan

- Start with a fully configured and populated (tools and data) Galaxy instance in the cloud.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center.
- We are using this today

Manage Your Cloud Cluster

CloudMan from Galaxy

Admin | Report bugs | Wiki | Screencast

CloudMan Console

s you to manage this instance cloud cluster and the services provided
econnected. Once the cluster has initialized, use the controls below to

odes ▼ | Remove nodes ▼ | Access Galaxy

quested: 5

Autoscaling is off.
Turn on?

Customizing Galaxy: Tool Shed

- Allow users to share tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies: biomedical app store
- Integration with Galaxy instances to automate tool installation and updates

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

2857 valid tools on Dec 16, 2014

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Repositories by Category

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	31
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	9
Combinatorial Selections	Tools for combinatorial selection	
Computational chemistry	Tools for use in computational chemistry	17
Convert Formats	Tools for converting data formats	34
Data Managers	Utilities for Managing Galaxy's built-in data cache	6
Data Source	Tools for retrieving data from external data sources	18
Fasta Manipulation	Tools for manipulating fasta data	63
Fastq Manipulation	Tools for manipulating fastq data	27
Genome-Wide Association Study	Utilities to support Genome-wide association studies	1
Genomic Interval Operations	Tools for operating on genomic intervals	33
Graphics	Tools producing images	28
Imaging	Utilities to support imaging	1
Metabolomics	Tools for use in the study of Metabolomics	4
Metagenomics	Tools enabling the study of metagenomes	41
Micro-array Analysis	Tools for performing micro-array analysis	7
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	75
Ontology Manipulation	Tools for manipulating ontologies	8

Install to Galaxy | Browse repository | Tool Shed Actions

Repository 'bowtie2'

Revision:
0:96d2e31a3938 *(2014-01-27)*

Dependencies of this repository

Repository dependencies – *installation of these additional repositories is required*

Repository **package_bowtie2_2_1_0** revision **017a00c265f1** owned by **devteam**

Repository **package_samtools_0_1_18** revision **171cd8bc208d** owned by **devteam**

Tool dependencies – *repository tools require handling of these dependencies*

Name	Version	Type
samtools	0.1.18	package
bowtie2	2.1.0	package

Contents of this repository

Valid tools – *click the name to preview the tool and use the pop-up menu to inspect all metadata*

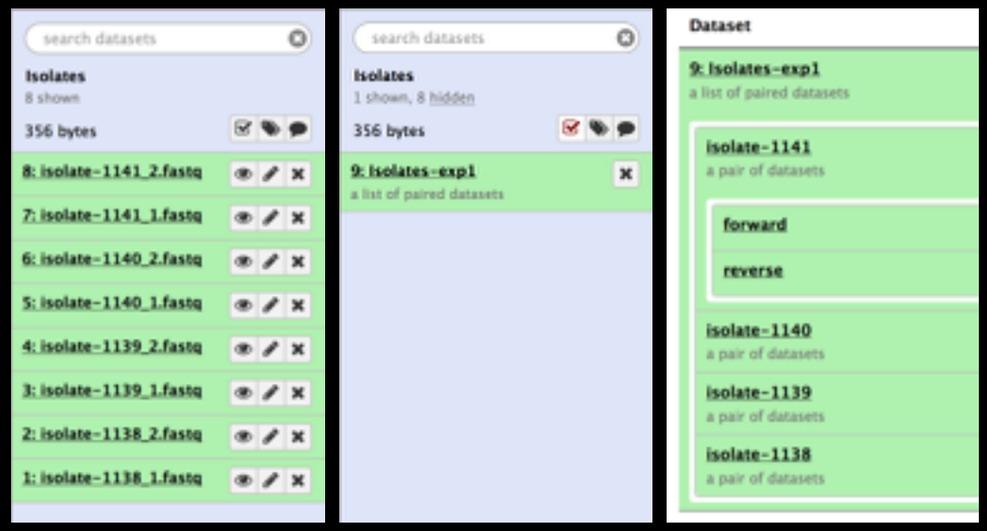
Automated tool test results

Test runs

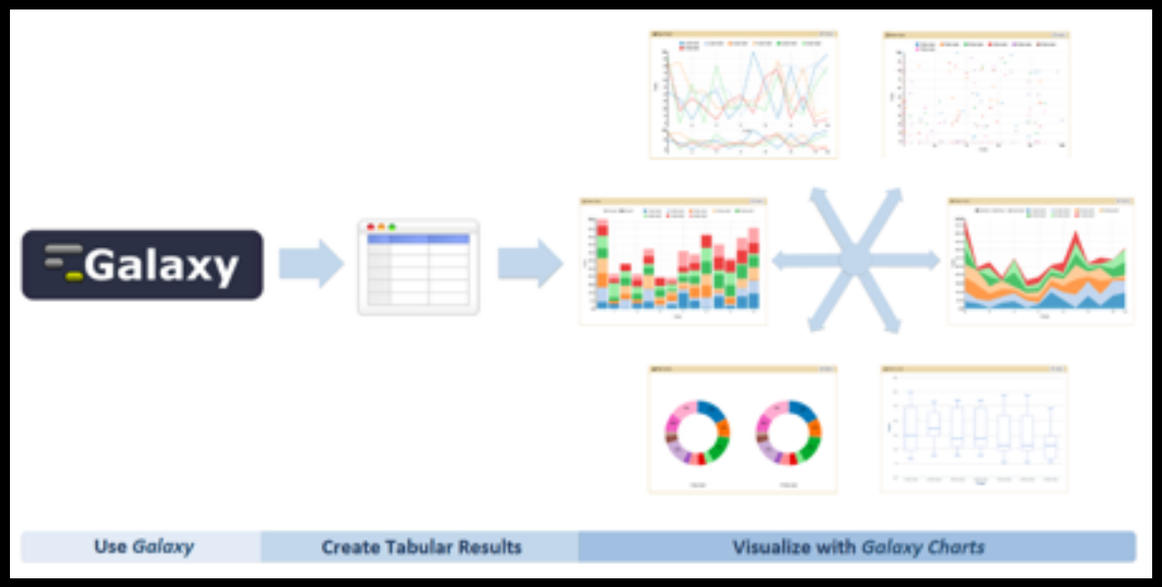
And much, much more…

Interactive environments

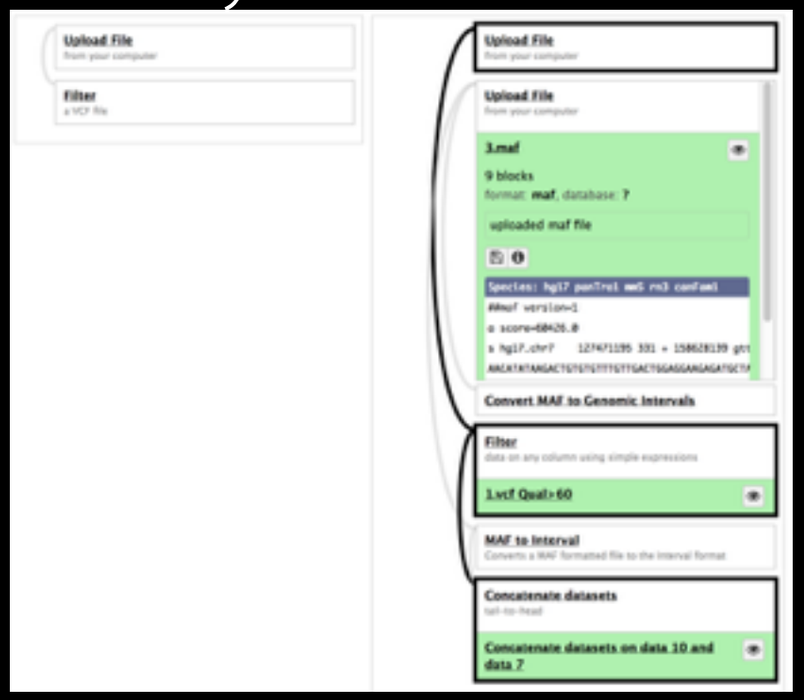
Dataset collections



Embedded charts



History structure



The Agenda

Galaxy project overview

RNA-Seq exercise

- RNA-Seq in brief
- QC
- Mapping

Galaxy Community

Differential expression

Not The Agenda

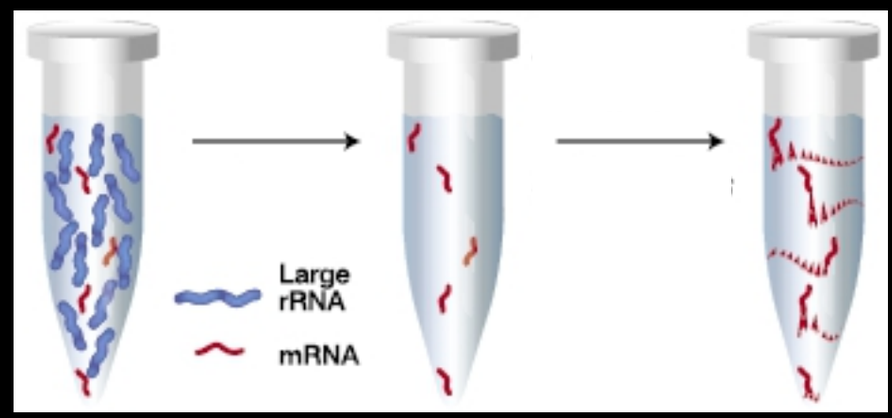
This workshop will *not* cover

- 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.

RNA

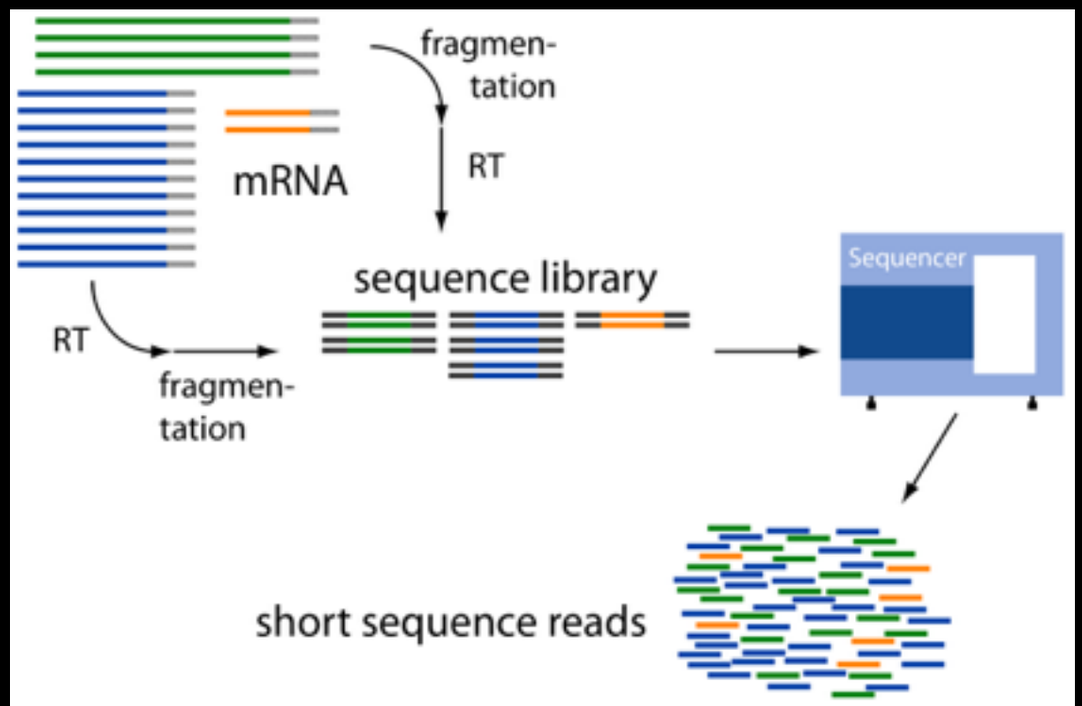
- RNA in cells consists of
  - 95% ribosomal rRNA and tRNA
  - other non-coding ncRNA
  - protein coding mRNA



- Sequence is transcribed from genome but
  - Introns spliced out
  - mRNA is polyadenylated (“A”s added to end)

RNA-Seq

- Deplete rRNA
- or select for polyadenylated RNA
- Fragmentation
- Reverse transcribe to cDNA
- Attach adaptor sequences
- Size selection
- Amplify by PCR
- High-throughput sequencing (eg Illumina)



Pepke *et al*, Nature Methods, 2009

RNA-Seq output

- Several million "reads" per sample

- Reads are RNA sequence starting from random locations within the original mRNA

- May read through into adaptor sequence

- Current typical length ~150 bases
  (really just needs to be long enough to locate in genome)

RNA-Seq applications

- Sequencing RNA transcripts, applications:
  - *de novo* transcriptome
  - **gene-wise differential expression**
  - novel splice variants
  - differential splicing
  - fusion genes
  - non-coding RNA
  - polyadenylation length
  - post-transcriptional modification
  - ...

