Introduction to Galaxy

Icahn School of Medicine at Mount Sinai
January 22, 2016

Dave Clements
Galaxy Team
Johns Hopkins University
http://galaxyproject.org/

#usegalaxy @galaxyproject

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

http://bit.ly/gxyismms2016

Goals

Provide an introduction to using Galaxy for bioinformatic analysis. Demonstrate how Galaxy can help you explore and learn options, perform analysis, and then share, repeat, and reproduce your analyses.

This workshop does cover RNA-Seq but you won't be an expert at the end of the workshop. You will know enough to get started, and how to use Galaxy to learn more.

What is Galaxy?

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

http://galaxyproject.org

What is Galaxy?

Keith Bradnam's definition:

"A web-based platform that provides a simplified interface to many popular bioinformormatics tools."

From

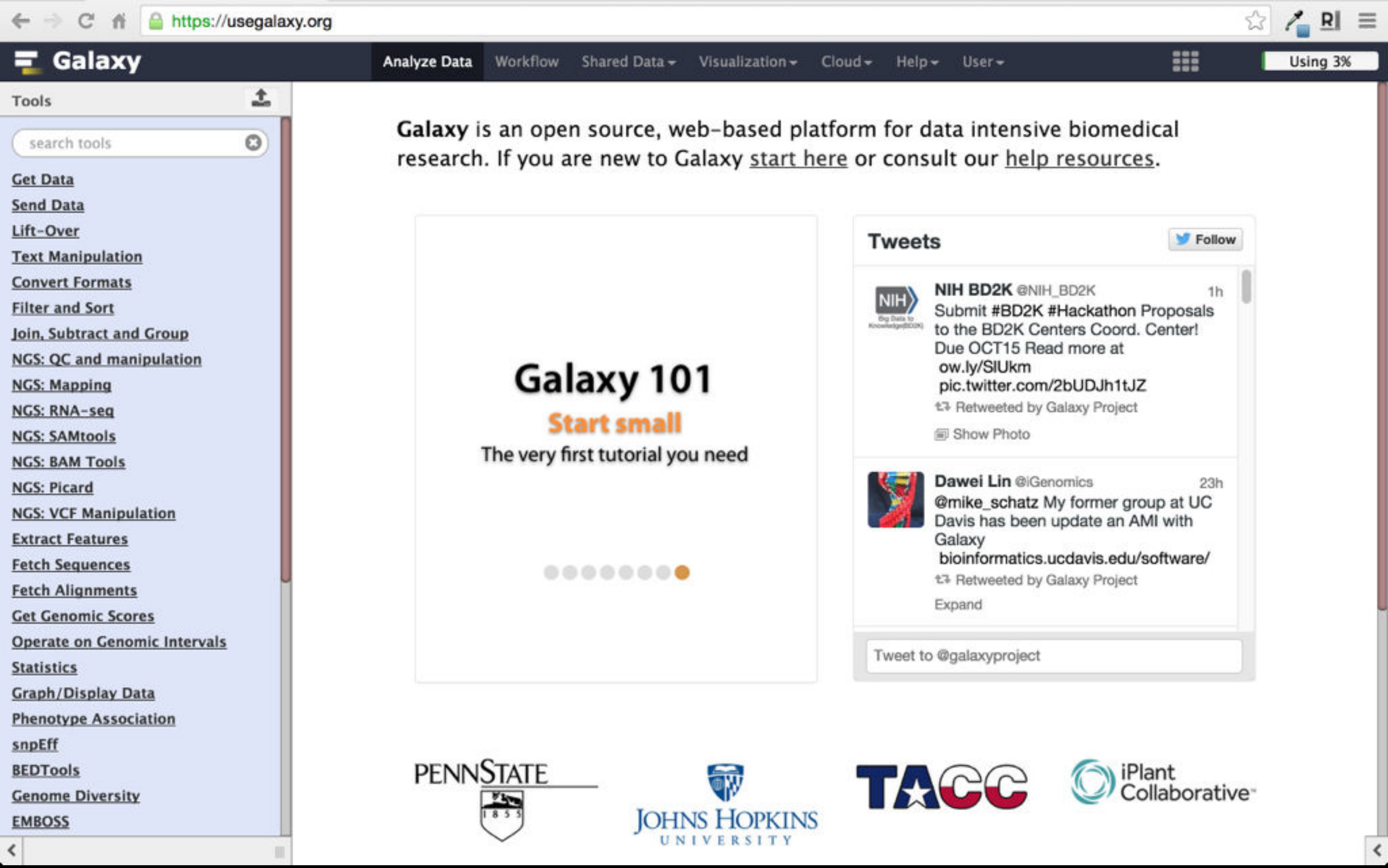
"13 Questions You May Have About Galaxy"

http://bit.ly/13questions

Galaxy is available several ways ...

http://galaxyproject.org

As a free for everyone service on the web: usegalaxy.org



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.

Galaxy 101

Start small

The very first tutorial you need

Tweets

NIH BD2K @NIH_BD2K 1h
Submit #BD2K #Hackathon Proposals to the BD2K Centers Coord. Center! Due OCT15 Read more at ow.ly/SlUkm pic.twitter.com/2bUDJh1tJZ
Retweeted by Galaxy Project

Dawei Lin @iGenomics 23h
@mike_schatz My former group at UC Davis has been update an AMI with Galaxy bioinformatics.ucdavis.edu/software/
Retweeted by Galaxy Project

Tweet to @galaxyproject

A free for everyone web service:

http://usegalaxy.org

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

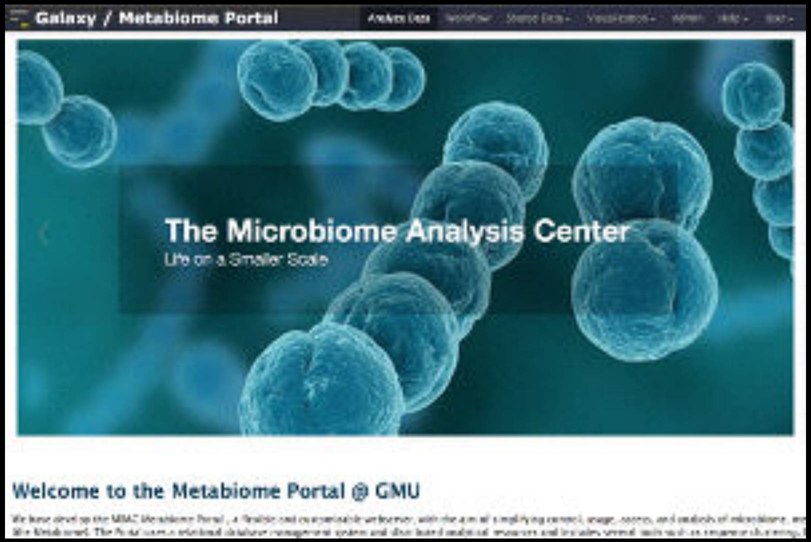
TACC

CYVERSE™

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

Explore the Galaxy with **RNA-Rocket**

PATHOGENPORTAL
THE BIOINFORMATICS RESOURCE CENTERS PORTAL



Galaxy / Metabiome Portal

The Microbiome Analysis Center
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GMU

香港中文大學 - 華大基因跨組學創新研究院
CUHK-BGI Innovation Institute of Trans-Omics
华大基因 BGI

(GIGA)^n Galaxy
by CBIIT

Integrated publishing of workflows from (GIGA)^n SCIENCE

Cistrome

A Galaxy Server dedicated to ChIP-* analysis

070+

Public Galaxy Servers
and *still* counting

The Genomic HyperBrowser

Powered by Galaxy

SCDE
STEM CELL DISCOVERY ENGINE

Experiments Connected

Whale Shark Galaxy!

South Green®
bioinformatics platform

Genomic analysis tools for southern and Mediterranean plants

bit.ly/gxyServers

Galaxy is available as Open Source Software

Galaxy is installed in locations around the world.

http://getgalaxy.org

Galaxy is available on the Cloud

We are using this today

http://aws.amazon.com/education
http://globus.org/
http://wiki.galaxyproject.org/Cloud

Galaxy on the Cloud: Galaxy CloudMan

http://usegalaxy.org/cloud

- Start with a **fully configured and populated** (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

http://bit.ly/gxyismms2016

Quick Poll: Are you ...

1. A bioinformatics novice
2. A bioinformatics apprentice
3. A bioinformatics guru

Yes, those are your only choices.

http://galaxyproject.org

Basic Analysis

Which exons have most overlapping Repeats?

Use Human, HG38, GENCODE v23, Chromosome 22

cloud1.galaxyproject.org

cloud2.galaxyproject.org

cloud3.galaxyproject.org

cloud4.galaxyproject.org

Exons & Repeats: A General Plan

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

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

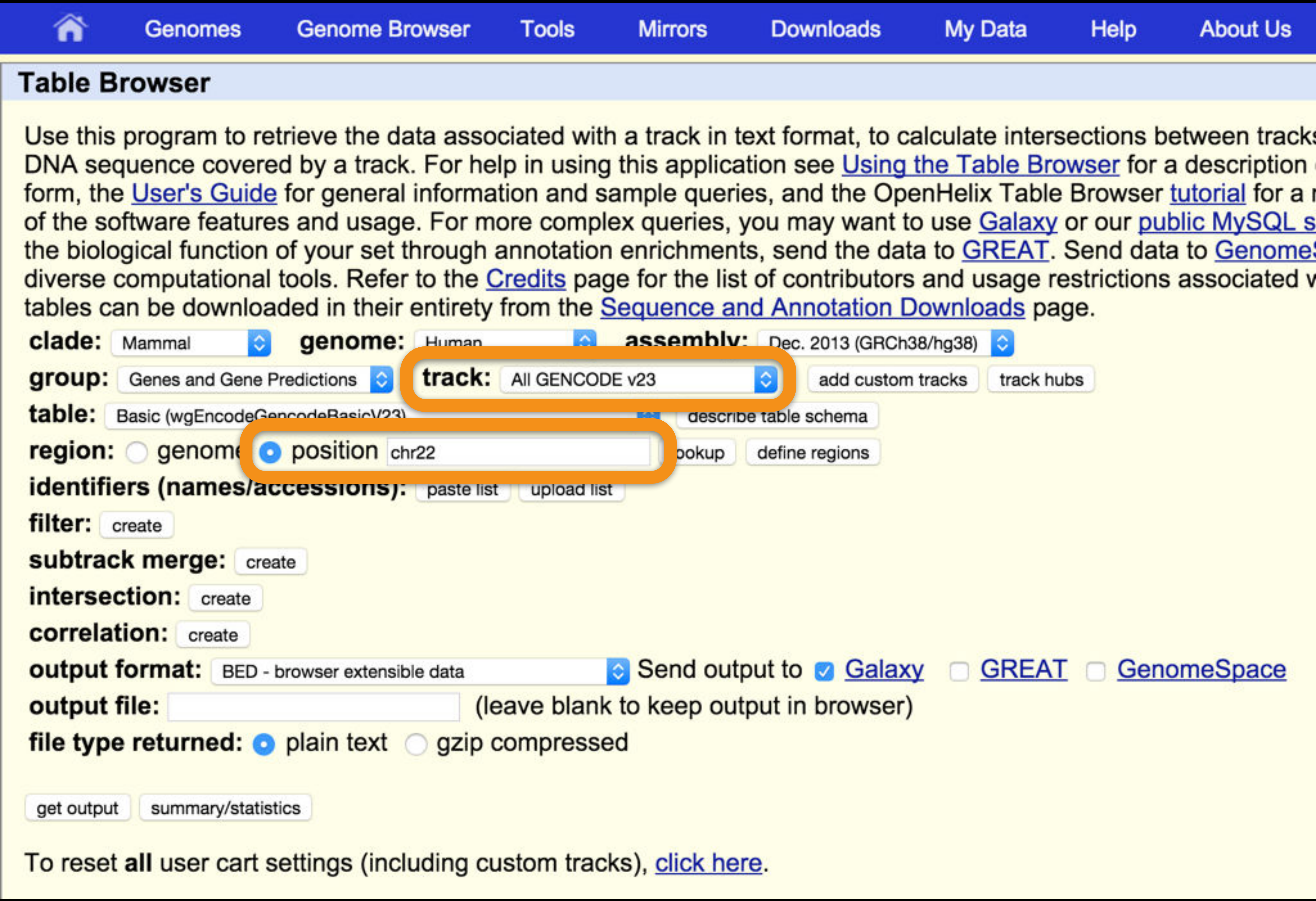


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a r of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL s the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeS diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated w tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions **track:** All GENCODE v23 add custom tracks track hubs

table: Basic (wgEncodeGencodeBasicV23) describe table schema

region: ○ genome ● position chr22 ookup define regions

identifiers (names/accessions): paste list upload list

filter: create

subtrack merge: create

intersection: create

correlation: create

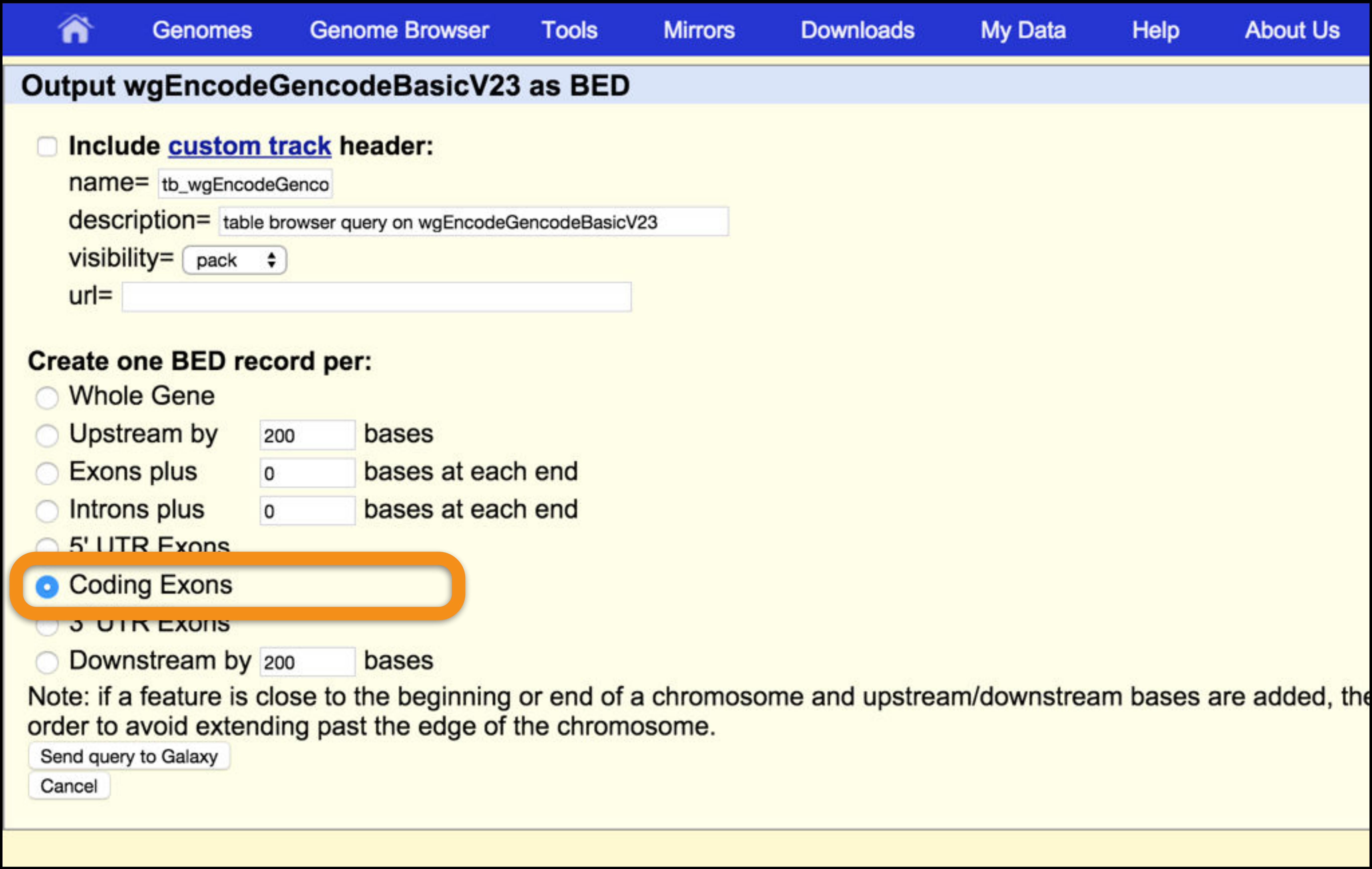
output format: BED - browser extensible data Send output to ☑ Galaxy □ GREAT □ GenomeSpace

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.