- Differential expression analysis most common
  - Commonly called “DGE”

This workshop

- Analysis of RNA-seq data to detect differential gene expression
- How much are the transcript expression levels changing between conditions
- We'll do a basic **A vs B** analysis

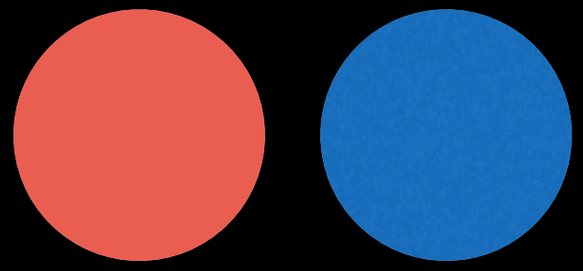
Example experiment

- 2 conditions of interest
  - eg. mutant versus wild-type, or
  - treated versus untreated, or, . . .
- Want to find all genes that are expressed differently between the 2 conditions

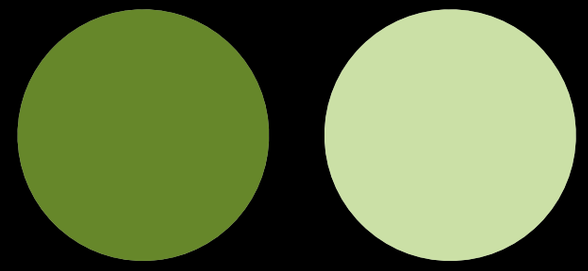
Replicas, replicas, replicas

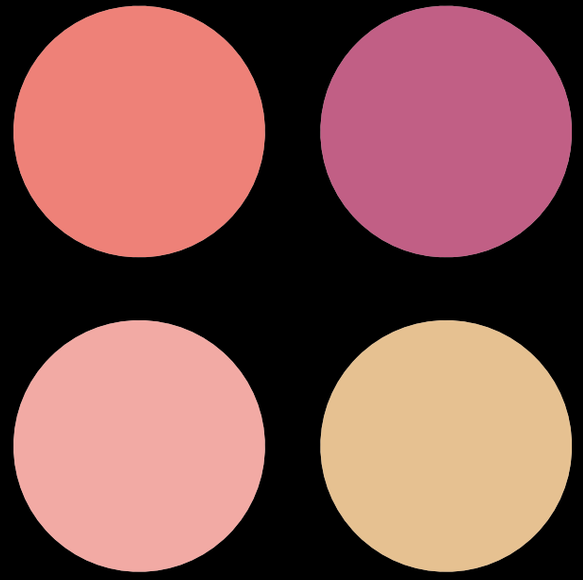
Why replicate?