Output wgEncodeGencodeBasicV23 as BED

□ **Include custom track header:**

name= tb_wgEncodeGenco

description= table browser query on wgEncodeGencodeBasicV23

visibility= pack

url=

Create one BED record per:

○ Whole Gene

○ Upstream by 200 bases

○ Exons plus 0 bases at each end

○ Introns plus 0 bases at each end

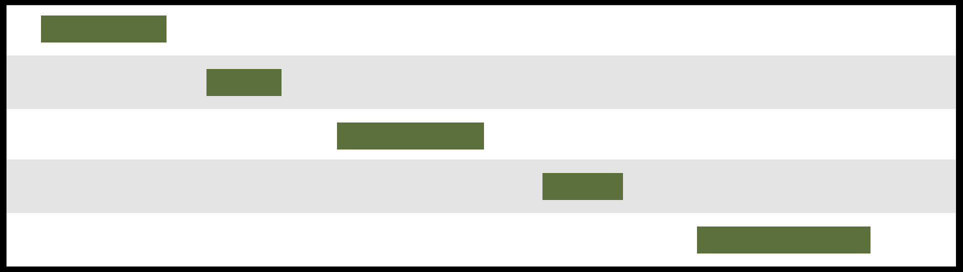
○ 5' UTR Exons

● Coding Exons

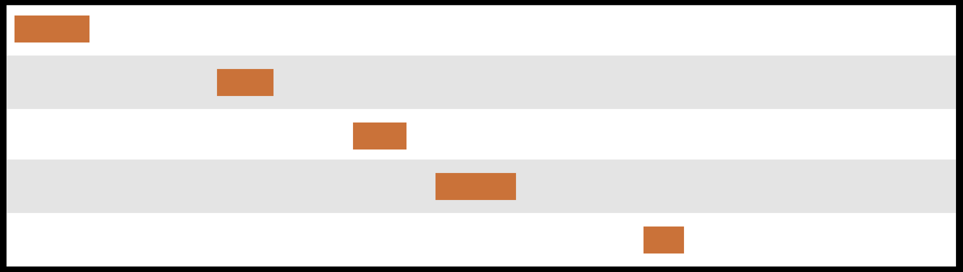
○ 3' UTR Exons

○ Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, the
order to avoid extending past the edge of the chromosome.

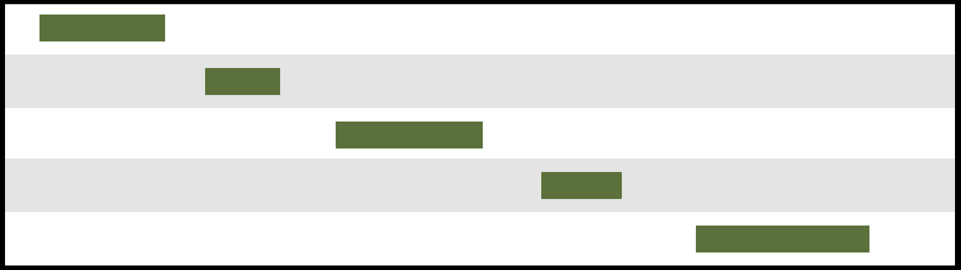


Exons

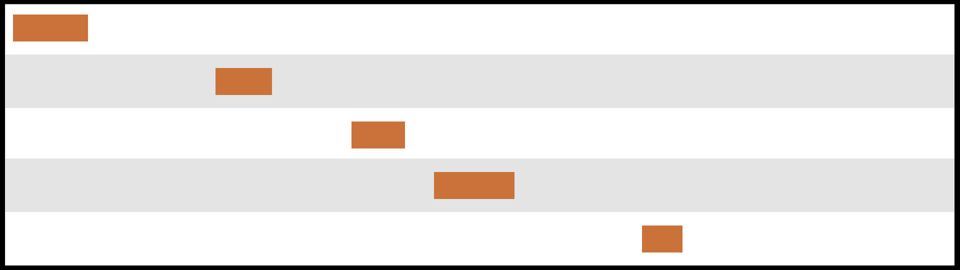


Repeats

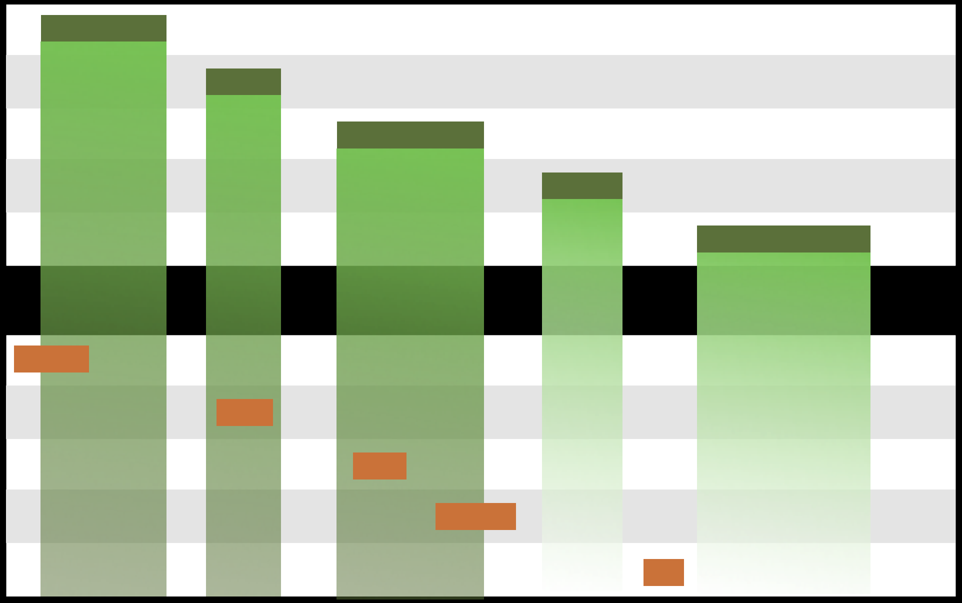
(Identify which exons have Repeats)



Exons



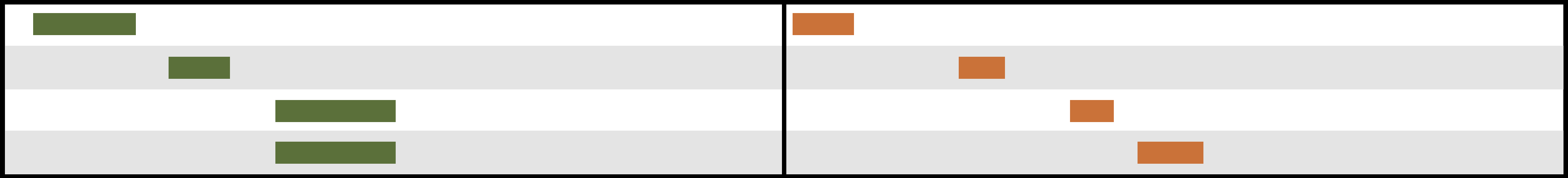
Repeats



Exons

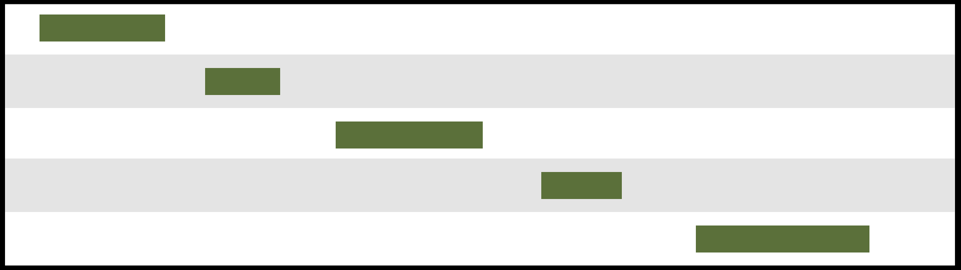
Repeats

Overlap pairings

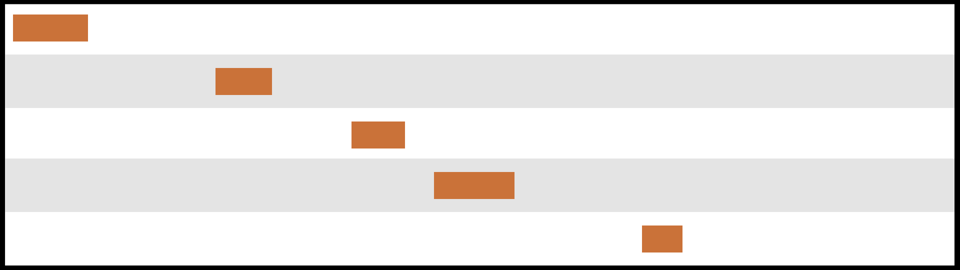


Operate on Genomic Intervals → **Join**

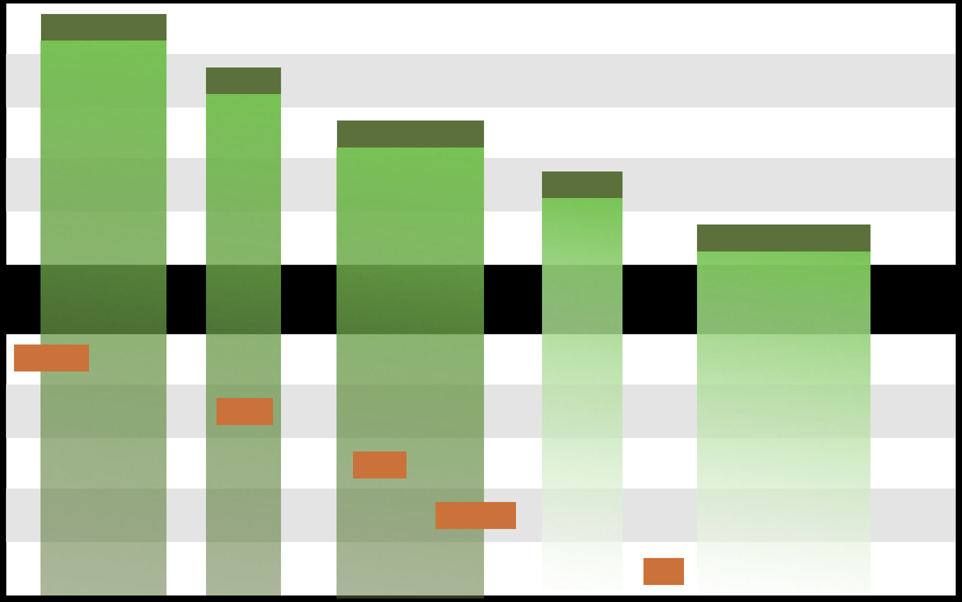
(Identify which exons have Repeats)



Exons



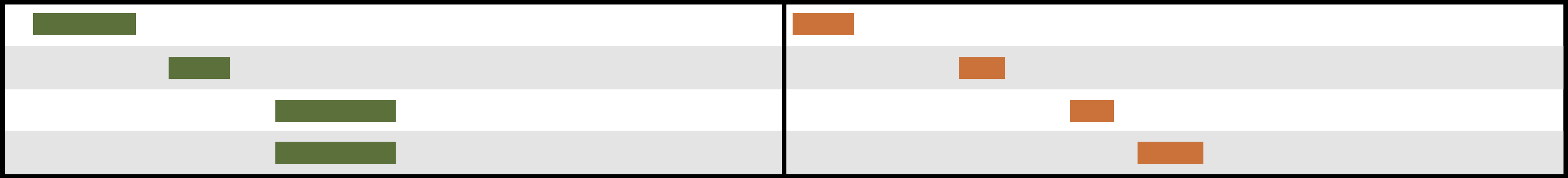
Repeats



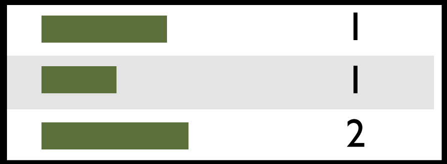
Exons

Repeats

Overlap pairings



(Count Repeats per exon)



Exon overlap counts

Join, Subtract, and Group → Group

Published History: Exons with overlapping repeats

Yay! But, a wee challenge

We have exon names and counts

Really want genes (or transcripts) and counts across the whole gene (or transcript)

Also see "101: Getting back exon info" at end of the slides

What we have: Computer generated Exon IDs

ENST00000073150.2_cds_0_0_chr22_15528159_f

Ensembl transcript ID and version number are embedded in Exon ID.

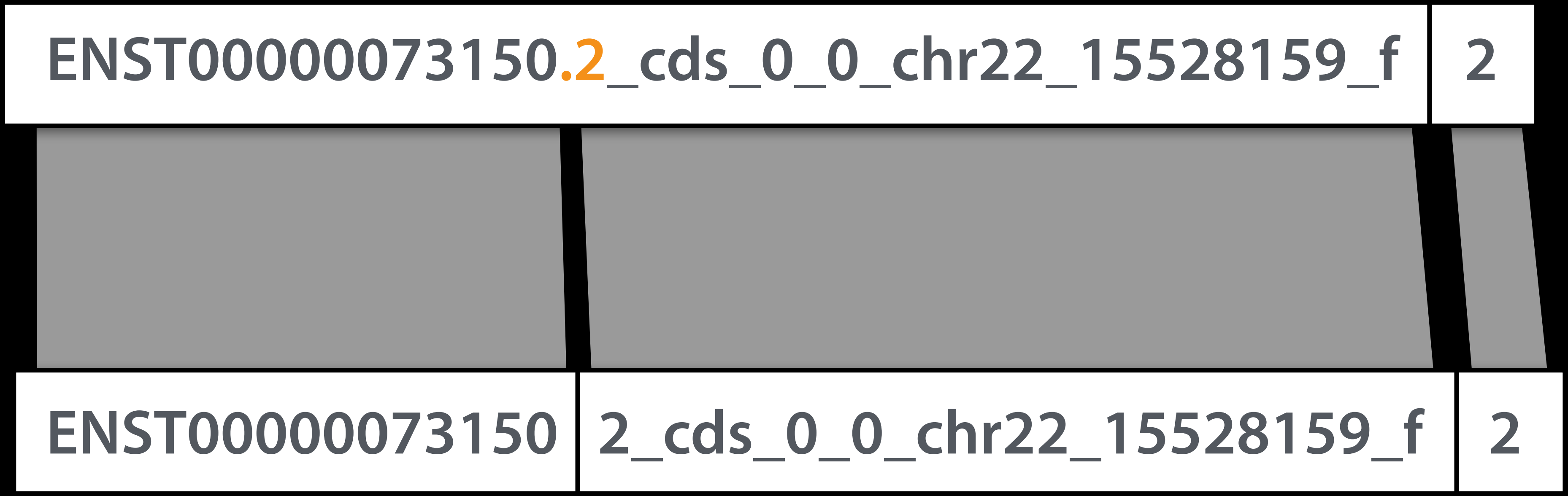
How can we extract the Transcript ID from the Exon ID?

(With the transcript ID we could summarize counts for each transcript and get the gene ID.)

Extract the transcript ID

Need to decide if we should keep transcript version or not.

Using our ability to see into the future, we decide not to keep it.

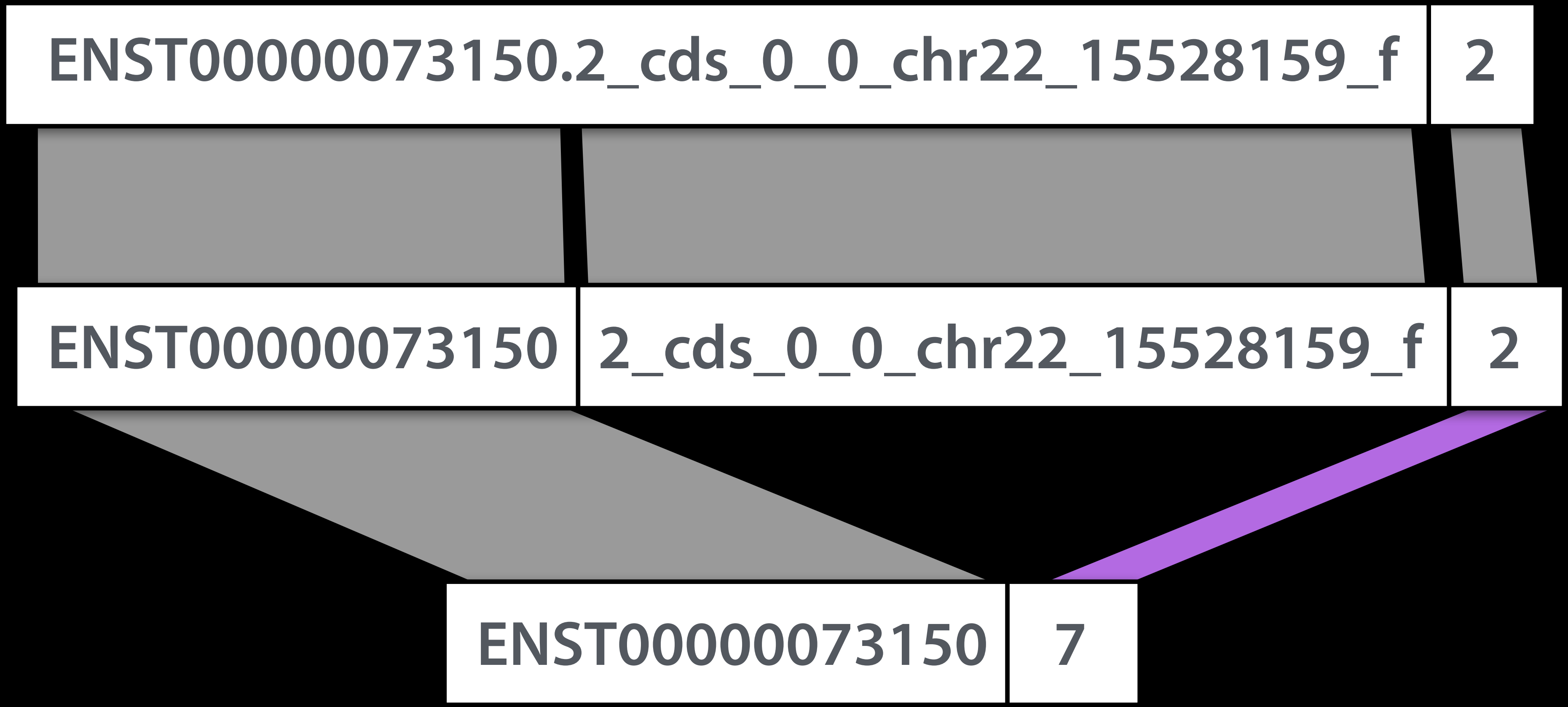


ENST00000073150.2_cds_0_0_chr22_15528159_f | 2

ENST00000073150 | 2_cds_0_0_chr22_15528159_f | 2

Text Manipulation → Convert delimiters to TAB
(converting dots instead of underscores)

Sum the scores for all exons in each transcript



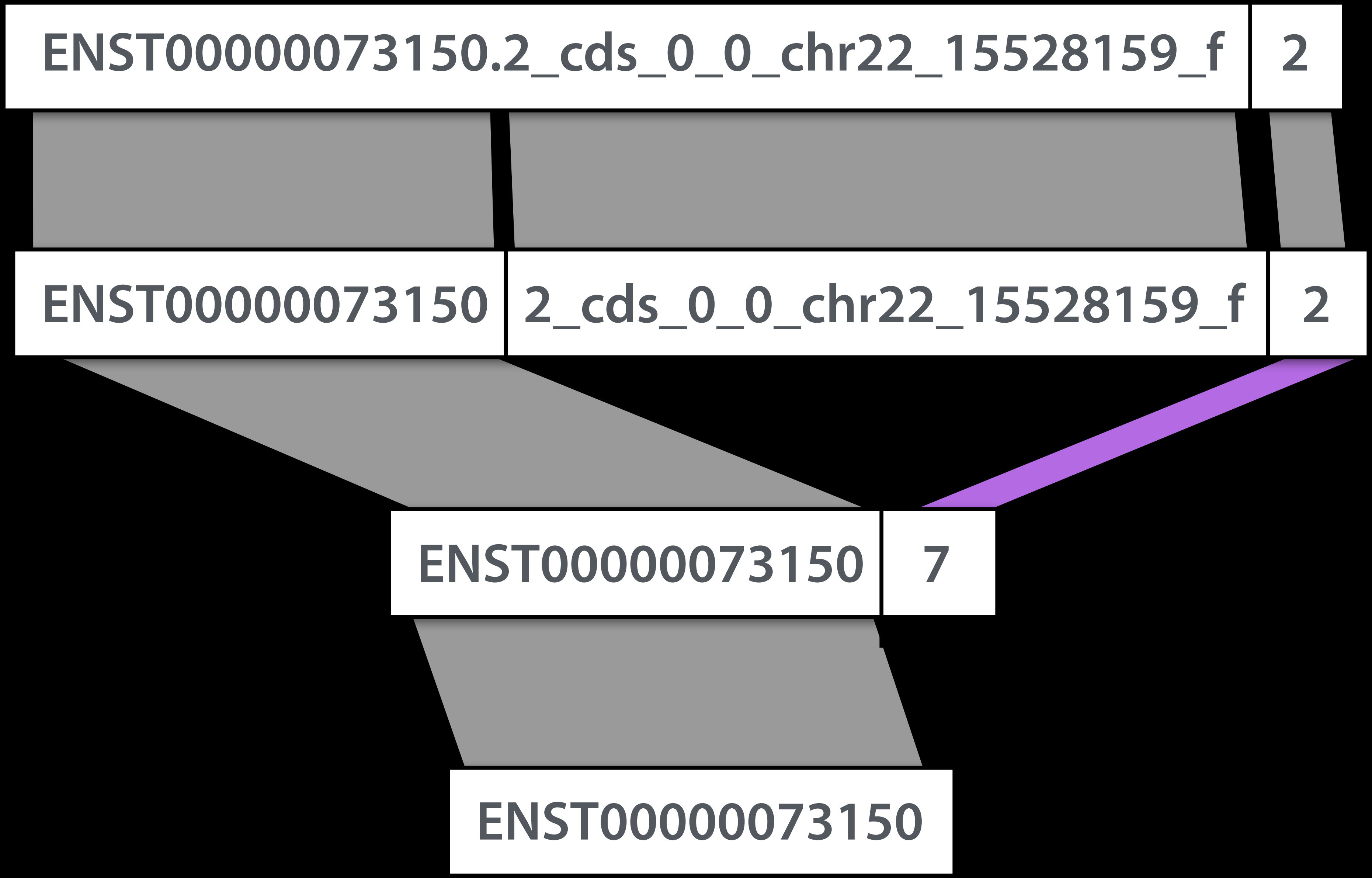
ENST00000073150.2_cds_0_0_chr22_15528159_f 2

ENST00000073150 2_cds_0_0_chr22_15528159_f 2

ENST00000073150 7

Join, Subtract and Group → Group, Sum score

Get list of transcript IDs



ENST00000073150.2_cds_0_0_chr22_15528159_f	2

ENST00000073150	2_cds_0_0_chr22_15528159_f	2

ENST00000073150	7

ENST00000073150

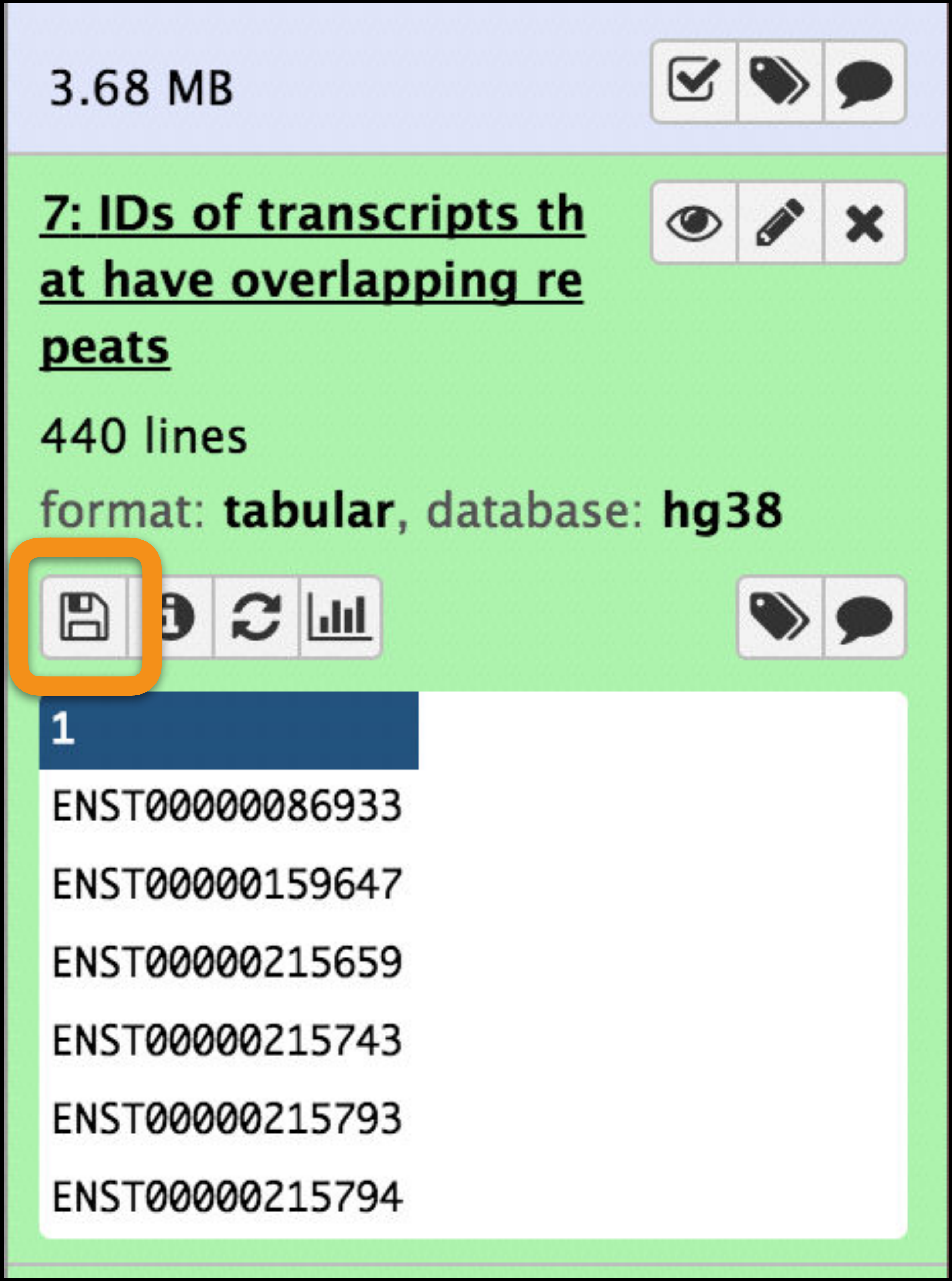
Text Manipulation → Cut

Published History: Transcripts with overlapping repeats

Have Transcripts, now get Gene IDs

Save list of Transcript IDs to a file.

We'll upload it to Ensembl BioMart



3.68 MB

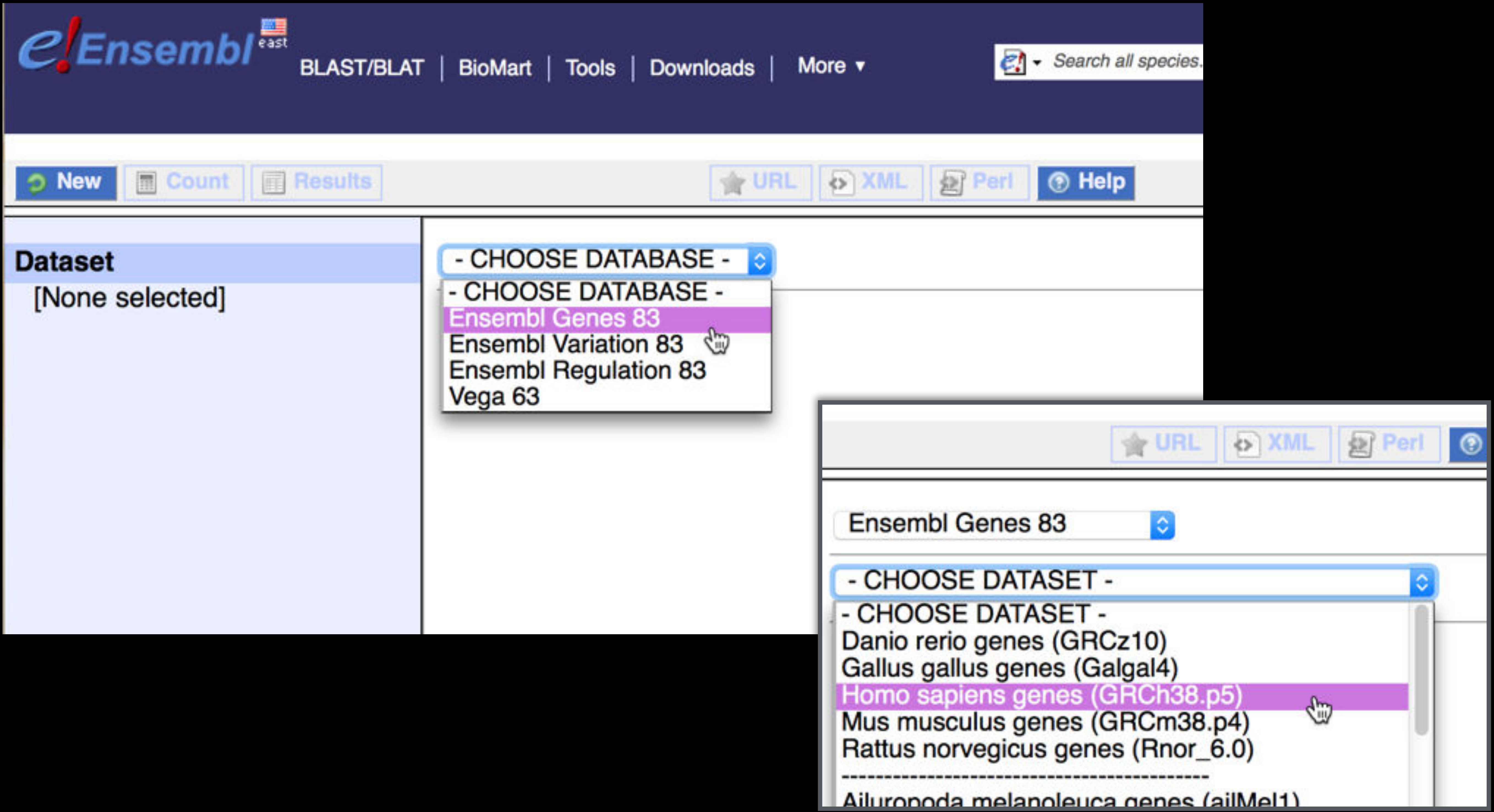
7: IDs of transcripts that have overlapping repeats