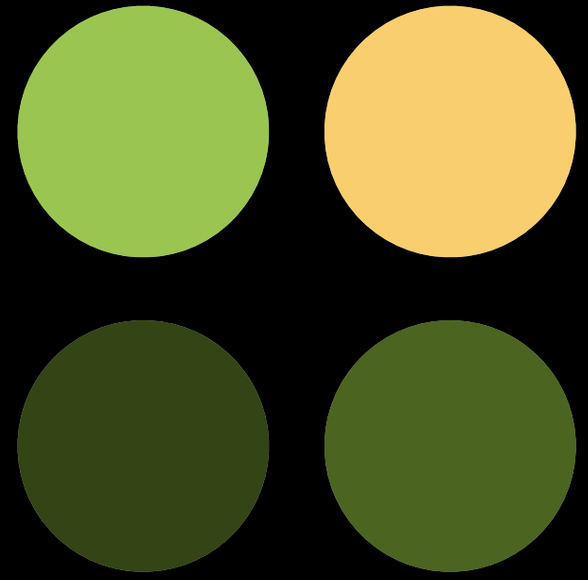
Mutant



Wild-type

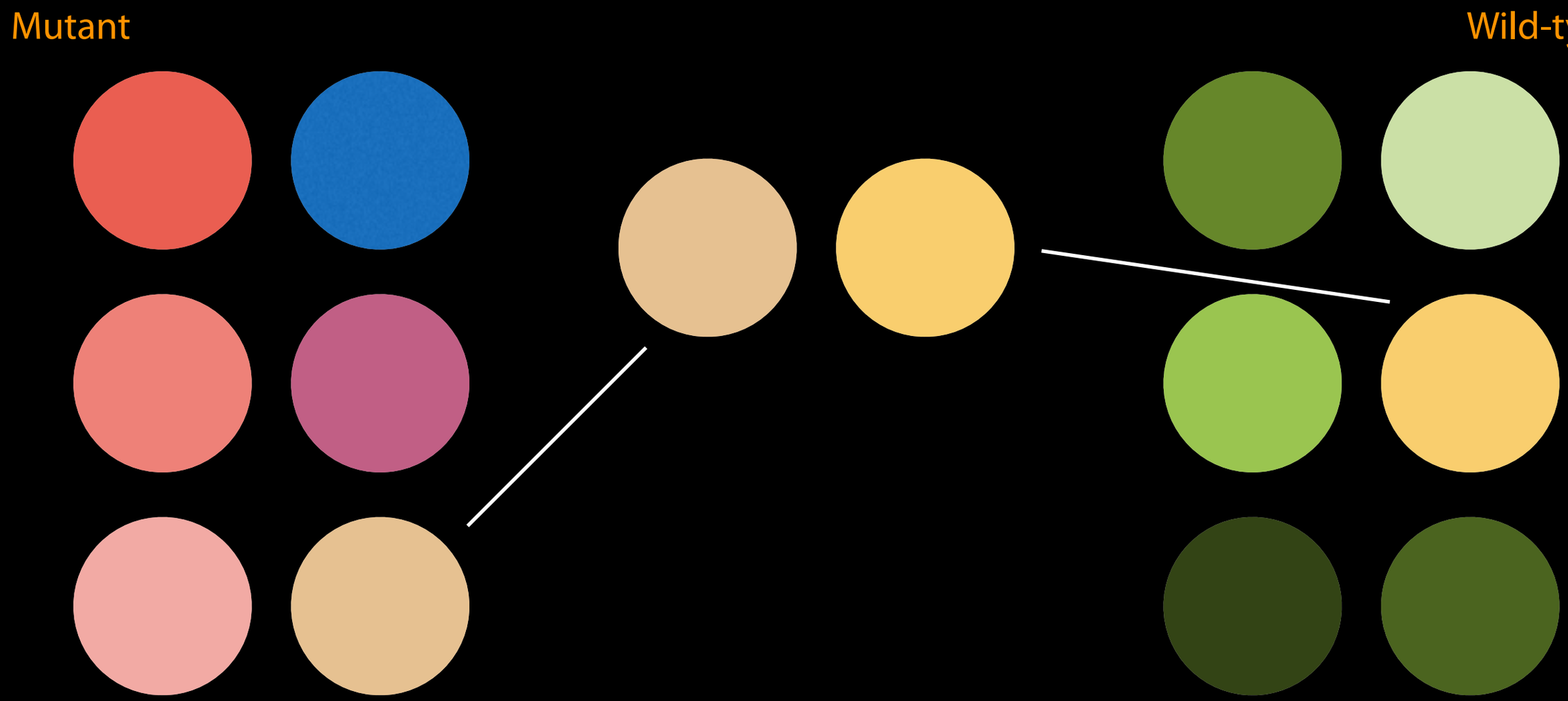






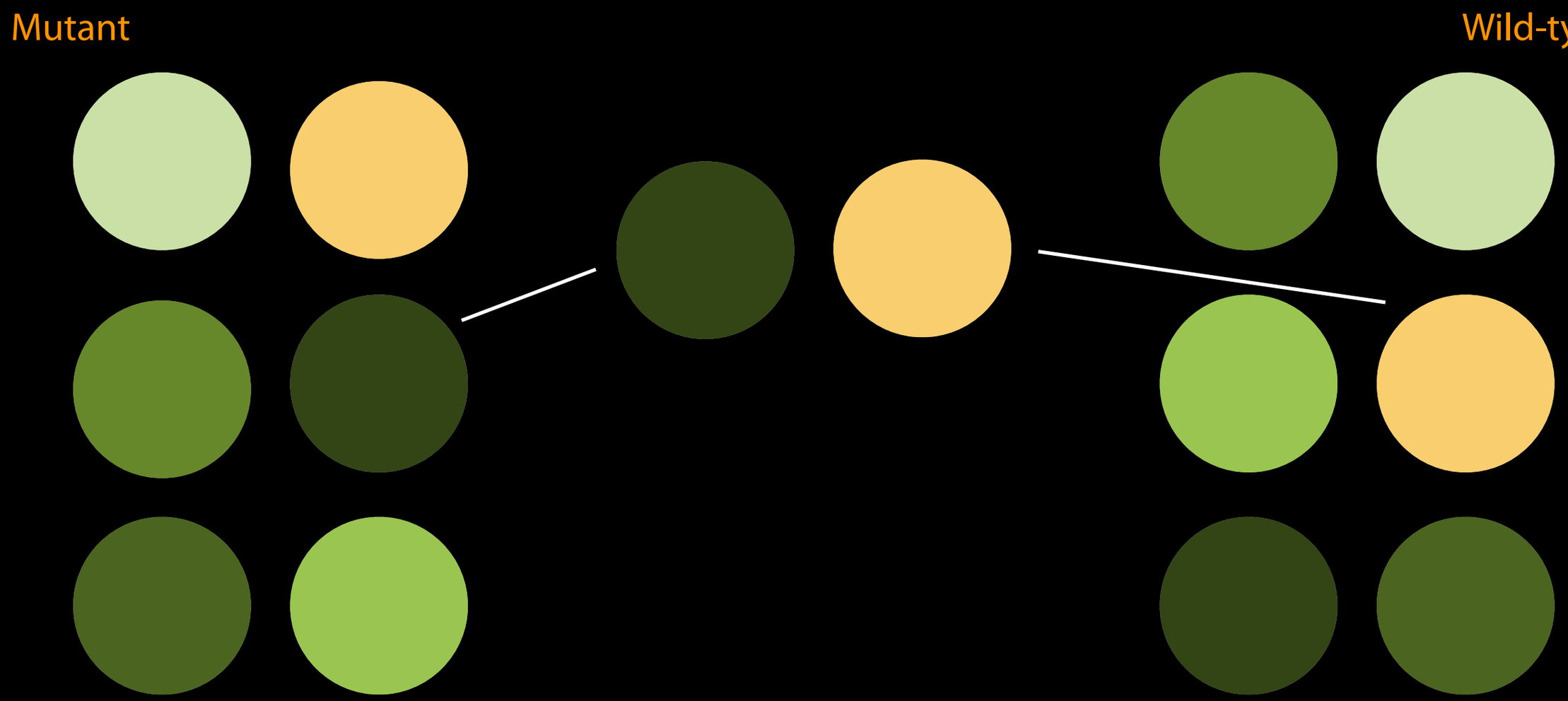
Biological heterogeneity

Unreplicated design



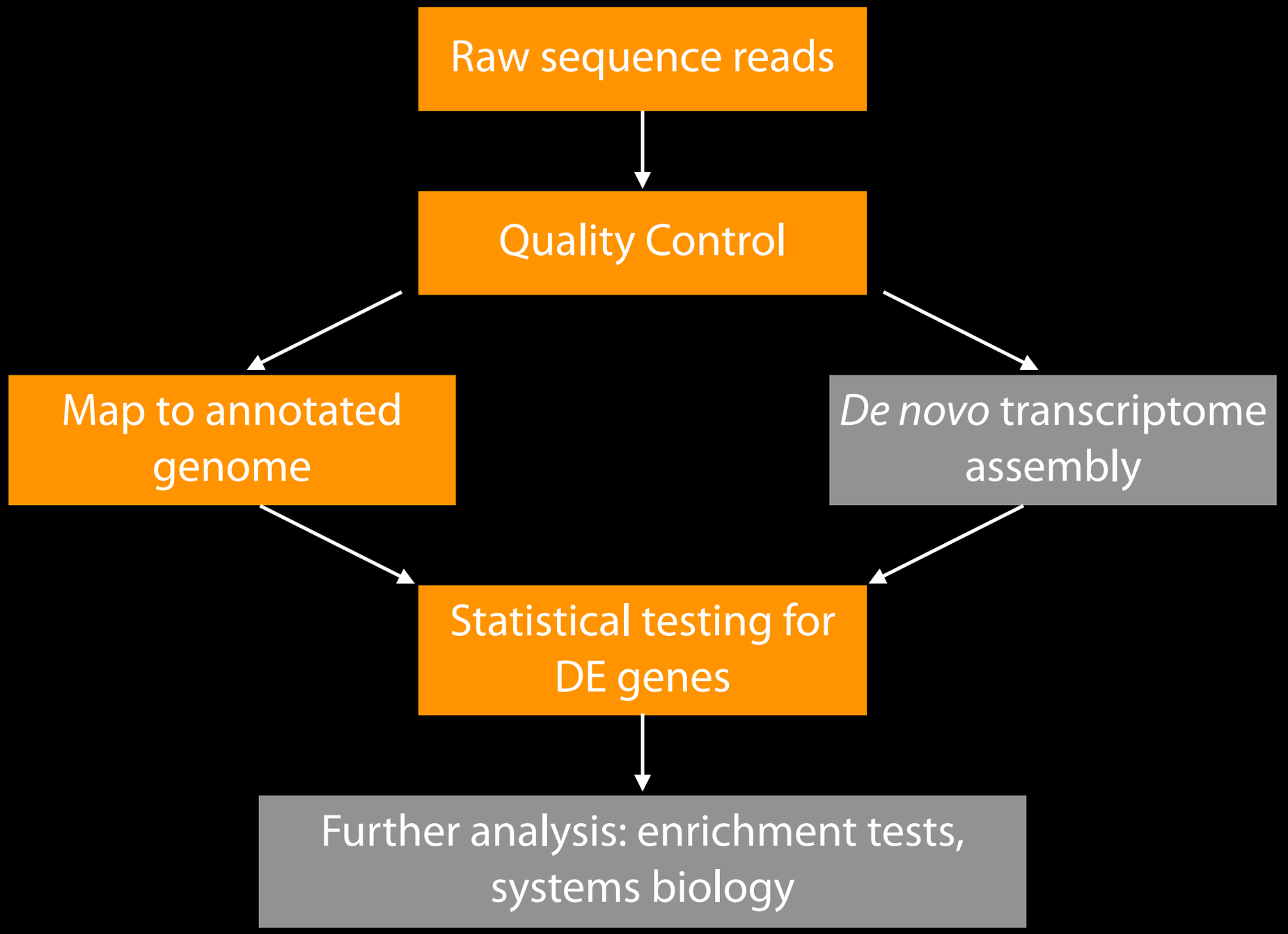
Here, groups differ, but single replicates from each group very similar

Unreplicated design



Here, groups are similar, but outlying observation from group on right makes it look like there's a big difference in unreplicated experiment

RNA-Seq DE analysis steps



NGS Data Quality Control

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

Quality Control is not sexy.

It is vital.

What is FASTQ?

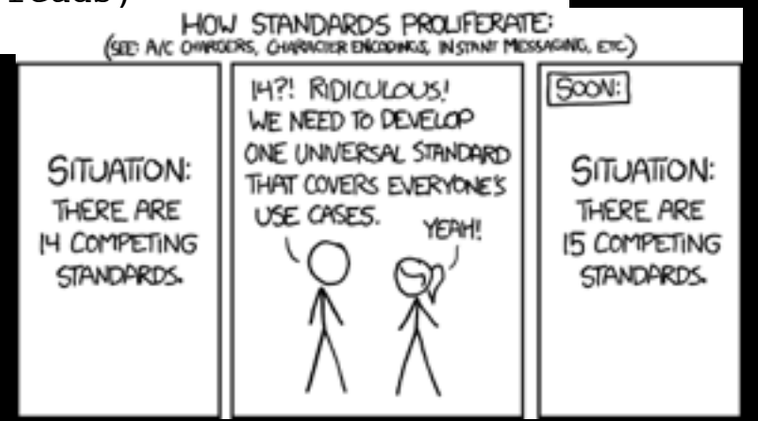
- Specifies sequence (FASTA) and quality scores (PHRED)
- Text format, 4 lines per entry

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

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

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  .               ..............XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./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



Our infrastructure for the day

http://cloud1.galaxyproject.org/
http://cloud2.galaxyproject.org/
http://cloud3.galaxyproject.org/
http://cloud4.galaxyproject.org/

NGS Data Quality Exercise

Create new history

(cog) → Create New

Get some data

Shared Data → Data Libraries

→ RNA-Seq Example Data*

→ Unfiltered Reads

→ Select MeOH_REP1_R1, MeOH_REP1_R2
and then Import to current history

UCDAVIS Genome Center Bioinformatics Core

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

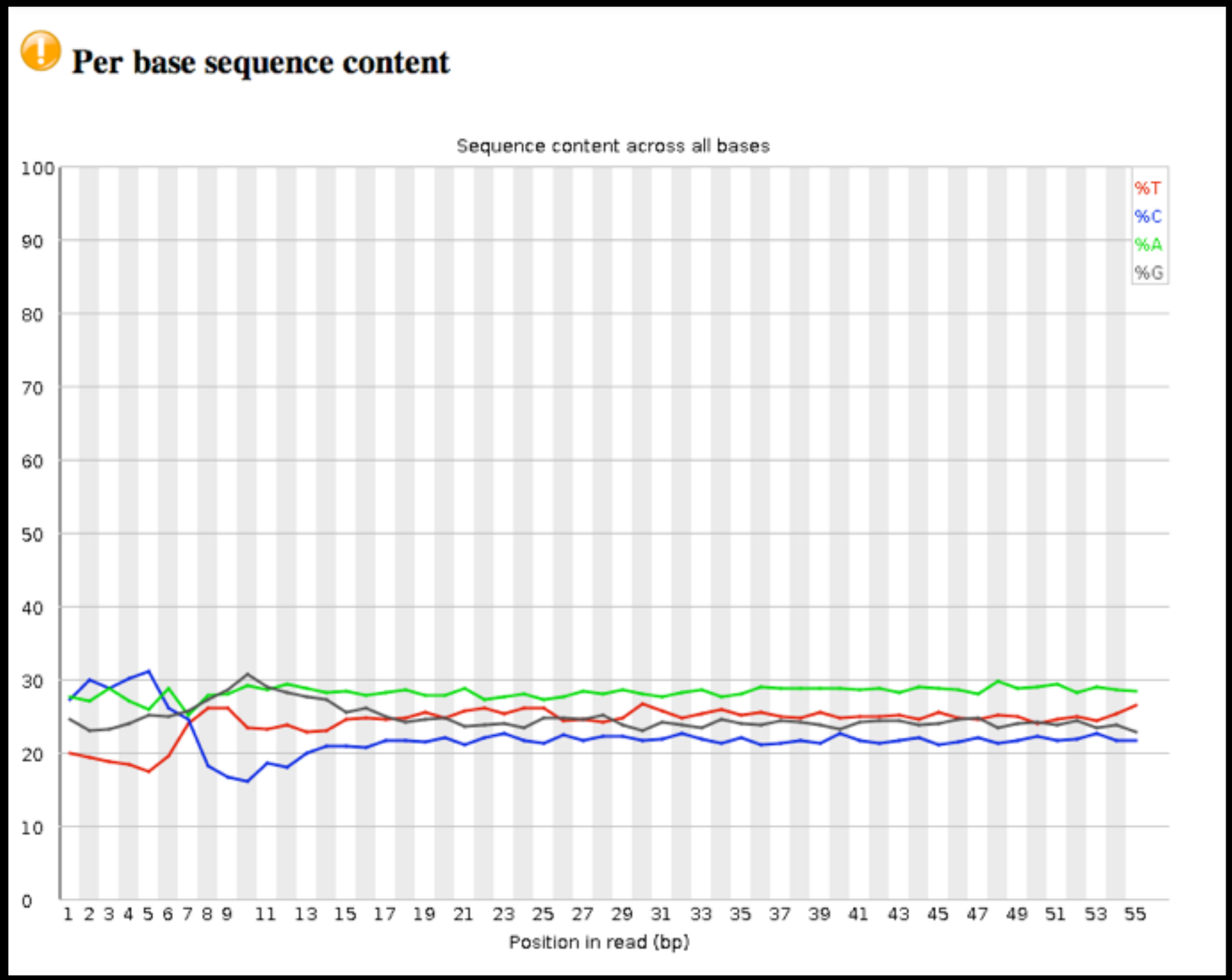
NGS Data Quality: Assessment tools

NGS QC and Manipulation → FastQC

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

http://bit.ly/FastQCBoxPlot

NGS Data Quality: Sequence bias at front of reads?