440 lines

format: **tabular**, database: **hg38**

1
ENST00000086933
ENST00000159647
ENST00000215659
ENST00000215743
ENST00000215793
ENST00000215794

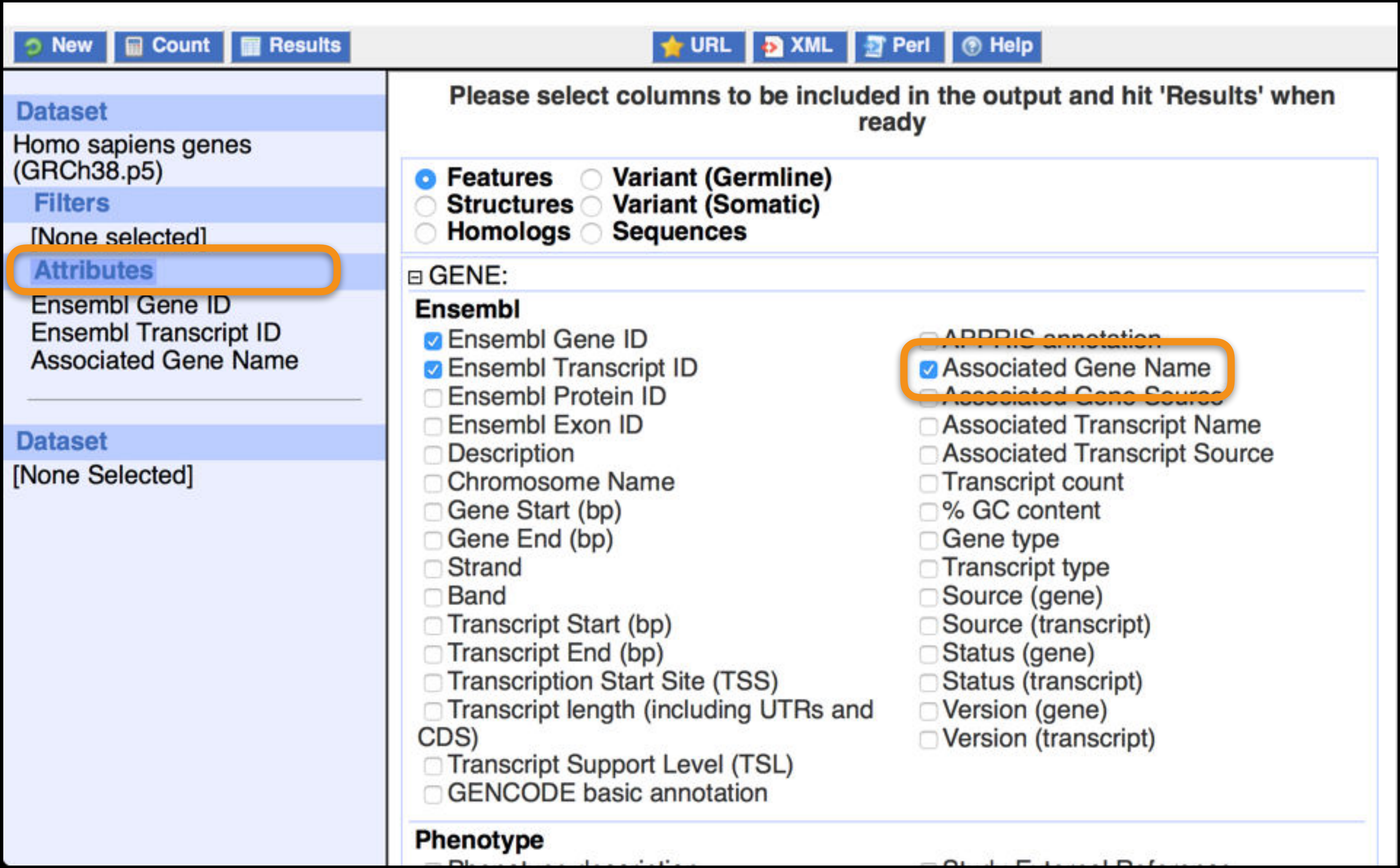
Ensembl BioMart



Specify Ensembl Genes 83, hg38

www.ensembl.org/biomart/martview

Ensembl BioMart:



Dataset

Homo sapiens genes (GRCh38.p5)

Filters

[None selected]

Attributes

Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name

Dataset

[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

● **Features** ○ **Variant (Germline)**
○ **Structures** ○ **Variant (Somatic)**
○ **Homologs** ○ **Sequences**

⊟ GENE:

Ensembl

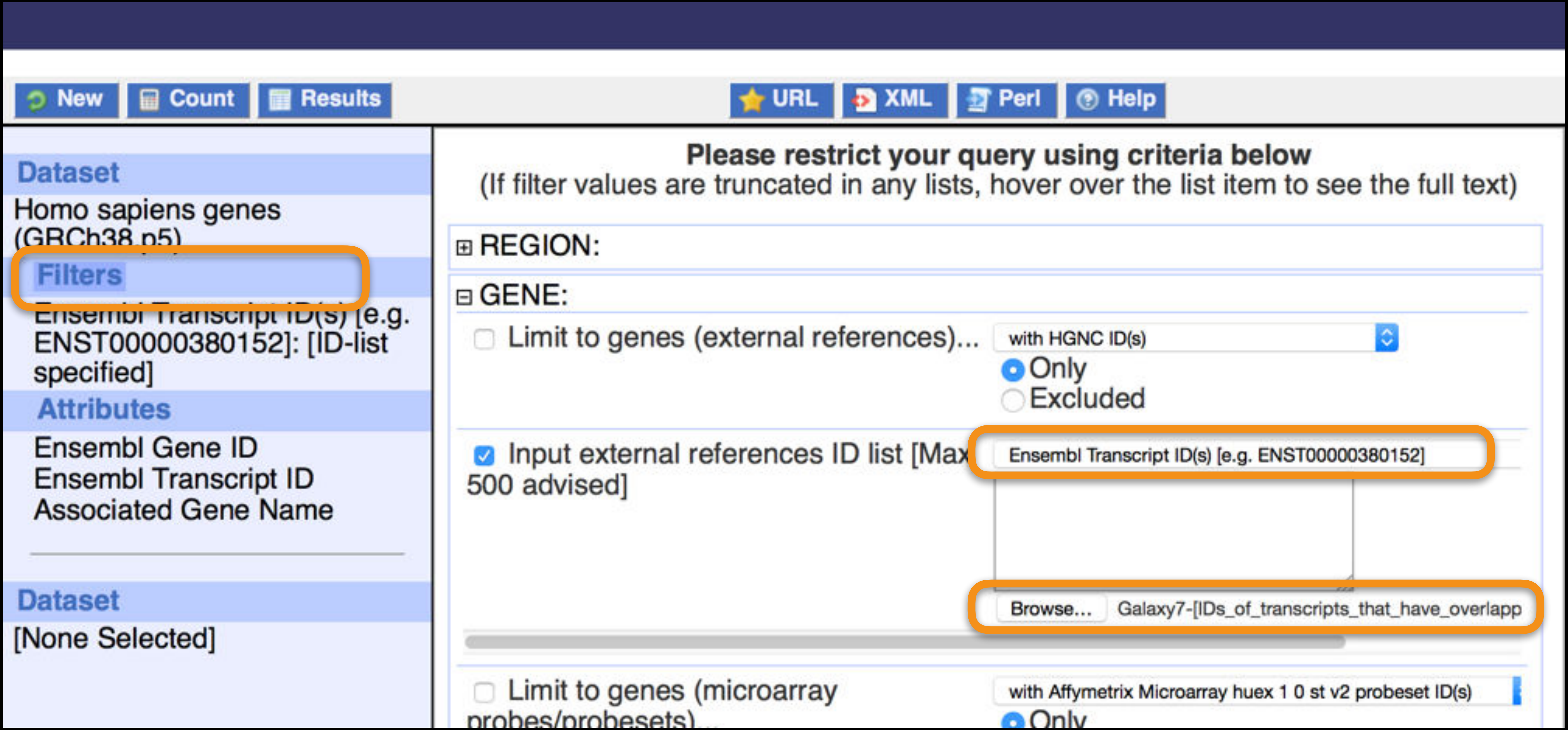
- ☑ Ensembl Gene ID
- ☑ Ensembl Transcript ID
- □ Ensembl Protein ID
- □ Ensembl Exon ID
- □ Description
- □ Chromosome Name
- □ Gene Start (bp)
- □ Gene End (bp)
- □ Strand
- □ Band
- □ Transcript Start (bp)
- □ Transcript End (bp)
- □ Transcription Start Site (TSS)
- □ Transcript length (including UTRs and CDS)
- □ Transcript Support Level (TSL)
- □ GENCODE basic annotation
- □ APPRIS annotation
- ☑ Associated Gene Name
- □ Associated Gene Source
- □ Associated Transcript Name
- □ Associated Transcript Source
- □ Transcript count
- □ % GC content
- □ Gene type
- □ Transcript type
- □ Source (gene)
- □ Source (transcript)
- □ Status (gene)
- □ Status (transcript)
- □ Version (gene)
- □ Version (transcript)

Phenotype

- □ Phenotype description
- □ Study External Reference

Specify attributes to put in output report

Ensembl BioMart:



Dataset
Homo sapiens genes (GRCh38.p5)

Filters
Ensembl Transcript ID(s) [e.g. ENST00000380152]: [ID-list specified]

Attributes
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name

Dataset
[None Selected]

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)

⊞ REGION:

⊟ GENE:

□ Limit to genes (external references)... with HGNC ID(s)
● Only
○ Excluded

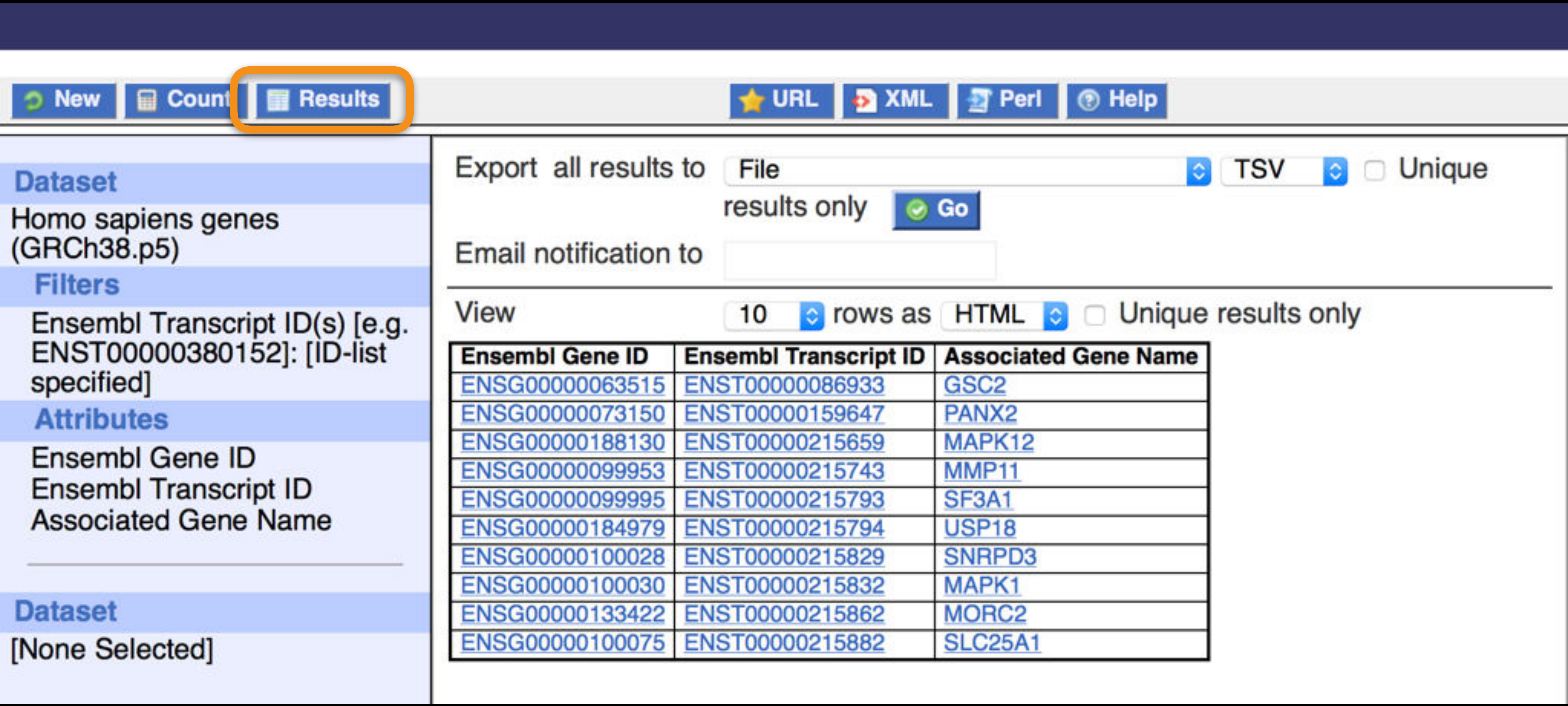
☑ Input external references ID list [Max 500 advised] Ensembl Transcript ID(s) [e.g. ENST00000380152]

Browse... Galaxy7-[IDs_of_transcripts_that_have_overlapp

□ Limit to genes (microarray probes/probesets)... with Affymetrix Microarray huex 1 0 st v2 probeset ID(s)
● Only

Specify which genes we want to this information for

Ensembl BioMart:



Dataset

Homo sapiens genes (GRCh38.p5)

Filters

Ensembl Transcript ID(s) [e.g. ENST00000380152]: [ID-list specified]

Attributes

Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name

Dataset

[None Selected]

Export all results to File TSV □ Unique results only

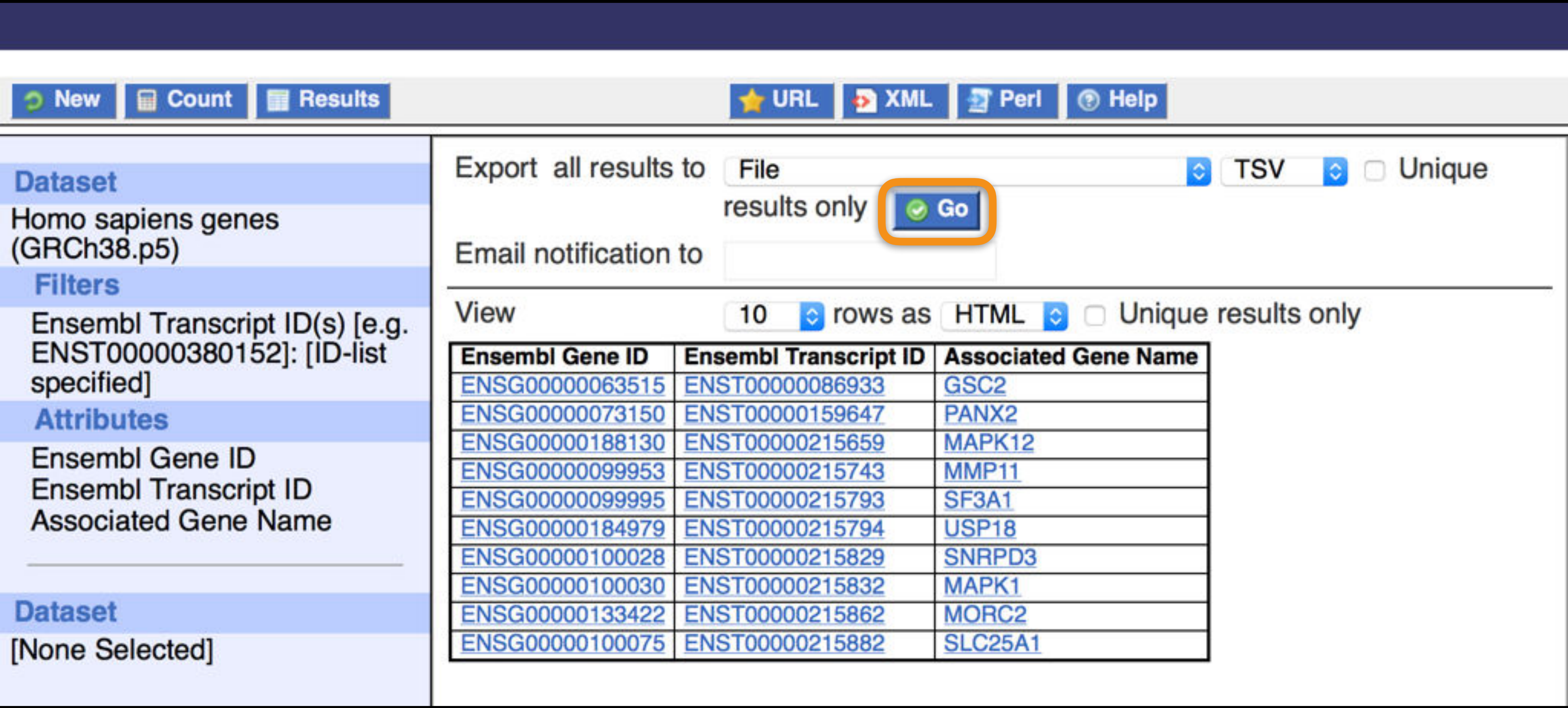
Email notification to

View 10 rows as HTML □ Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name
ENSG00000063515	ENST00000086933	GSC2
ENSG00000073150	ENST00000159647	PANX2
ENSG00000188130	ENST00000215659	MAPK12
ENSG00000099953	ENST00000215743	MMP11
ENSG00000099995	ENST00000215793	SF3A1
ENSG00000184979	ENST00000215794	USP18
ENSG00000100028	ENST00000215829	SNRPD3
ENSG00000100030	ENST00000215832	MAPK1
ENSG00000133422	ENST00000215862	MORC2
ENSG00000100075	ENST00000215882	SLC25A1

See the report

Ensembl BioMart:



Dataset

Homo sapiens genes (GRCh38.p5)

Filters

Ensembl Transcript ID(s) [e.g. ENST00000380152]: [ID-list specified]

Attributes

Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name

Dataset

[None Selected]

Export all results to File TSV □ Unique results only

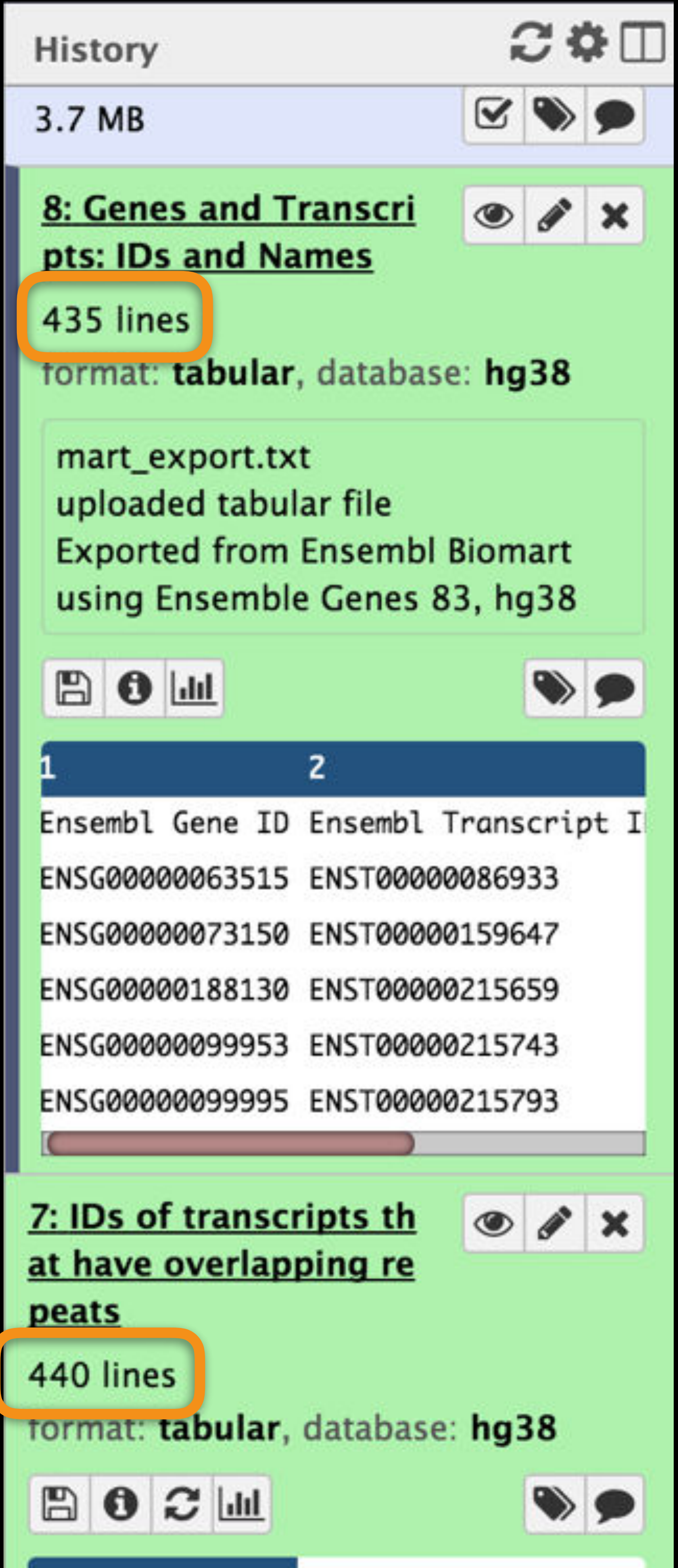
Email notification to

View 10 rows as HTML □ Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name
ENSG00000063515	ENST00000086933	GSC2
ENSG00000073150	ENST00000159647	PANX2
ENSG00000188130	ENST00000215659	MAPK12
ENSG00000099953	ENST00000215743	MMP11
ENSG00000099995	ENST00000215793	SF3A1
ENSG00000184979	ENST00000215794	USP18
ENSG00000100028	ENST00000215829	SNRPD3
ENSG00000100030	ENST00000215832	MAPK1
ENSG00000133422	ENST00000215862	MORC2
ENSG00000100075	ENST00000215882	SLC25A1

Download the data

Get Gene IDs into Galaxy



History

3.7 MB

8: Genes and Transcripts: IDs and Names

435 lines

format: **tabular**, database: **hg38**

mart_export.txt
uploaded tabular file
Exported from Ensembl Biomart
using Ensemble Genes 83, hg38

1	2
Ensembl Gene ID	Ensembl Transcript I
ENSG00000063515	ENST00000086933
ENSG00000073150	ENST00000159647
ENSG00000188130	ENST00000215659
ENSG00000099953	ENST00000215743
ENSG00000099995	ENST00000215793