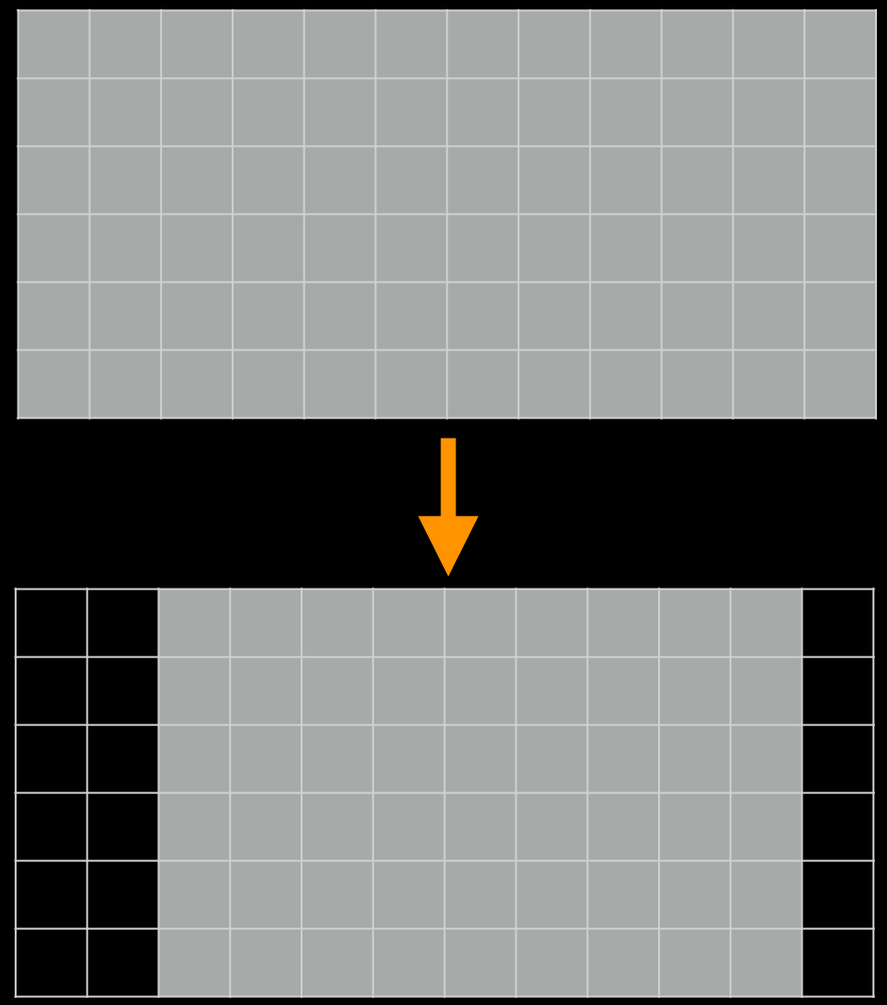


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

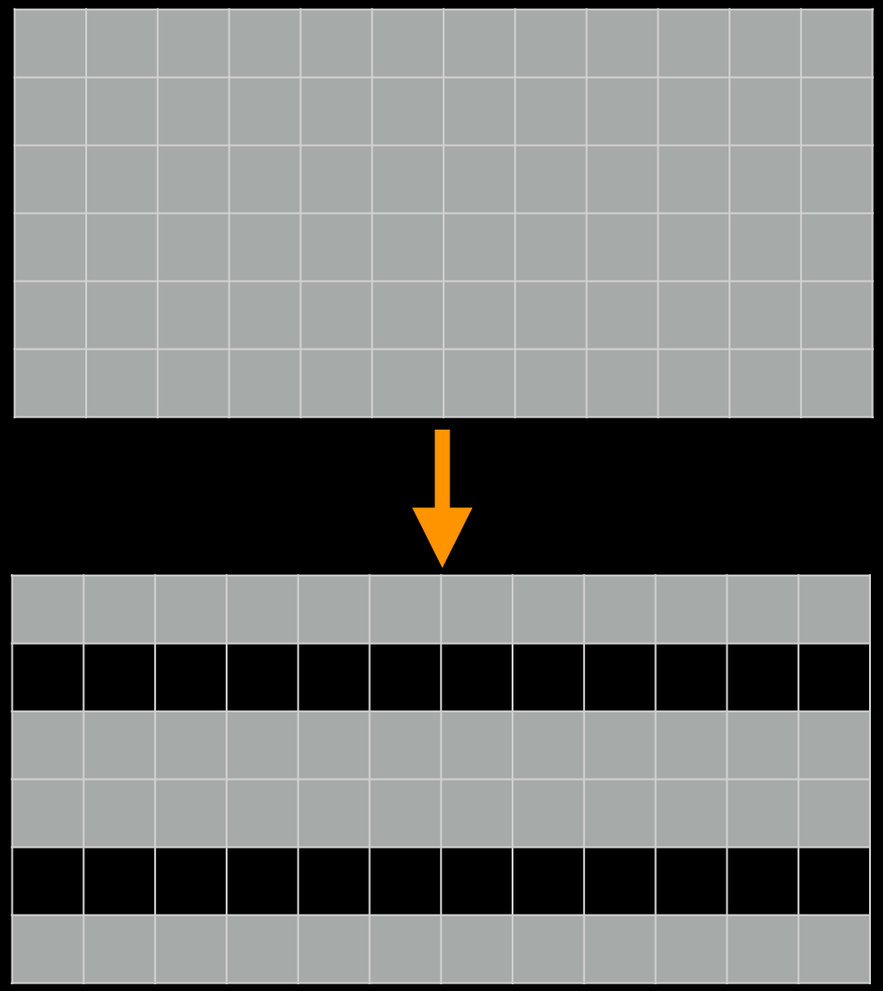
NGS Data Quality: Trim as we see fit

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



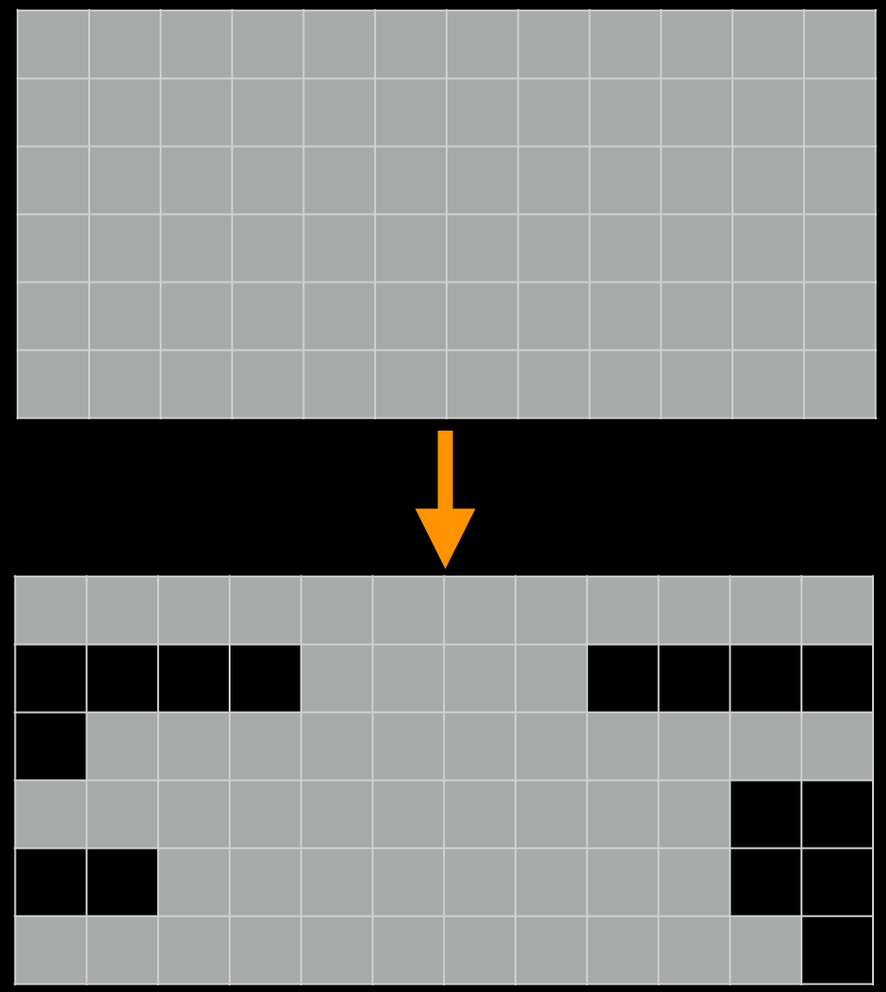
NGS Data Quality: Base Quality Trimming

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

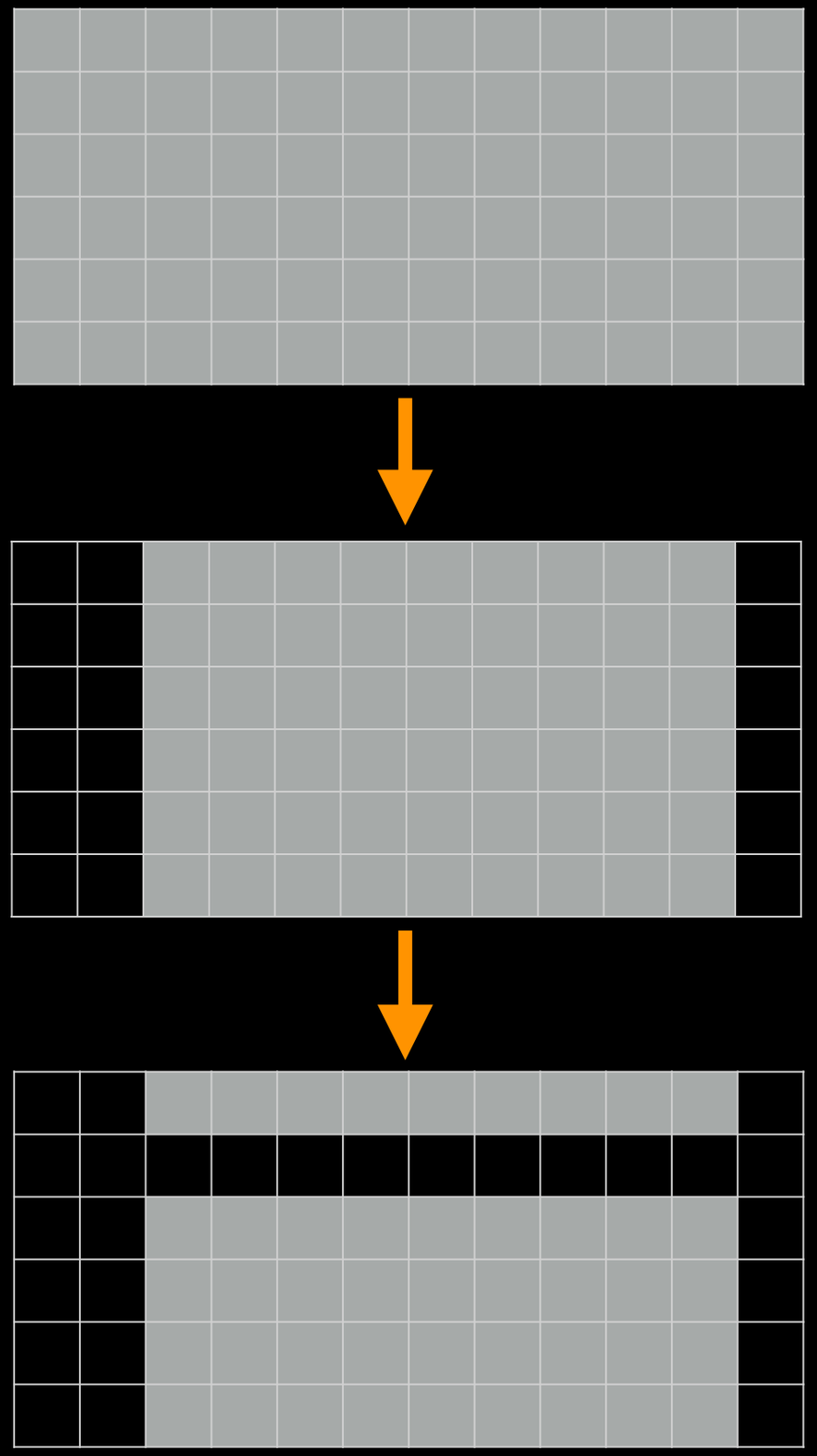


NGS Data Quality: Base Quality Trimming

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



Options are not mutually exclusive



Option 1
(by column)

+

Option 2
(by entire row)

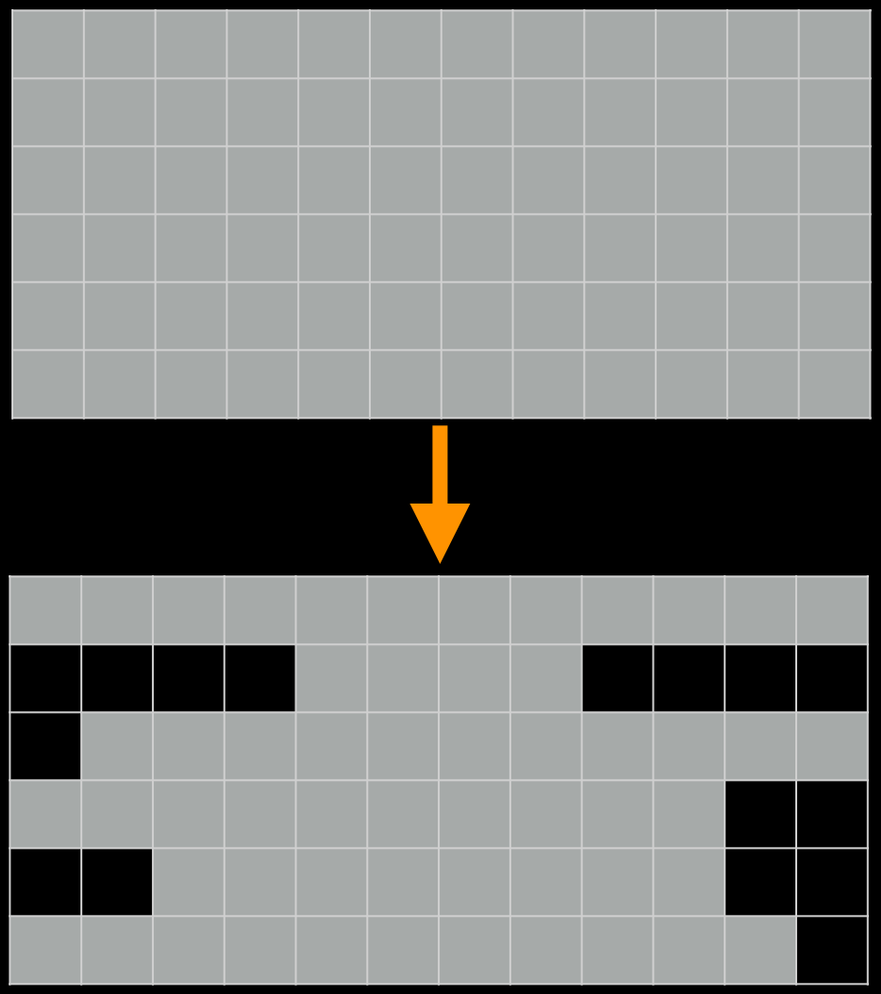
Trim? *As we see fit?*

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

Trim? *As we see fit?*

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

NGS Data Quality: Base Quality Trimming



I really want to use Option 3:

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

but ...

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

Tophat Manual

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

Jen Jackson

Galaxy User Support Person Extraordinaire

"Dang."

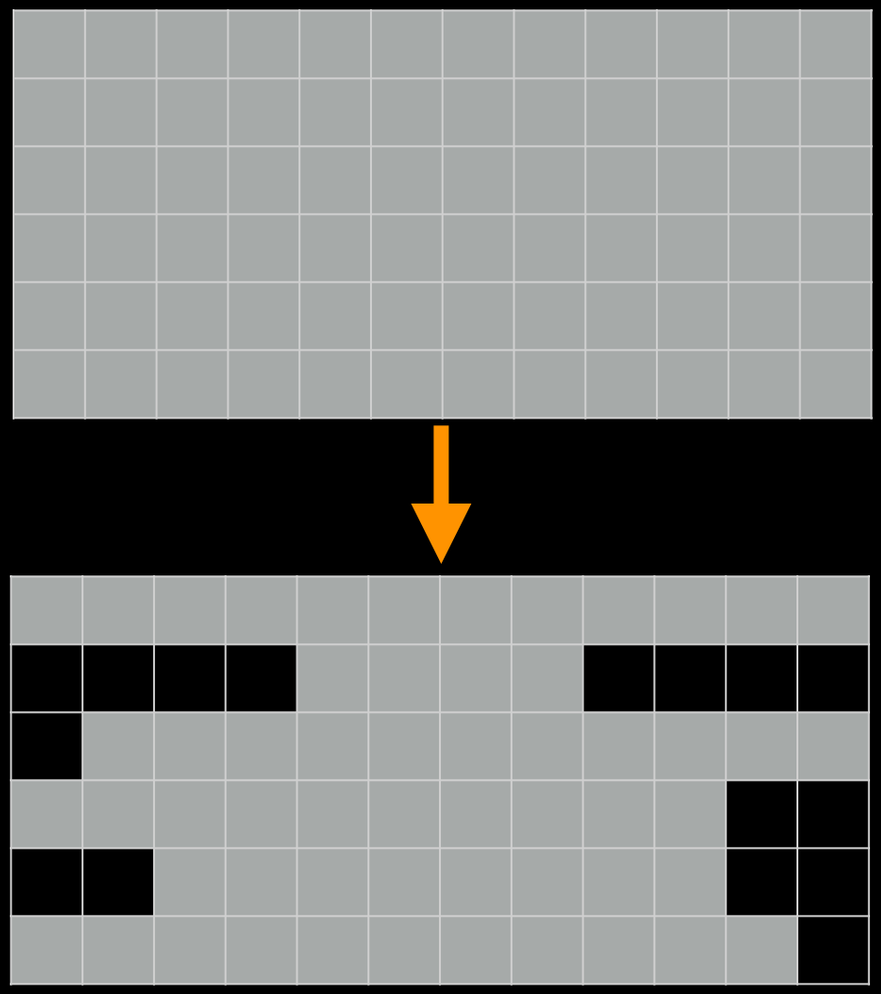
Most of us

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

Keeping paired ends paired: Options

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

NGS Data Quality: Base Quality Trimming



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

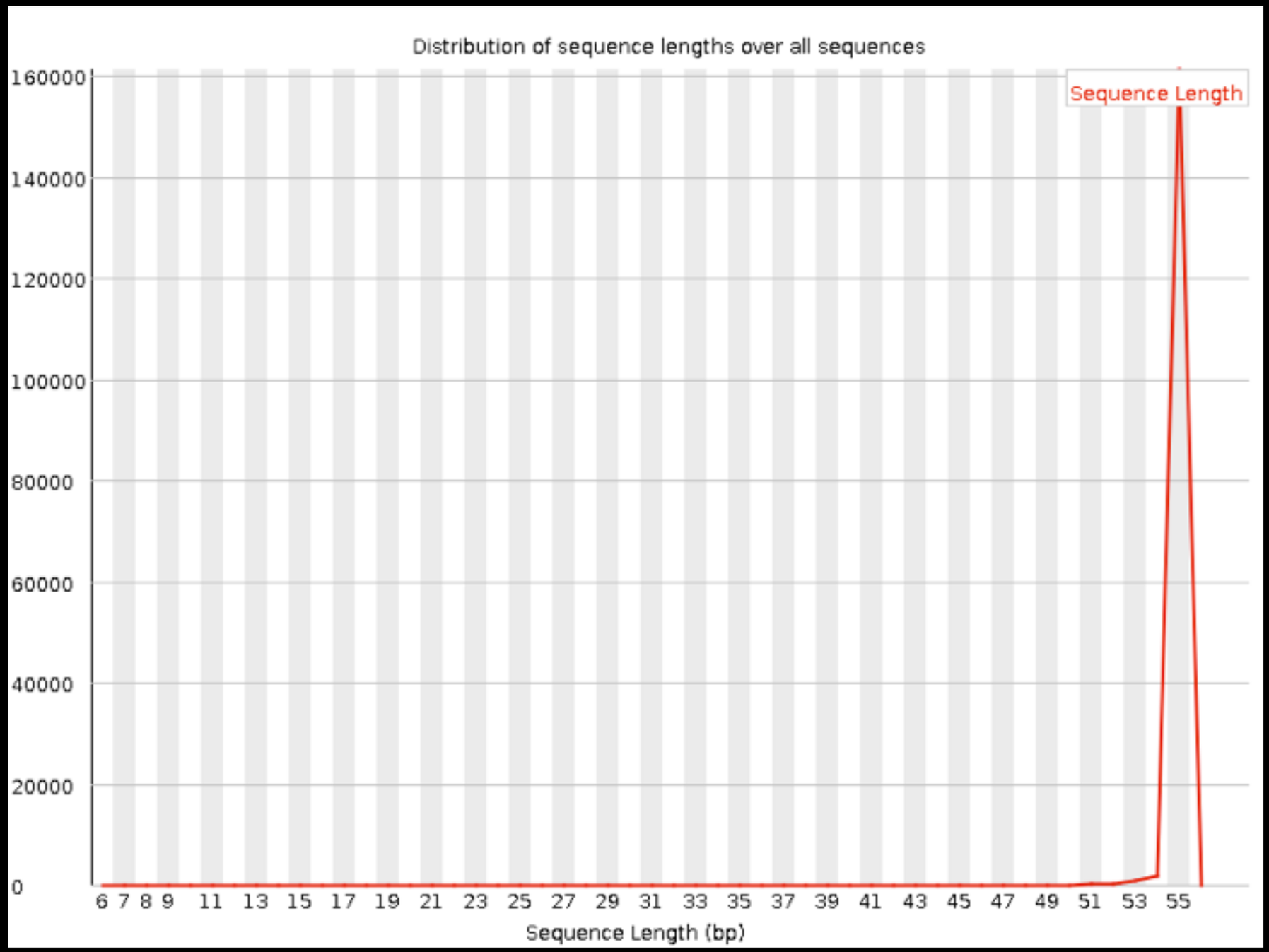
- NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**

Check **"Keep reads with zero length"**

Run again:

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

NGS Data Quality: Base Quality Trimming



New Problem?

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

Option 2 can fix this (but break pairings).

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

NGS Data Quality: Sequencing Artifacts

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

... and there's a problem in Overrepresented sequences:

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGTGTATTTGTCAATTTTCTTCTCCACGTTCTTCTCGGCCTGTTTCCGTAGCCT	590	0.3541692929220167	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	342	0.2052981325073385	No Hit
CGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	325	0.1950932545756871	No Hit
CTGCATTATAAAAAGGACAGCCAGATATCAACTGTTACAGAAATGAAATAAGACG	230	0.1380659955458709	No Hit
CGGCCGCAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	199	0.1194571004940361	No Hit
GTCAGCTCAACTTGTAGGCCCCAAAAGAAAACAGCGTCTTACTGGGGAGGGATAT	197	0.1182565266197242	No Hit

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

But this will break pairings.

NGS Data Quality: Done with 1st Replicate!

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

Workflows:

Create a QC workflow that does all these steps

Or, cheat and import the shared workflow.

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

NGS Data Quality: Further reading & Resources

FastQC Documenation

Read Quality Assessment & Improvement

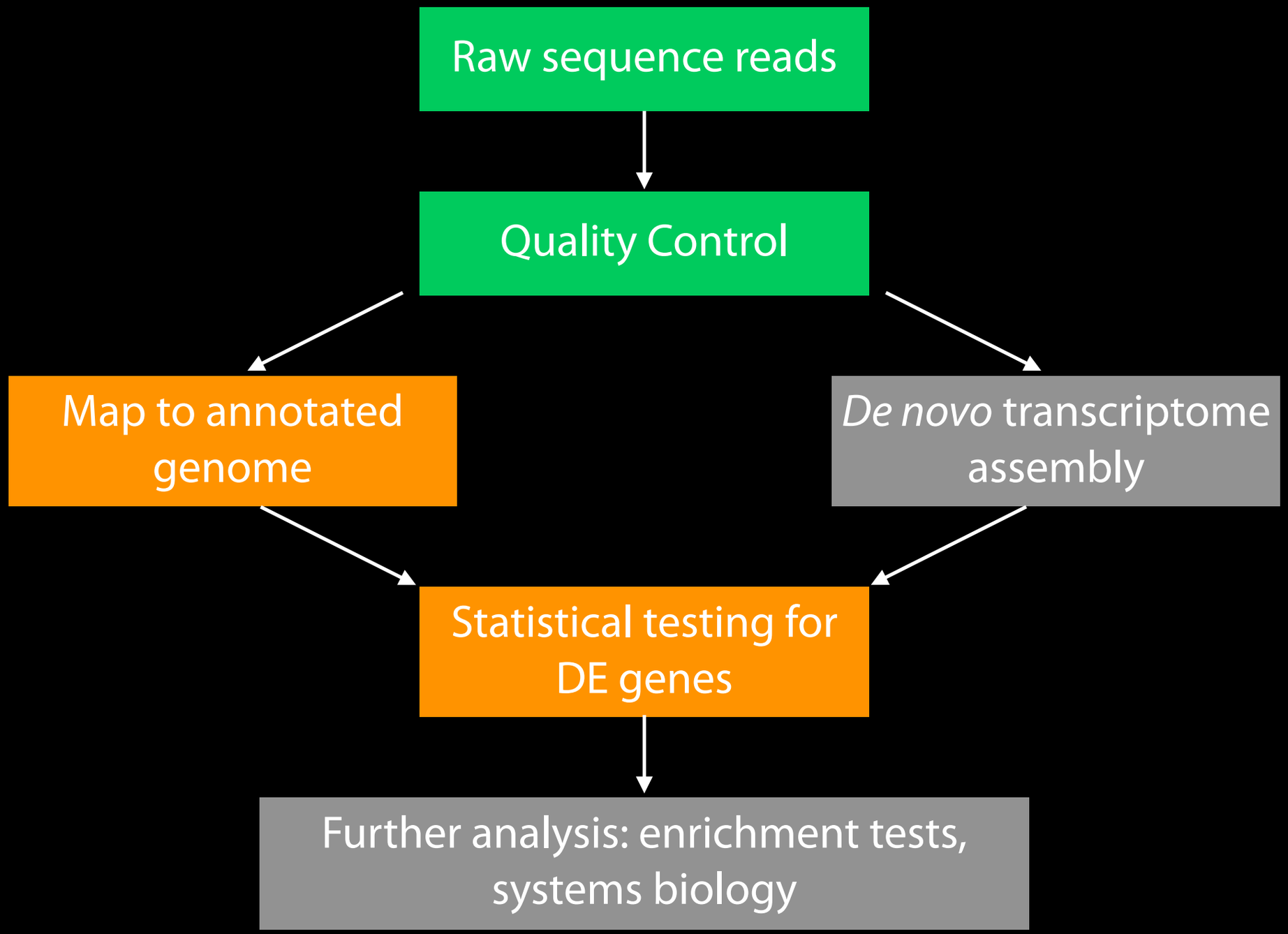
by Joe Fass

From the UC Davis 2013 Bioinformatics Short Course

Manipulation of FASTQ data with Galaxy

by Blankenberg, *et al*.

RNA-Seq DE analysis steps



RNA-seq Exercise: Mapping with TopHat

Create a new history

Import all datasets from library:

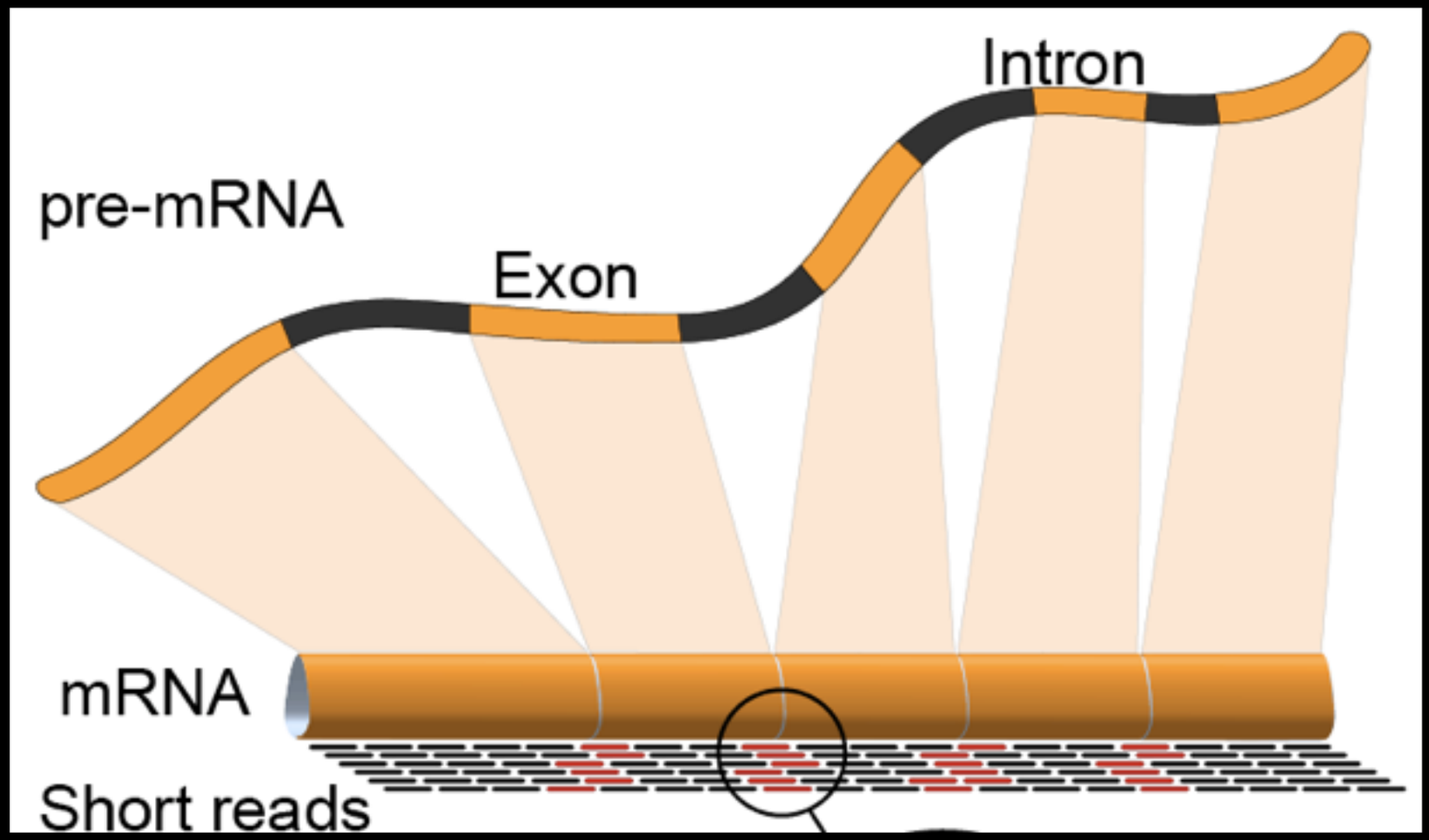
RNA-Seq Example → Reads, Post-QC, Re-Paired

All trimmed FASTQ and genes_chr12.gtf (from Reference directory)

NGS: RNA Analysis → TopHat2

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.*

Alignment output files

SAM

- plain text file, tab separated columns
- "a huge spreadsheet"
- inefficient to read and store

BAM

- a compressed version of SAM (~80% less storage)
- can be indexed (fast access to subsections)
- needs to be sorted to be useful however

Standardized format

- readable by most software

Anatomy of SAM file

```
Read1 113 1 497  37 37M    15 38662 0 CGGGTCTGACC  0;==-==9;>>> NM:i:0
Read2 213 1 497  37 37M    15 37662 0 CGGGTCTGACC  0;==-=45;>>> NM:i:1 XM:i:3
Read3 337 1 497  37 37M     15   38662 0 CGGGTCTGACC  ;===9;>>><>; NM:i:0
Read4 615 1 497  37 36MD1  15    447 0 CGGGTCTGACC  0;==5"=69;>> NM:i:0
Read5 844 1 497  37 37M    15  1445 0 CGGGTCTGACC  =-======9;>>> NM:i:0
```

One line per original read sequence

- Big!
- Where it aligned (if at all)
- How much of it aligned (soft/hard clipping)
- Mapping quality, likelihood correctly aligned
- Any differences to the reference (CIGAR string)
- Lots of other stuff (aligner dependent)
- Does not contain the reference sequence

TopHat2 basic parameters

- We're using Paired-end data
- Specifying two mate-pair FASTQ input files
- Mapping against hg19 reference genome

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
- Only look for supplied junctions → Yes

Mapping with TopHat: Make it quicker?

Warning: Here be dragons!

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

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

TopHat Manual

Mapping with TopHat: Max # of Alignments Allowed

Some reads align to more than one place equally well.

For such reads, how many should TopHat include?

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

TopHat break ties randomly.

TopHat assigns equal fractional credit to all *n*

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

TopHat Manual

Mapping with TopHat: Lets do it some more!

NGS: RNA Analysis → TopHat2

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

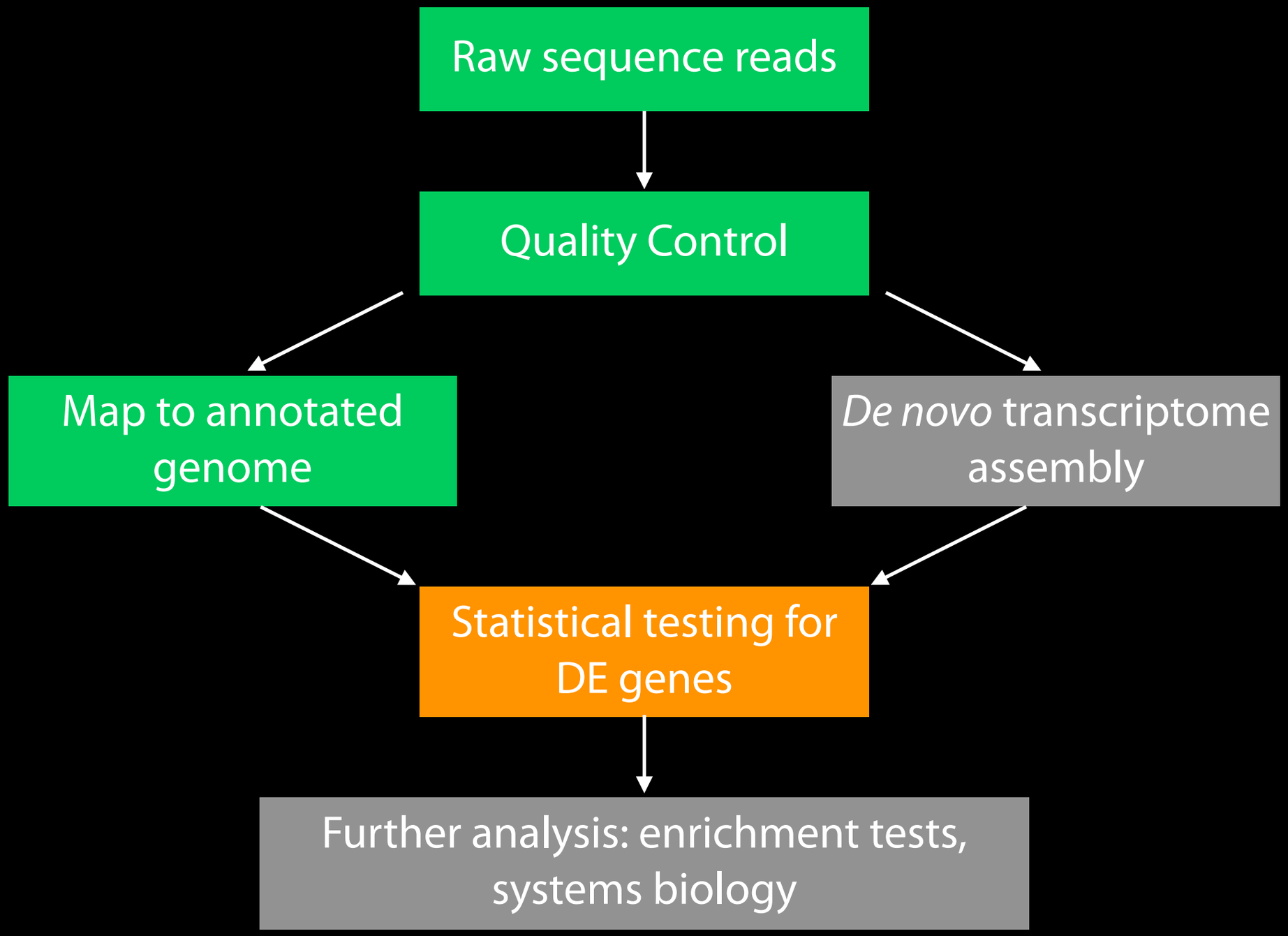
RNA-Seq Analysis with Galaxy
by Jeroen F.J. Laros, Wibowo Arindrarto, Leon Mei

from the GCC2013 Training Day

RNA-Seq Analysis with Galaxy
by Curtis Hendrickson, David Crossman, Jeremy Goecks

from the GCC2012 Training Day

RNA-Seq DE analysis steps




HELP

http://wiki.galaxyproject.org

Galaxy

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

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

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

Use Galaxy

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

usegalaxy.org

Deploy Galaxy

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

- Admin
- Cloud
- Galaxy Appliance

getgalaxy.org

Community & Project

Galaxy has a large and active user community and many ways to Get Involved.

- Community
- News
- Events
- Support
- Galaxy Project

Contribute

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

Use Galaxy

Servers • Learn
Main • Share • Search

Communicate

Support • News
Events • Twitter
Mailing Lists (search)

Deploy Galaxy

Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

SLIPSTREAM APPLIANCE Galaxy made easy.

Contribute

Tool Shed • Share
Issues & Requests
Teach • Support

Events

News

Galaxy Wiki

EnisAfgan 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 d
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
send it to outreach@glaxyproject.org.

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

Upcoming Events

VIB Galaxy Baltimore Ga ADM

Date	Topic/Event	Venue/Location	
January 15-16	*Accessible and Reproducible Genomics at Scale with Galaxy*	Revolutionizing Next-Gene Belgium	
January 19-20	*NGS pipelines with Galaxy*	e-Infrastructures for Massi Sweden	
January 22	*Baltimore Area Galaxy Meetup*	Baltimore, Maryland, Unite	
January 28	*ChIP-Seq Analysis and Visualization using Galaxy and IGB Tutorial*	Online	
February 9-13	*Analyse bioinformatique de séquences sous Galaxy*	Montpellier, France	J.F. Dufayard
	Accessible and Reproducible Large-Scale	Genome and Transcriptome Analysis, part of Molecular Medicine Tri-	

Galaxy Wiki

Login | Search:

News

News

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

The Galaxy News is also available as an RSS feed.

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

See also

- Galaxy News Briefs
- Galaxy Updates
- Galaxy on Twitter
- Events
- Learn
- Support
- About the Galaxy Project

News Items

January 2014 CloudMan Release
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

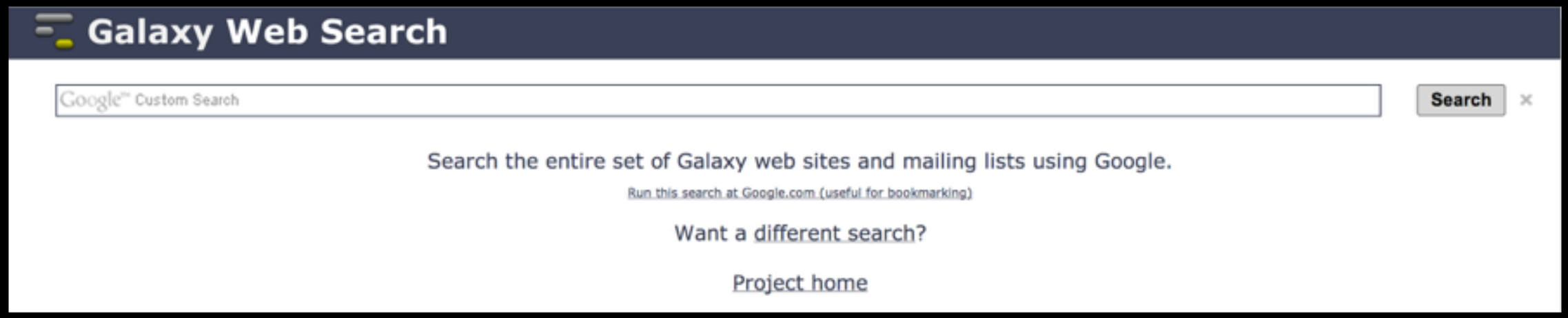
January 2014 CloudMan Release

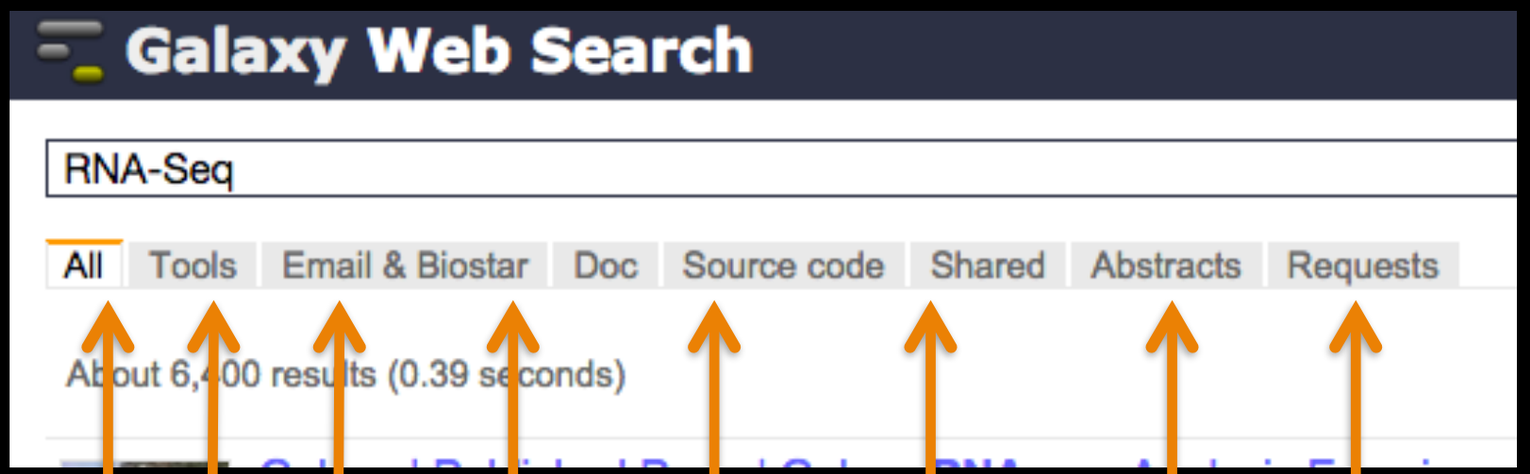
We just released an update to Galaxy CloudMan. CloudMan offers an easy way to get a personal and completely functional instance of Galaxy in the cloud in just a few minutes, without any manual configuration.