7: IDs of transcripts that have overlapping repeats

440 lines

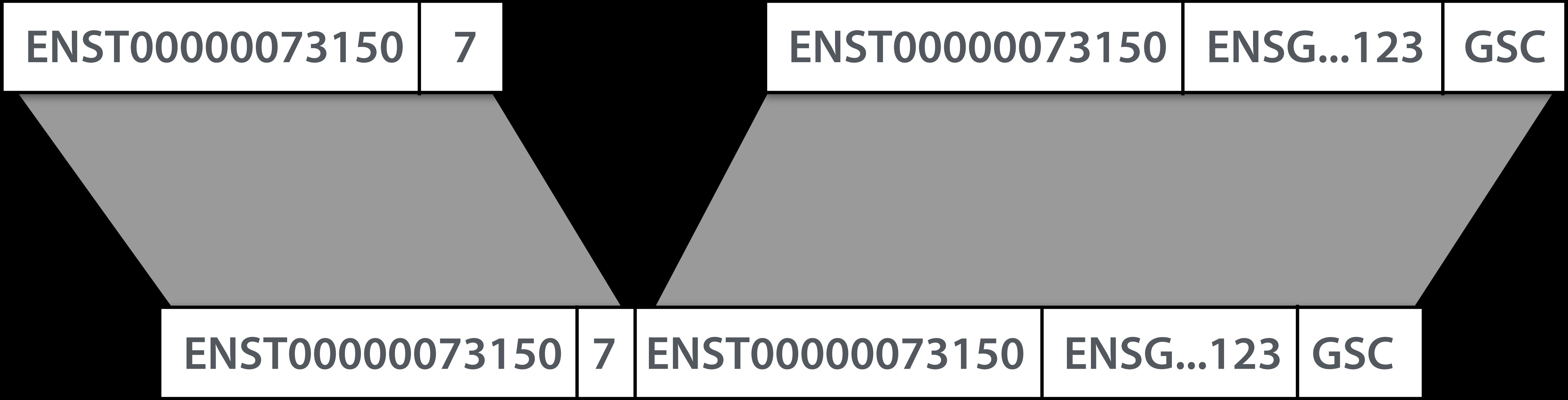
format: **tabular**, database: **hg38**

Upload the file from BioMart.
Note that we lost 5-6 transcripts

Unite our Transcript Scores with Biomart info

Transcript Scores

Biomart Info



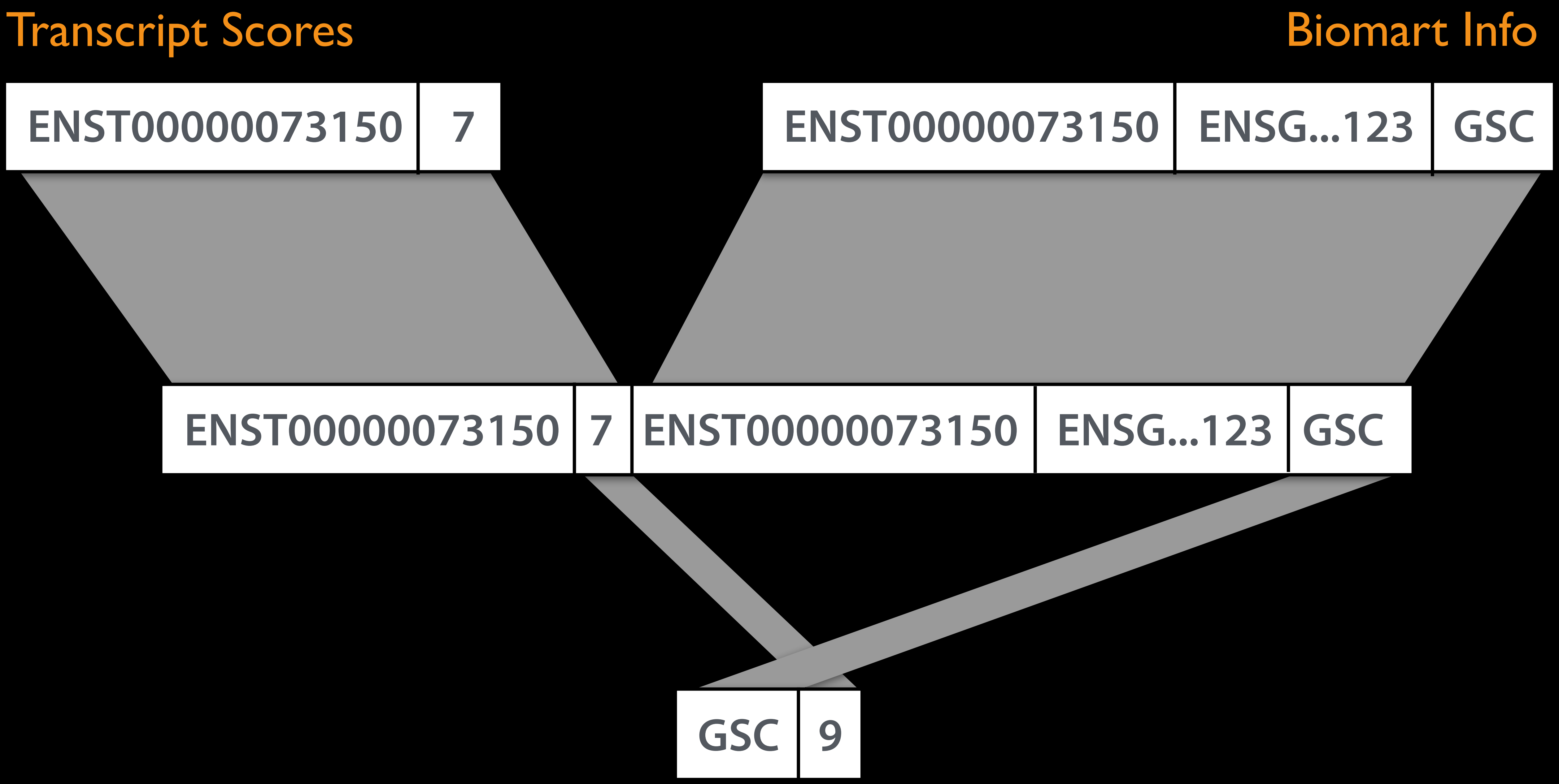
ENST00000073150	7

ENST00000073150	ENSG...123	GSC

ENST00000073150	7	ENST00000073150	ENSG...123	GSC

Join, Subtract and Group → Join

Assign scores to genes



Transcript Scores

ENST00000073150 | 7

Biomart Info

ENST00000073150 | ENSG...123 | GSC

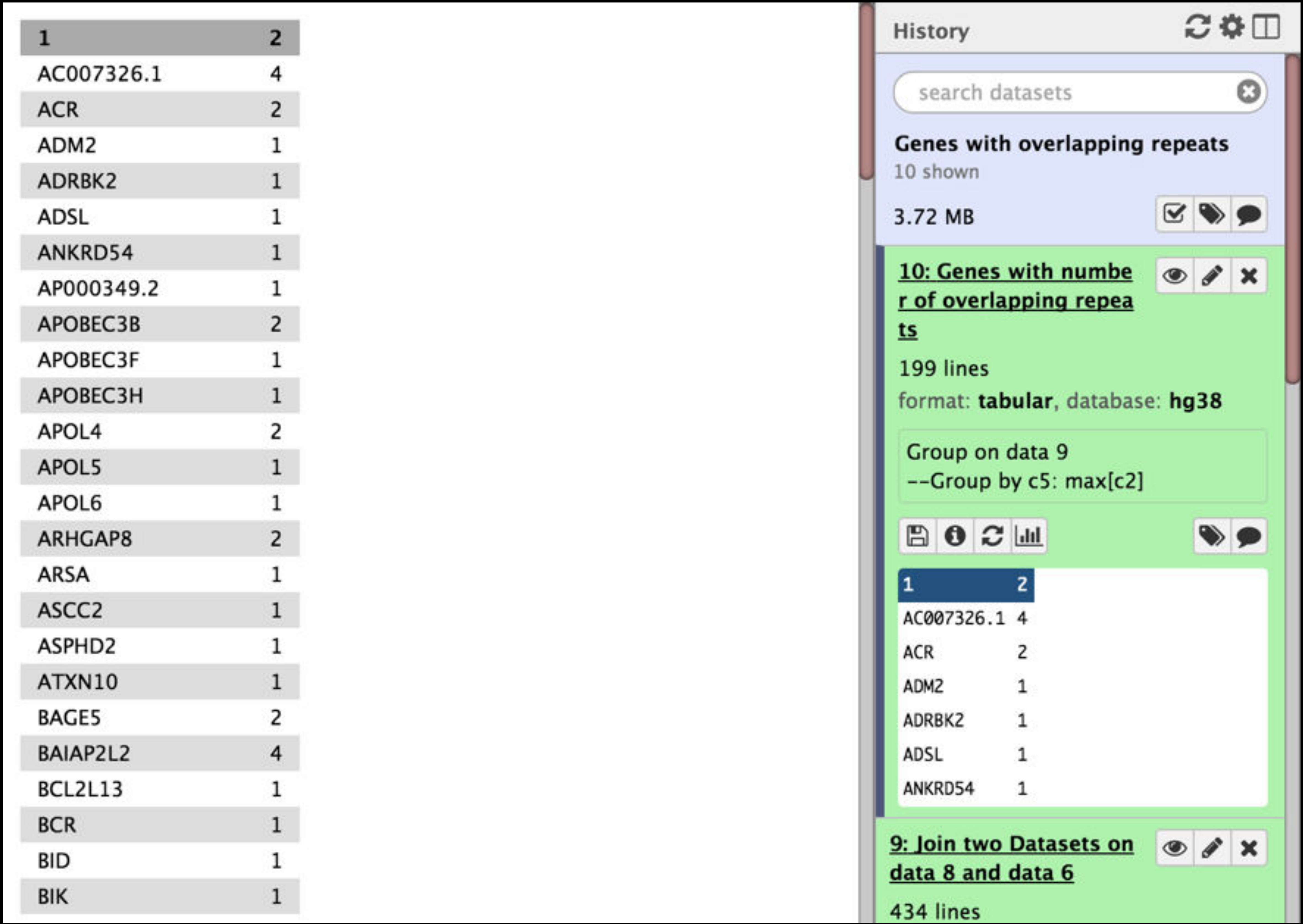
ENST00000073150 | 7 | ENST00000073150 | ENSG...123 | GSC

GSC | 9

Join, Subtract and Group → Group, Max

Published History: Genes with overlapping repeats

Now have a list of genes with # overlapping repeats



1	2
AC007326.1	4
ACR	2
ADM2	1
ADRBK2	1
ADSL	1
ANKRD54	1
AP000349.2	1
APOBEC3B	2
APOBEC3F	1
APOBEC3H	1
APOL4	2
APOL5	1
APOL6	1
ARHGAP8	2
ARSA	1
ASCC2	1
ASPHD2	1
ATXN10	1
BAGE5	2
BAIAP2L2	4
BCL2L13	1
BCR	1
BID	1
BIK	1

Genes with overlapping repeats

10 shown

3.72 MB

10: Genes with number of overlapping repeats

199 lines

format: **tabular**, database: **hg38**

Group on data 9
--Group by c5: max[c2]

1	2
AC007326.1	4
ACR	2
ADM2	1
ADRBK2	1
ADSL	1
ANKRD54	1

9: Join two Datasets on data 8 and data 6

434 lines

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

http://bit.ly/gxyismms2016

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

http://bit.ly/gxyismms2016

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
  - And then rolling that up to genes
- But, ...
  - is there anything inherent in the analysis about humans, exons or repeats?

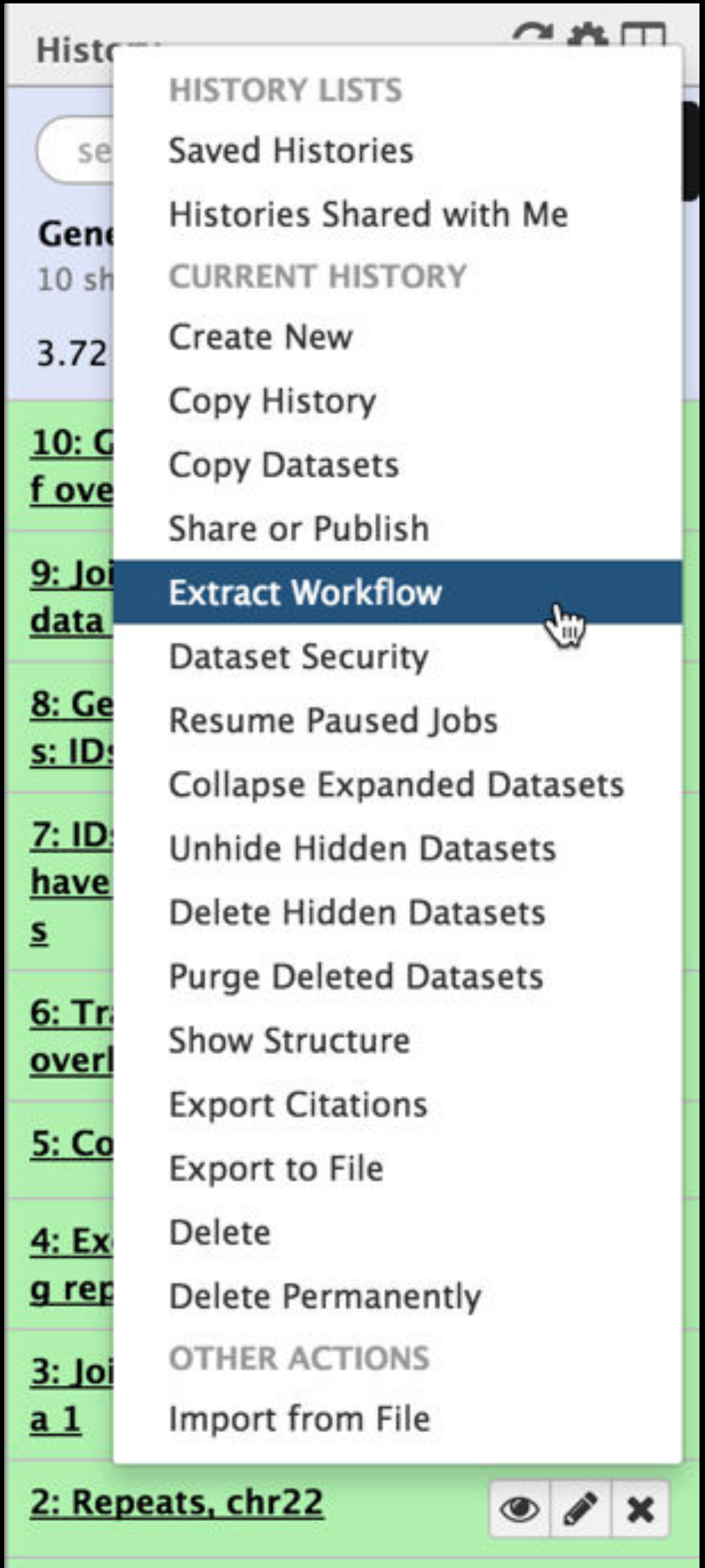
Create a Workflow from a History

Extract Workflow from history

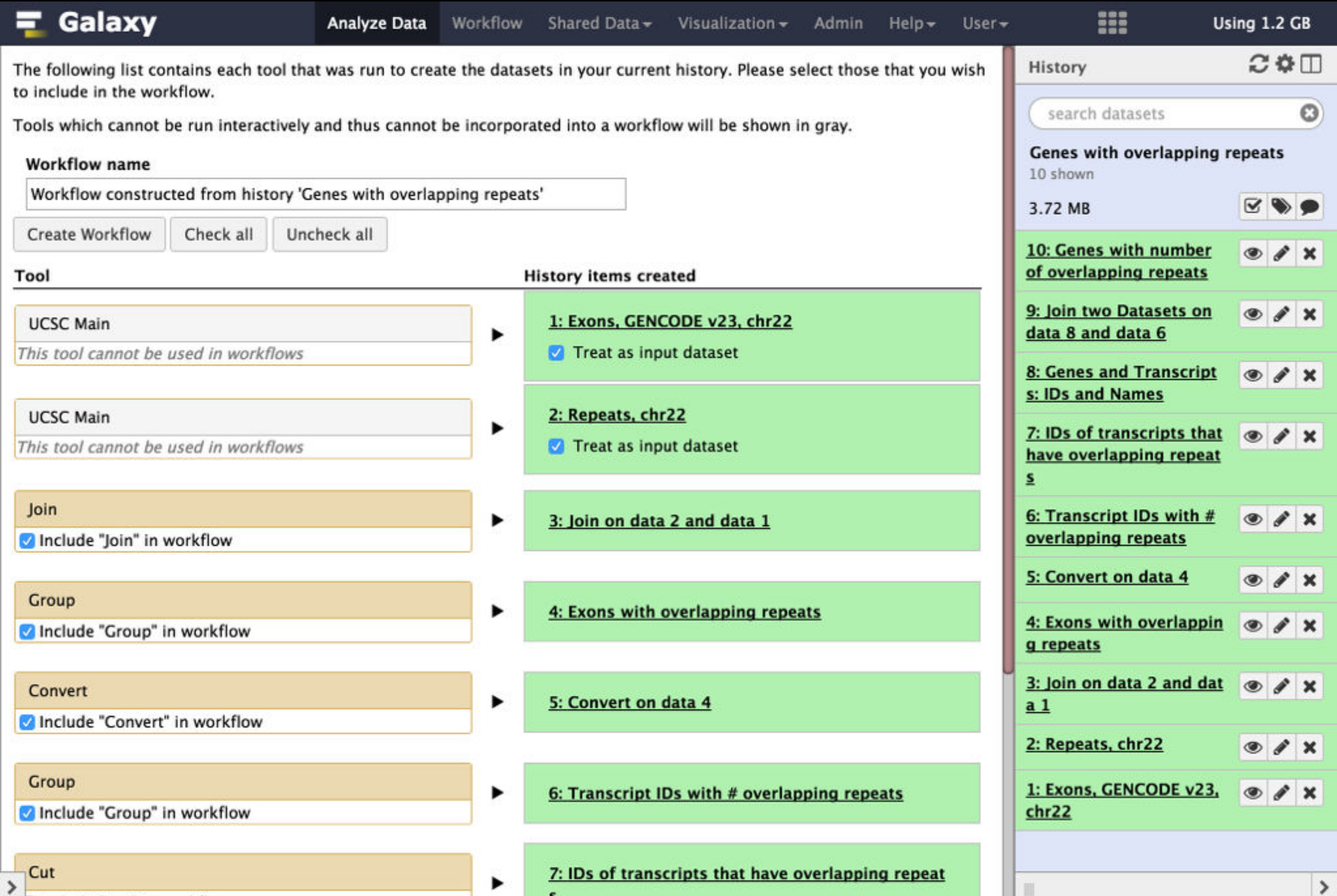
Create a workflow from this history.

Edit it to make some things clearer.

⚙ (cog) → Extract Workflow



Create a Workflow from a History: ...



The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name

Workflow constructed from history 'Genes with overlapping repeats'

Tool	History items created
UCSC Main *This tool cannot be used in workflows*	1: Exons, GENCODE v23, chr22 ☑ Treat as input dataset
UCSC Main *This tool cannot be used in workflows*	2: Repeats, chr22 ☑ Treat as input dataset
Join ☑ Include "Join" in workflow	3: Join on data 2 and data 1
Group ☑ Include "Group" in workflow	4: Exons with overlapping repeats
Convert ☑ Include "Convert" in workflow	5: Convert on data 4
Group ☑ Include "Group" in workflow	6: Transcript IDs with # overlapping repeats
Cut	7: IDs of transcripts that have overlapping repeat

History

search datasets

Genes with overlapping repeats

10 shown

3.72 MB

- 10: Genes with number of overlapping repeats
- 9: Join two Datasets on data 8 and data 6
- 8: Genes and Transcripts: IDs and Names
- 7: IDs of transcripts that have overlapping repeats
- 6: Transcript IDs with # overlapping repeats
- 5: Convert on data 4
- 4: Exons with overlapping repeats
- 3: Join on data 2 and data 1
- 2: Repeats, chr22
- 1: Exons, GENCODE v23, chr22

Wait ...