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

CloudMan

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





Find

Everything on …

Tools for …

Email about …

Source code for …

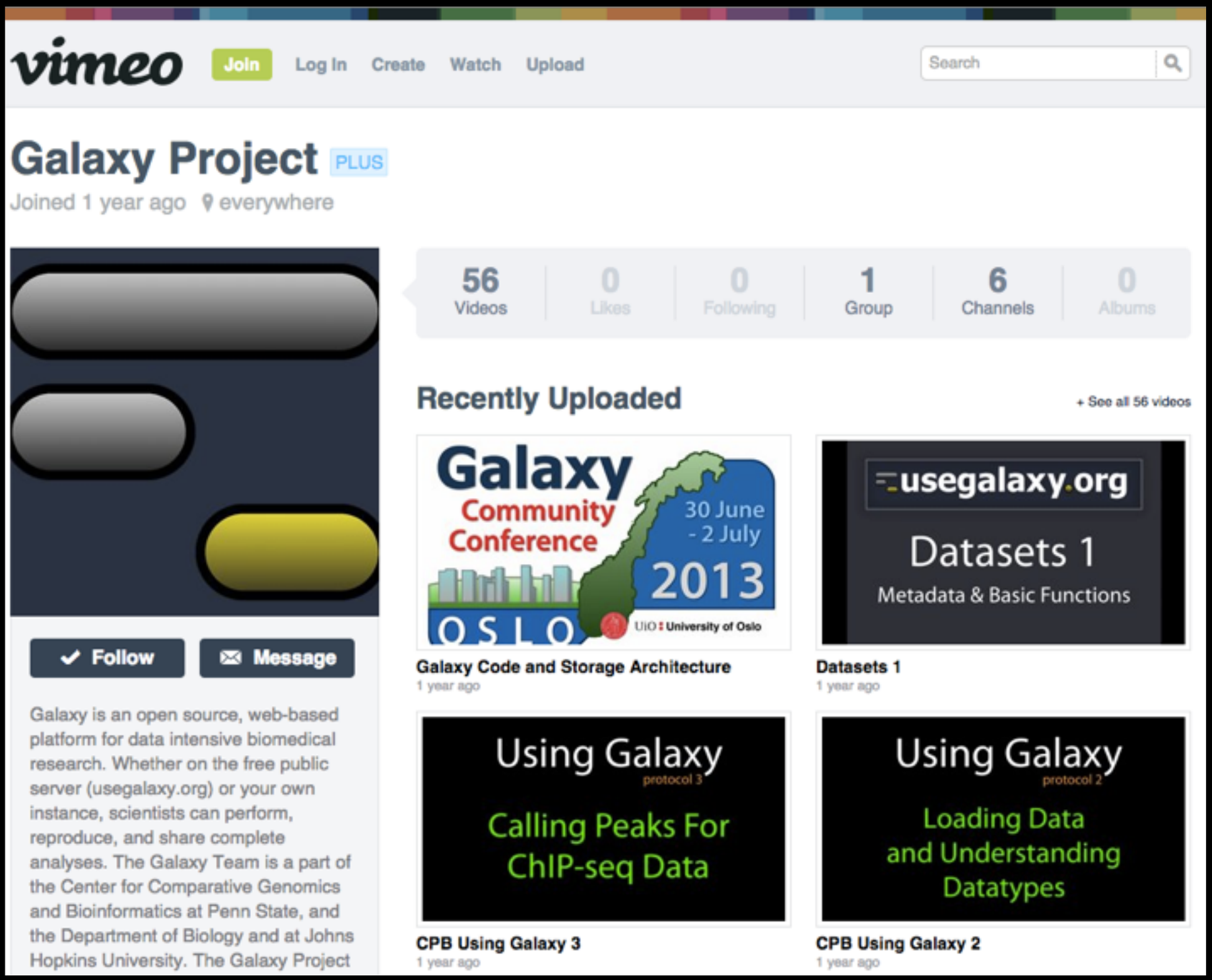
Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

Galaxy Resources & Community: Videos



"How to" screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject

Galaxy Resources and Community: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (47 posts in 2014, 4100+ members)

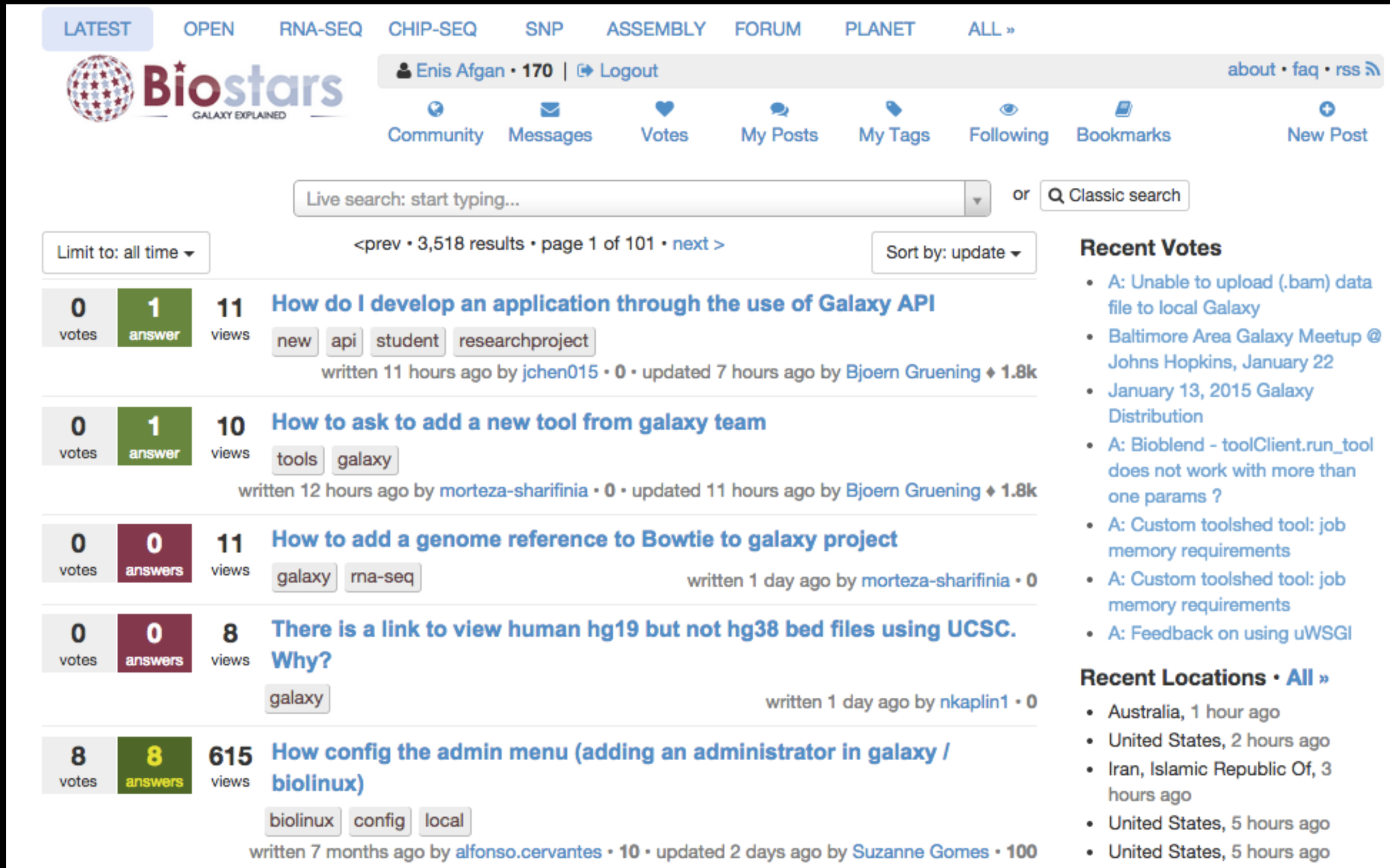
Galaxy-Dev

Questions about developing for and deploying Galaxy

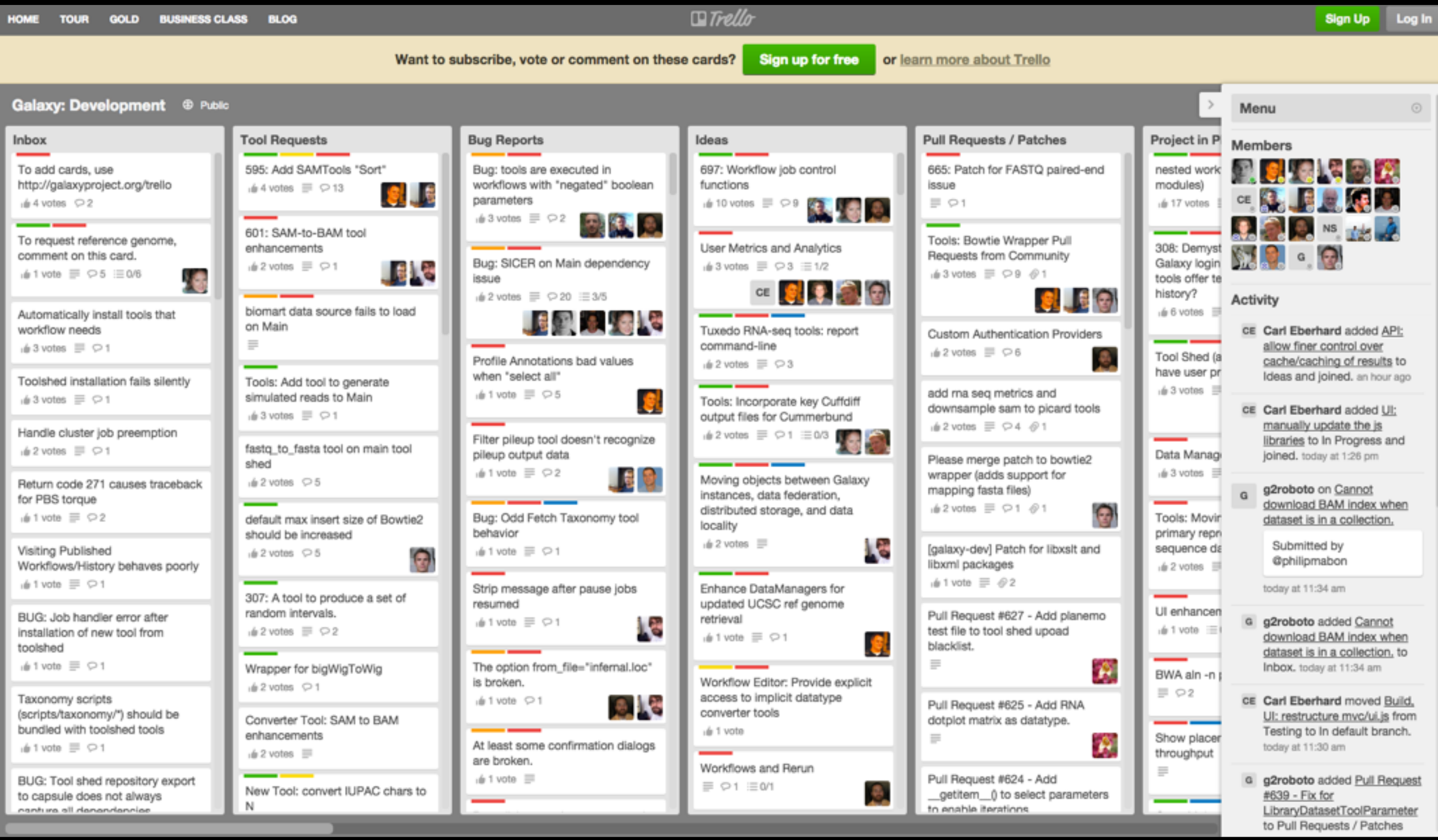
High volume (2700 posts in 2014, 1000+ members)

More interactive help?

https://biostar.usegalaxy.org/



Community can create, vote and comment on issues



http://bit.ly/gxytrello

wiki.galaxyproject.org/Events/GCC2015

Galaxy Resources & Community: CiteULike Group

citeulike

CiteULike | Group: Galaxy | Search | Register | Log in

Group: Galaxy - library 2020 articles

Search | Copy | Export | Sort | Hide Details

✔ **Open pipelines for integrated tumor genome profiles reveal differences between pancreatic cancer tumors and cell li**
Cancer Medicine (December 2014), pp. n/a-n/a, doi:10.1002/cam4.360
by Jeremy Goecks, Bassel F. El-Rayes, Shishir K. Maithel, H. Jean Khoury, James Taylor, Michael R. Rossi
posted to methods shared usemain by galaxyproject to the group Galaxy on 2015-01-09 00:22:41 ★★★★★

✔ **RNA-Seq reveals a xenobiotic stress response in the soybean aphid, Aphis glycines, when fed aphid-resistant soybea**
BMC Genomics, Vol. 15, No. 1. (16 November 2014), 972, doi:10.1186/1471-2164-15-972
by Raman Bansal, M. A. R. Mian, Omprakash Mittapalli, Andy P. Michel
posted to methods uselocal by galaxyproject to the group Galaxy on 2015-01-09 00:15:52 ★★
■ Abstract

✔ **Draft Genome Sequence of the Cellulolytic Bacterium Clavibacter sp. CF11, a Strain Producing Cold-Active Cellulase**
Genome Announcements, Vol. 3, No. 1. (26 February 2015), e01304-14, doi:10.1128/genomea.01304-14
by Ying Du, Bo Yuan, Yonghui Zeng, et al.
posted to methods usemain by galaxyproject to the group Galaxy on 2015-01-09 00:03:34 ★★
■ Abstract

✔ **Identifying transcriptional cis -regulatory modules in animal genomes**
Wiley Interdisciplinary Reviews: Developmental Biology (December 2014), pp. n/a-n/a, doi:10.1002/wdev.168
by Kushal Suryamohan, Marc S. Halfon
posted to refpublic by galaxyproject to the group Galaxy on 2015-01-09 00:01:02 ★★

✔ **Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis**
Genome Research, Vol. 24, No. 12. (01 December 2014), pp. 1945-1962, doi:10.1101/gr.164830.113
by Weisheng Wu, Christapher S. Morrissey, Cheryl A. Keller, et al.
posted to methods by galaxyproject to the group Galaxy on 2015-01-08 23:56:32 ★★
■ Abstract

✔ **Rampant software errors undermine scientific results**
F1000Research (11 December 2014), doi:10.12688/f1000research.5930.1
by David A. W. Soergel
posted to workbench by galaxyproject to the group Galaxy on 2015-01-08 23:49:39 ★★ along with 2 people and 1 group

✔ **Reproducibility in Science**
Circulation Research, Vol. 116, No. 1. (02 January 2015), pp. 116-126, doi:10.1161/circresaha.114.303819
by C. Glenn Begley, John P. A. Ioannidis
posted to reproducibility by galaxyproject to the group Galaxy on 2015-01-08 23:21:23 ★★ along with 1 person
■ Abstract

Group Tags

All tags in the group Galaxy

Filter:

[Display as Cloud]

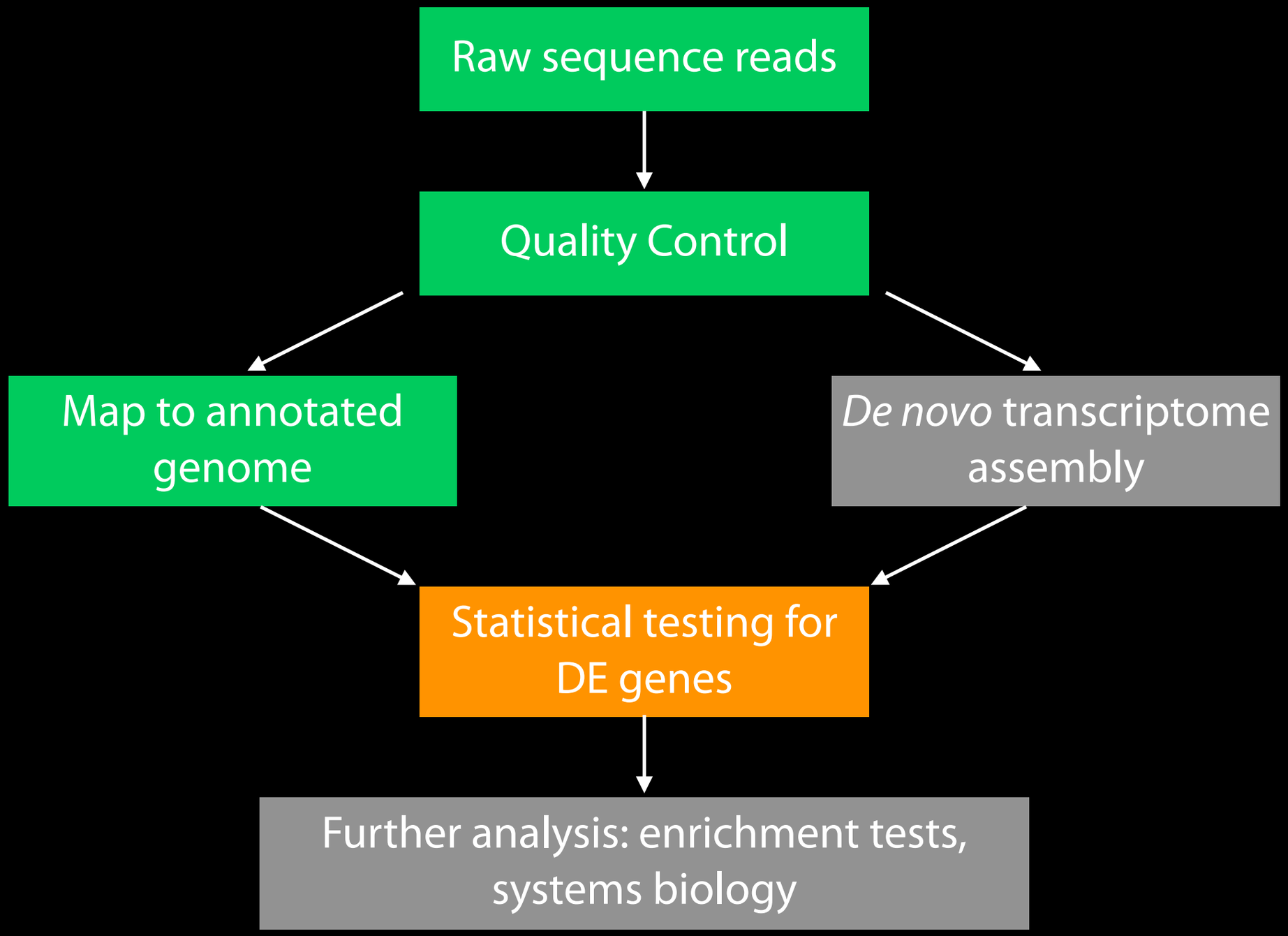
methods	973
workbench	633
usemain	190
tools	146
isgalaxy	115
cloud	77
usepublic	76
uselocal	70
shared	70
other	54
unknown	51
reproducibility	42
howto	41
project	38
refpublic	38
visualization	12
usecloud	3

Over 2000 papers

17 tags

http://bit.ly/gxycul

RNA-Seq DE analysis steps



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

- Running with 2 Groups: MeOH and R3G
- Each group has 3 replicates
- genes_chr12.gtf as the transcript

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.

Try running Cuffdiff with less replicas.

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

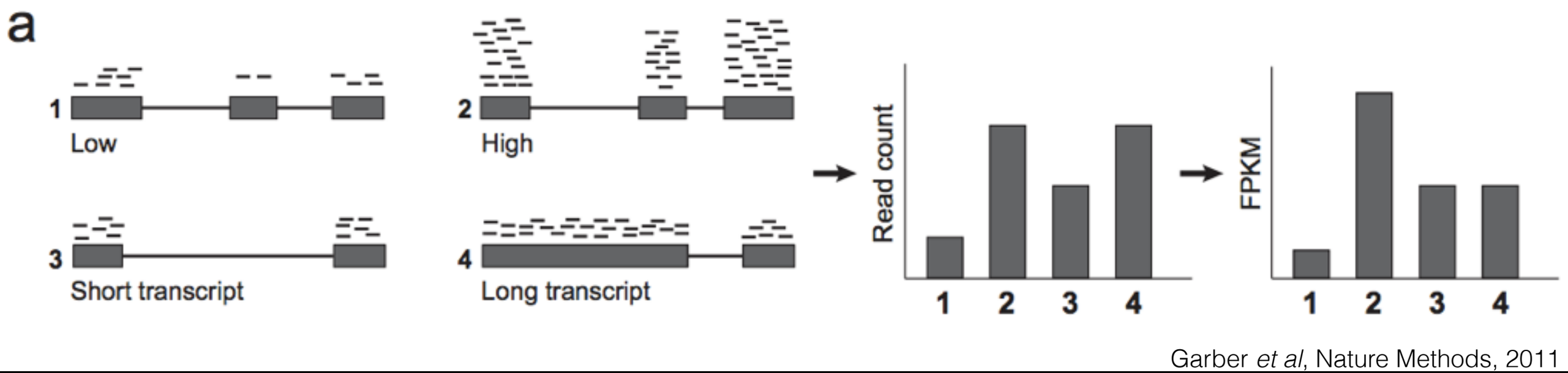
RNA-Seq Analysis with Galaxy and Alternative Tools

by Saskita Hiltemann, Youri Hoogstrate & Leon Mei

from the GCC2014Training Day

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.



Garber *et al*, Nature Methods, 2011

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

Voom+Limma **EdgeR** **DESeq**

R packages available as Galaxy tools; isolate between features based on expression counts and show a total count of a gene including all its isoforms

Cuffdiff Alternatives

Take 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

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

Join the 6 HTSeq datasets together on gene name

Cut out the duplicate gene name columns

NGS: RNA Analysis → Differential Count

Cuffdiff Alternatives: Differential Counts

Output from three tools: Voom, EdgeR, 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
for each of the tools?

After DGE RNA-seq?

You get your "gene list", finished?

- Validate
- Typically expect some false-positives
- Genes not in your list may be differentially expressed

Important to always remember

- Your list of genes is produced with an arbitrary significance threshold!

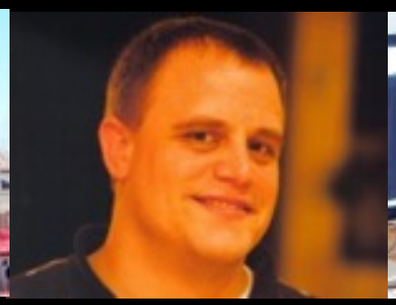
Next?

- Gene-set enrichment tests
- Novel transcripts, novel splice-variants, ...

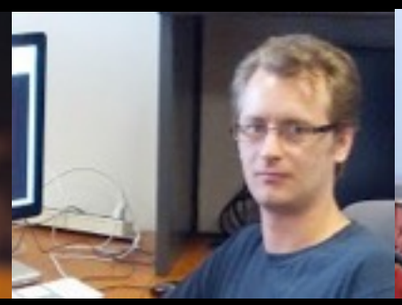
The Galaxy Team

Enis Afgan

Dannon Baker



Dan Blankenberg



Dave Bouvier

Marten Cech

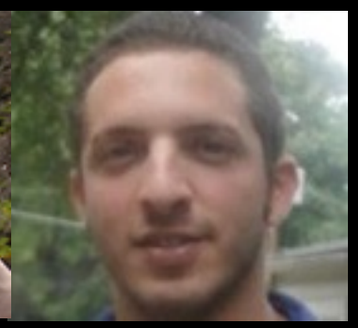
John Chilton

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

Thanks

Enis Afgan
Dave Clements
Clare Sloggett
Andrew Sharp

Galaxy Project
Johns Hopkins University
outreach@galaxyproject.org

amazon web services