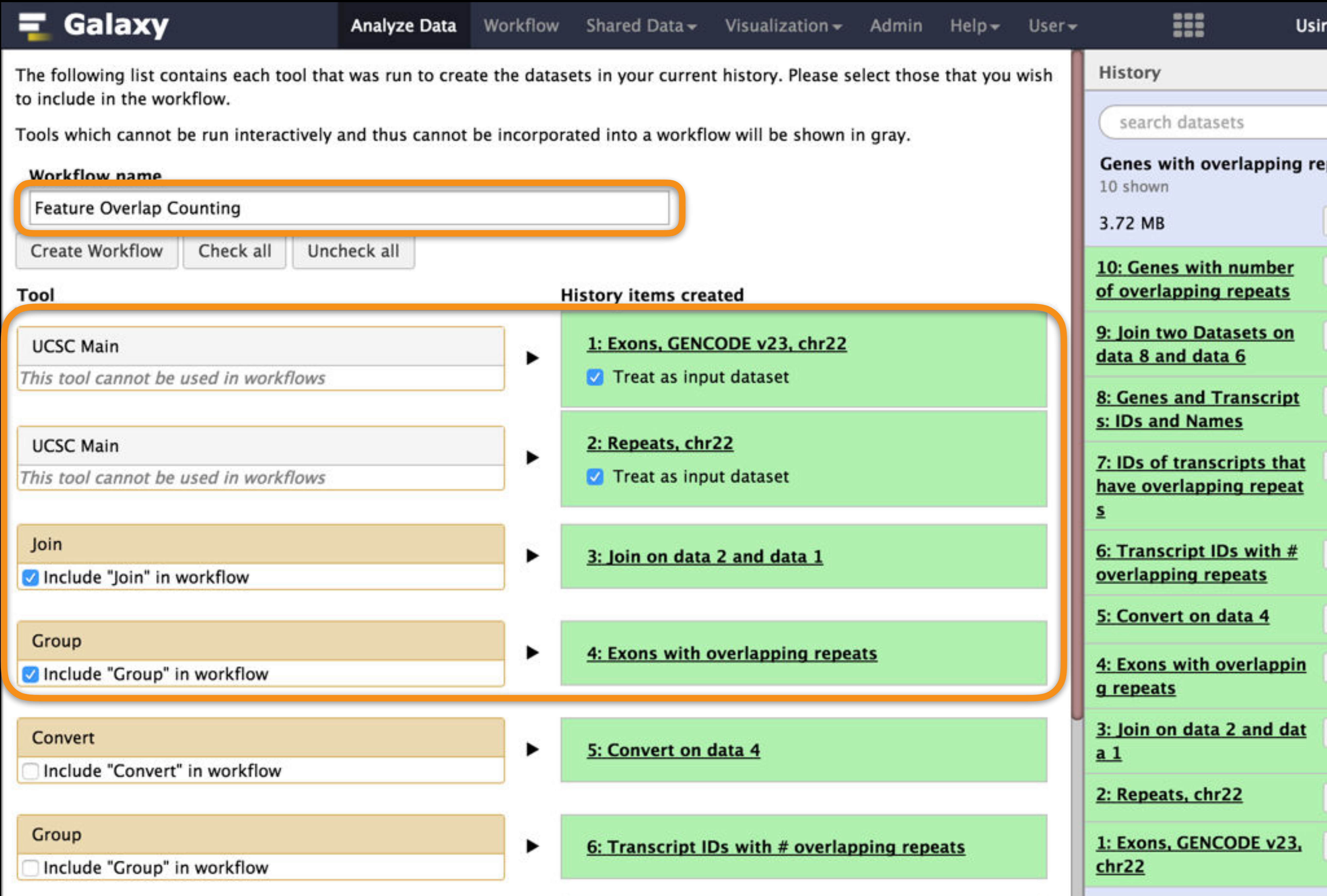
Can this whole analysis be a useful workflow?
(No.)

Are there parts of this analysis are a good candidate for a workflow - something to be reused on other data?

Steps 5 and 6 extract a Transcript ID from a UCSC encoded Exon name.
Not clean, and not widely useful

The first 4 items count overlaps between features.
That might be useful.

Create a Workflow from a History: ...



The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name

Feature Overlap Counting

Create Workflow | Check all | Uncheck all

Tool		History items created
UCSC Main *This tool cannot be used in workflows*	▶	1: Exons, GENCODE v23, chr22 ☑ Treat as input dataset
UCSC Main *This tool cannot be used in workflows*	▶	2: Repeats, chr22 ☑ Treat as input dataset
Join ☑ Include "Join" in workflow	▶	3: Join on data 2 and data 1
Group ☑ Include "Group" in workflow	▶	4: Exons with overlapping repeats
Convert ☐ Include "Convert" in workflow	▶	5: Convert on data 4
Group ☐ Include "Group" in workflow	▶	6: Transcript IDs with # overlapping repeats

History

search datasets

Genes with overlapping re
10 shown

3.72 MB

10: Genes with number of overlapping repeats

9: Join two Datasets on data 8 and data 6

8: Genes and Transcripts: IDs and Names

7: IDs of transcripts that have overlapping repeats

6: Transcript IDs with # overlapping repeats

5: Convert on data 4

4: Exons with overlapping repeats

3: Join on data 2 and data 1

2: Repeats, chr22

1: Exons, GENCODE v23, chr22

Workflows

Run / test it

Guided: rerun with same inputs

Workflow → Run

Did that work?

On your own:

Count # of exons overlapping each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions

Published Workflow: Feature Overlap Counting

Workflows: Sweet spots

Short, well-defined tasks, with well-defined inputs and outputs.

Analysis pipelines for large experiments with many samples where sample and data preparation protocols are the same throughout.

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

http://bit.ly/gxyismms2016

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

http://bit.ly/gxyismms2016

RNA-Seq Analysis: Get the Data

Create new history

⚙ (cog) → Create New

Import:

Shared Data → Data Libraries → Training → RNA-Seq*

→ UC-Davis → Raw Reads

Select first two

MeOH_REP1_R1, MeOH_REP1_R2

UCDAVIS Bioinformatics Core
Genome Center

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

NGS Data Quality Control

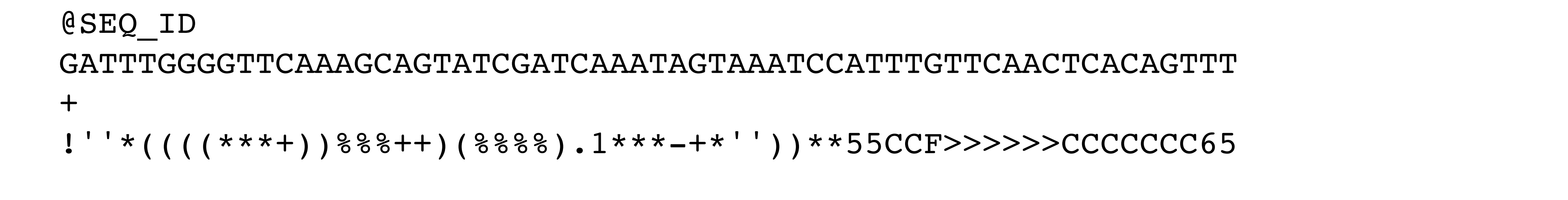
- FASTQ format
- Examine quality in an RNA-Seq dataset
- Trim/filter as we see fit, hopefully without breaking anything.

Quality Control is not sexy.

But it is vital.

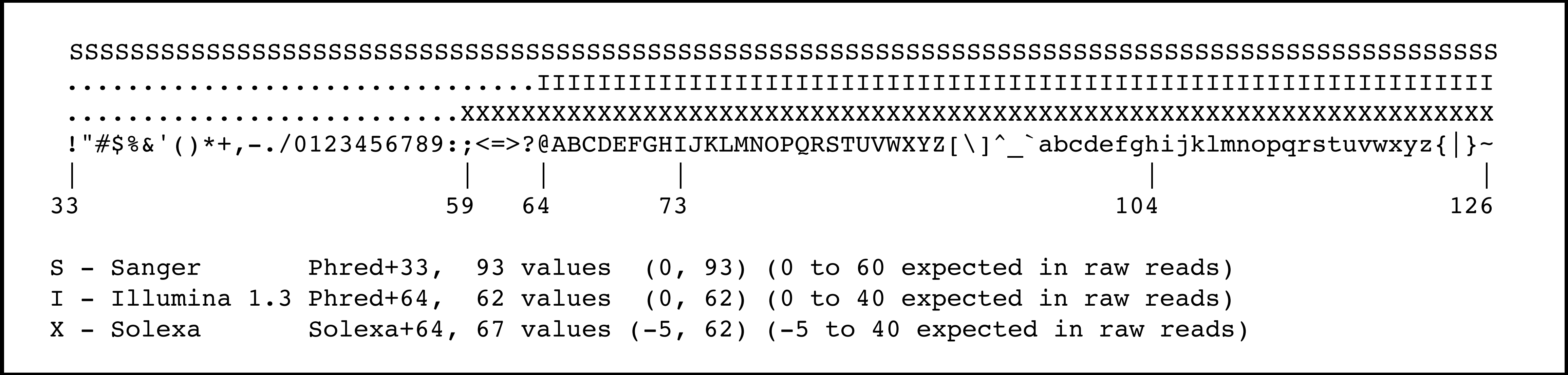
What is FASTQ?

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



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

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



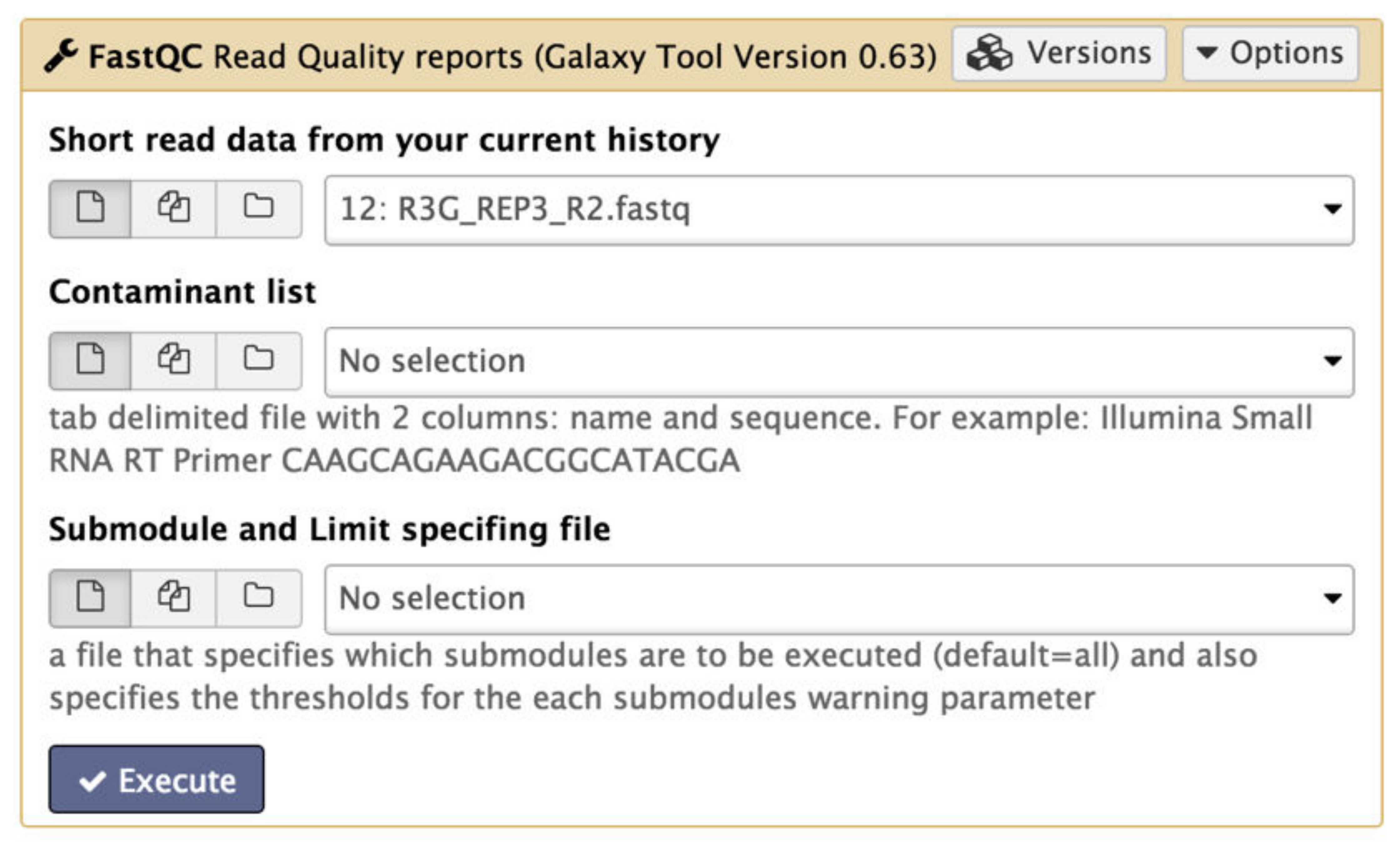
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@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: Assessment tools

NGS QC and Manipulation → FastQC

Generates summary quality information.



FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

12: R3G_REP3_R2.fastq

Contaminant list

No selection

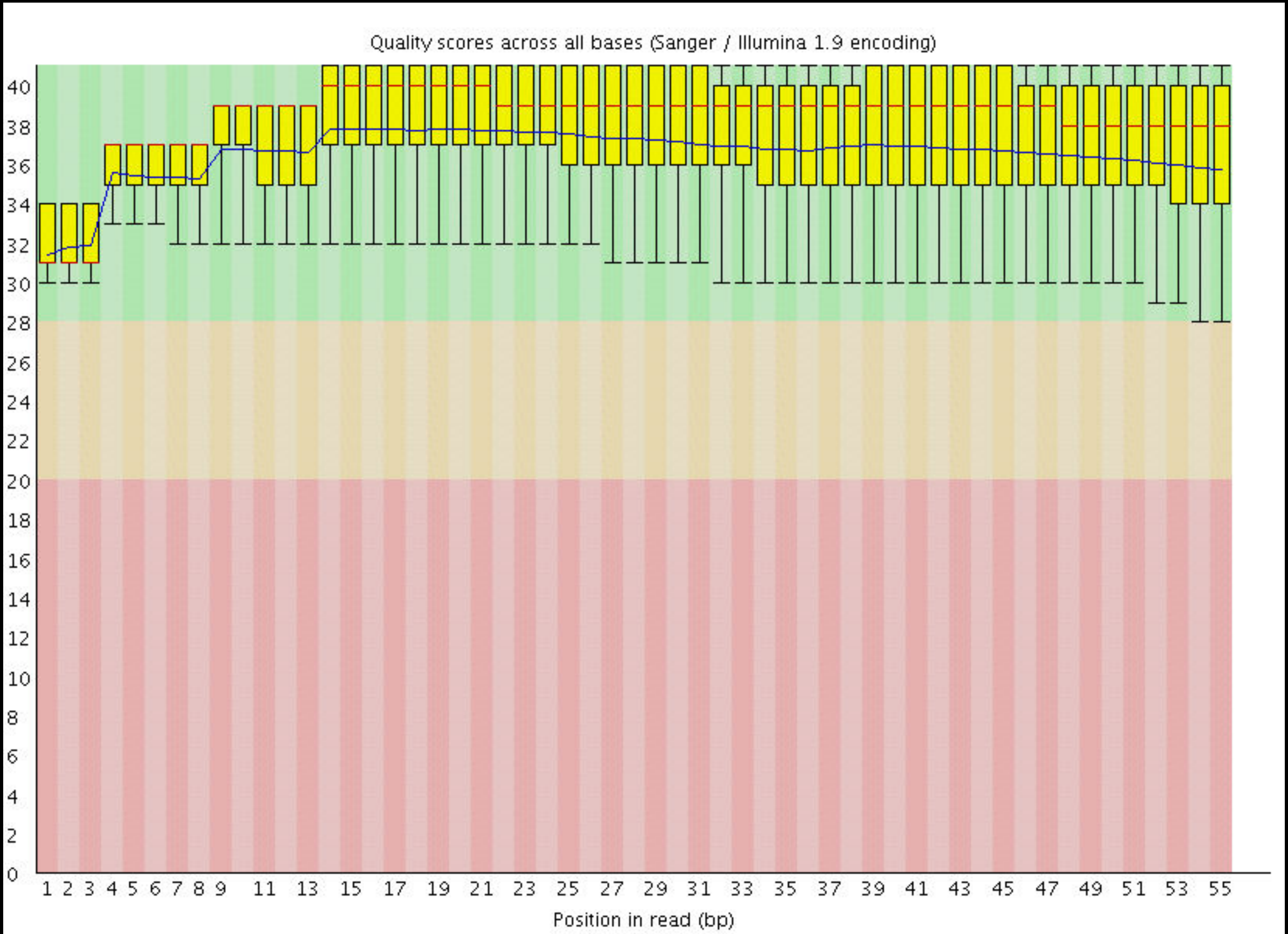
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file

No selection

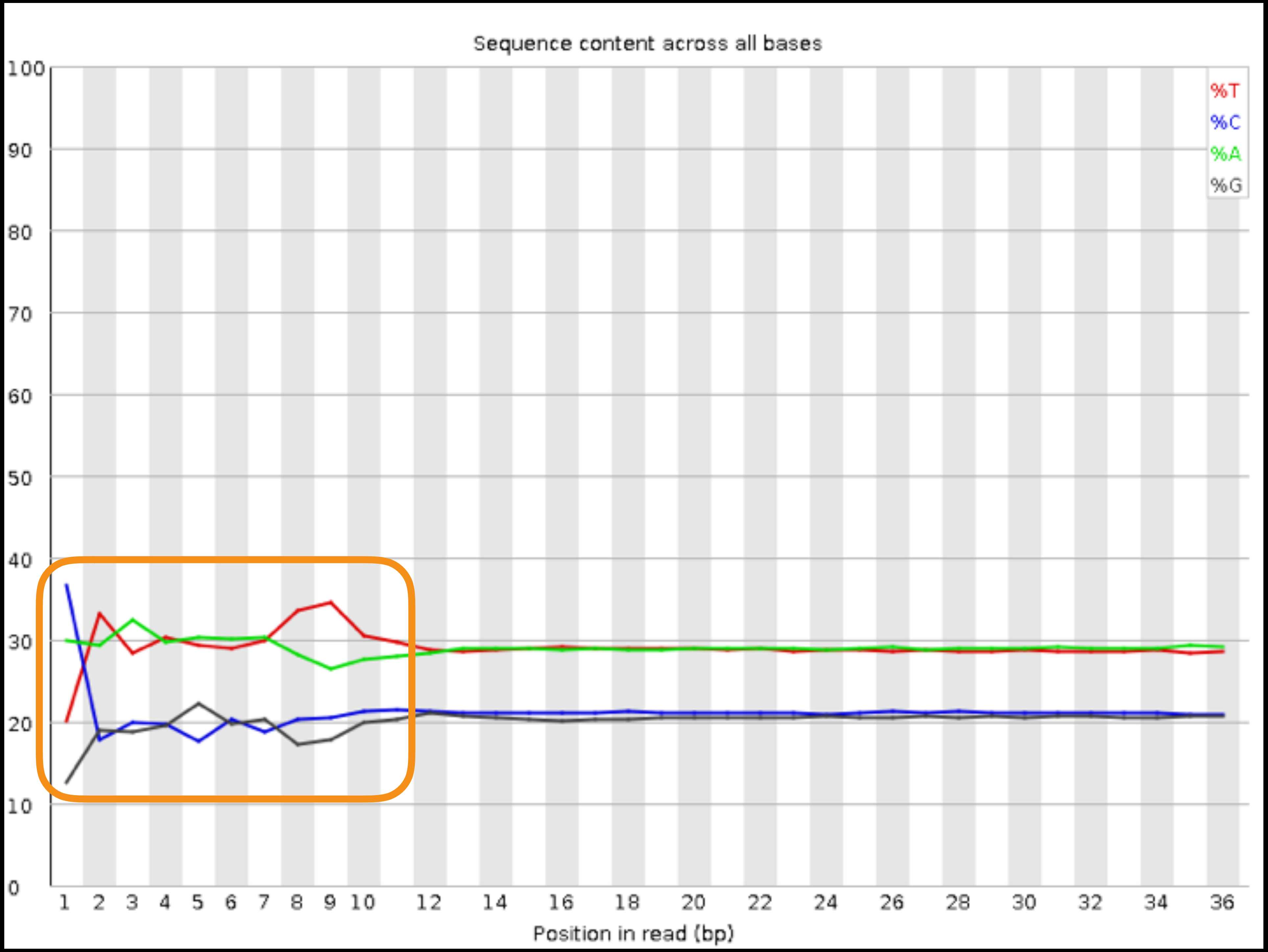
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

NGS Data Quality: Assessment tools



http://bit.ly/FastQCBoxPlot

NGS Data Quality: Sequence bias at front of reads?

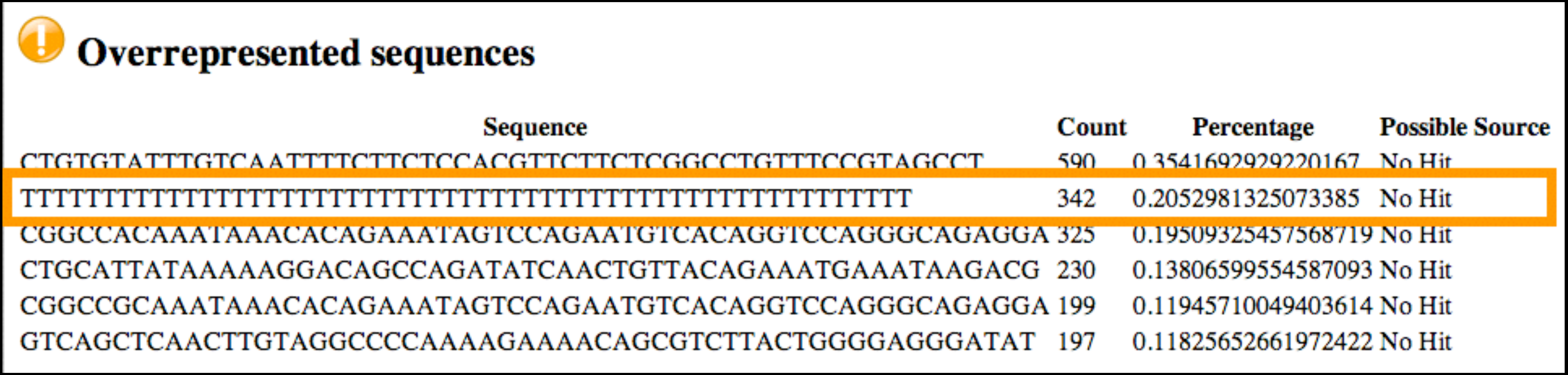


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

NGS Data Quality: Sequencing Artifacts

And only now we notice a problem with MeOH Rep1 R2 (the reverse reads)



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGTGTATTTGTCAATTTTCTTCTCCACGTTCTTCTCGGCCTGTTTCCGTAGCCT	590	0.3541692929220167	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	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. More on that in a bit.)

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

Common Trimming options

- Drop the first n columns from your reads
- Drop the last n columns from your reads
- Sliding window approach: only keep regions that are above a specified quality threshold
- Keep or drop whole read based on overall quality

Common Trimming Pitfalls

Broken Pairs

Often, one side of a pair passes QC, while the other does not.

Broken pairings can affect results in subtle or drastic ways

Short short reads.

QC may reduce reads to a length at which their mapping is no longer meaningful.

Need help with Trimming? (and anything else)

That's a whole lotta options...

Choices you make now have impact on downstream tools

NGS = a whole lotta options in general

What to do?

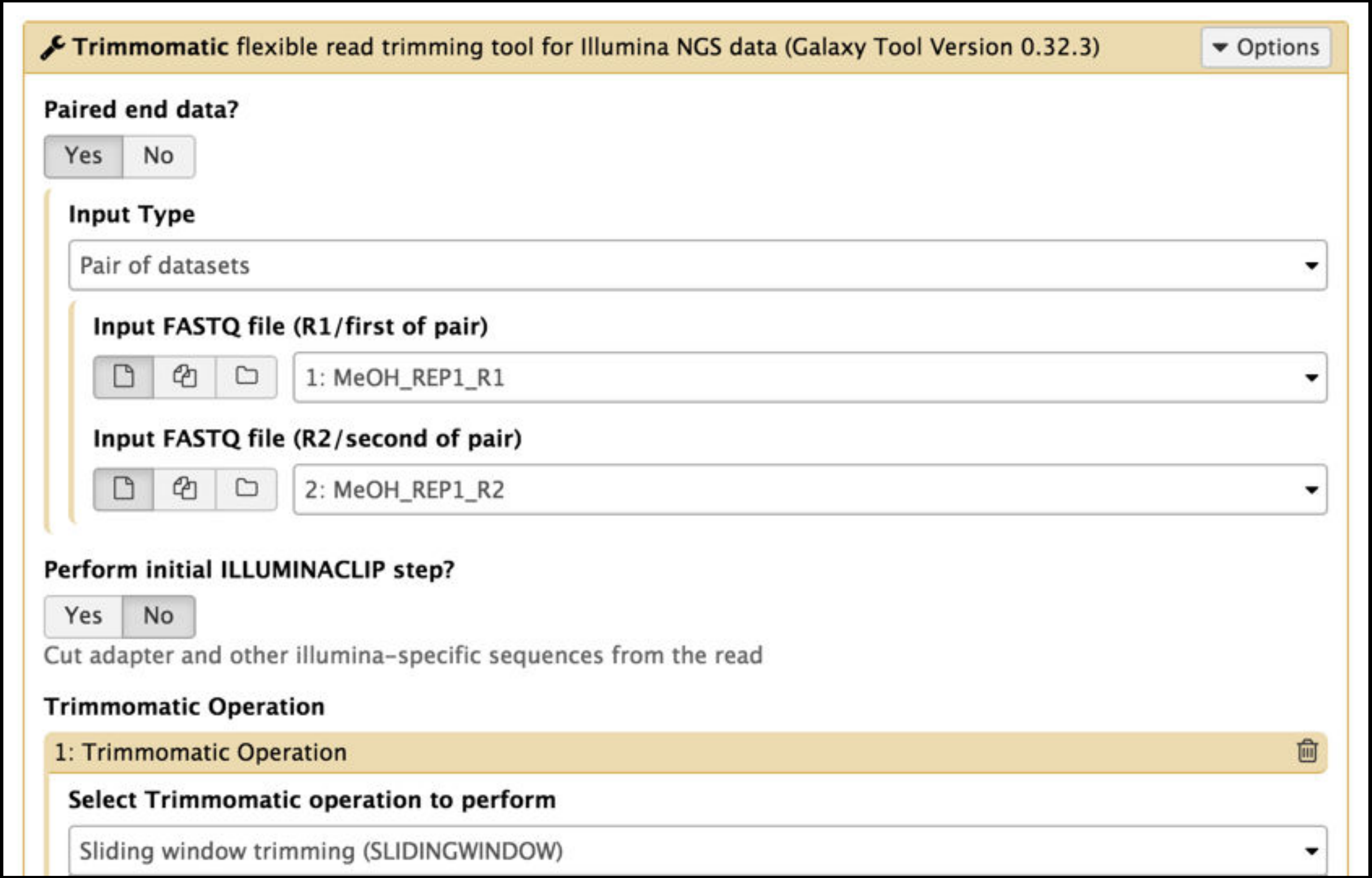
How to better understand bioinformatics & Galaxy

- Experiment. (You are already used to the idea and) Galaxy makes it easy
- Read tool documentation and tool and method review papers
- Get Help!
  - http://biostars.org/
  - http://seqanswers.com/
  - https://biostar.usegalaxy.org/
  - http://galaxyproject.org/search

SEQanswers
the next generation sequencing community

Biostar
BIOINFORMATICS EXPLAINED

Trimmomatic to the rescue



Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Tool Version 0.32.3)

Paired end data?

[✓ Yes] No

Input Type

Pair of datasets

Input FASTQ file (R1/first of pair)

1: MeOH_REP1_R1

Input FASTQ file (R2/second of pair)

2: MeOH_REP1_R2

Perform initial ILLUMINACLIP step?

Yes [✓ No]

Cut adapter and other illumina-specific sequences from the read

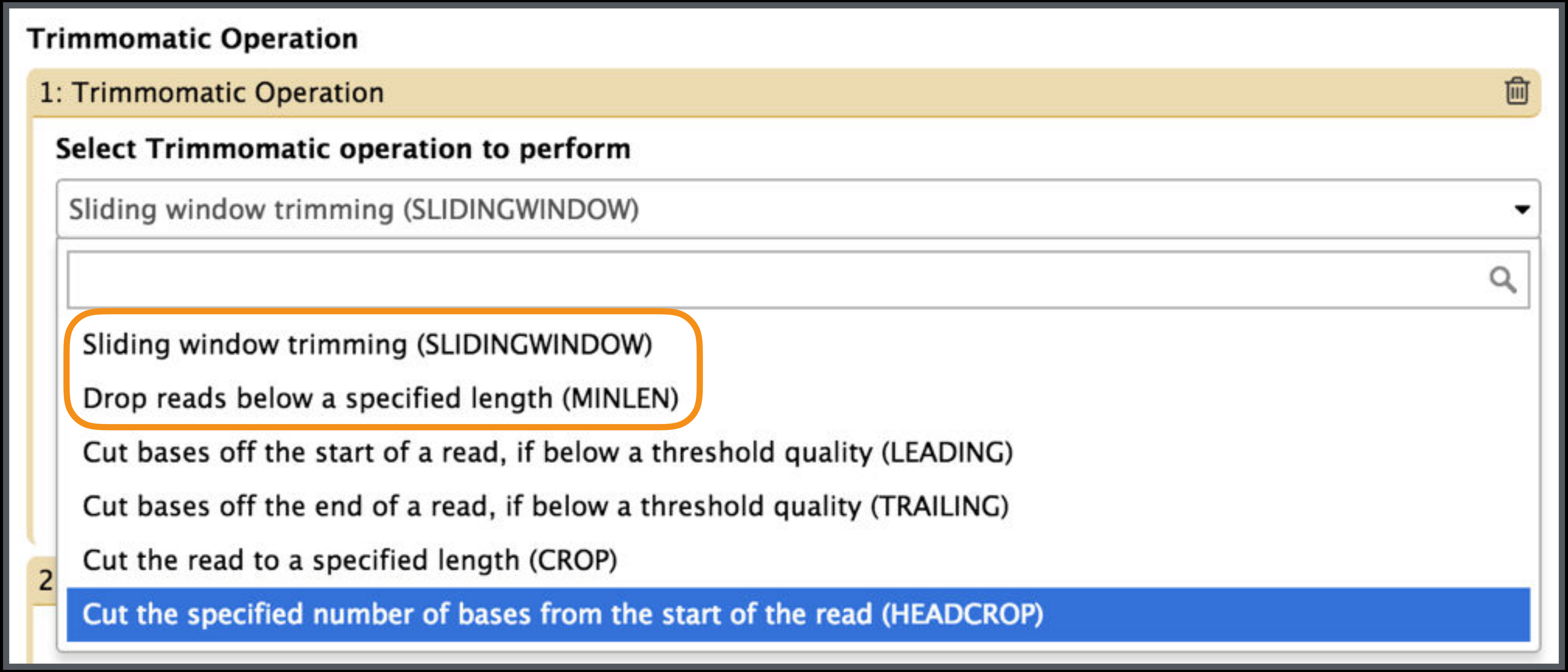
Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Bolger, A.M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, doi: 10.1093/bioinformatics/btu170



Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Sliding window trimming (SLIDINGWINDOW)

Drop reads below a specified length (MINLEN)

Cut bases off the start of a read, if below a threshold quality (LEADING)

Cut bases off the end of a read, if below a threshold quality (TRAILING)

Cut the read to a specified length (CROP)

Cut the specified number of bases from the start of the read (HEADCROP)

Trimmomatic preserves read pairing

Multiple filters can be run in arbitrary order

We'll use sliding window, followed by minimum length.

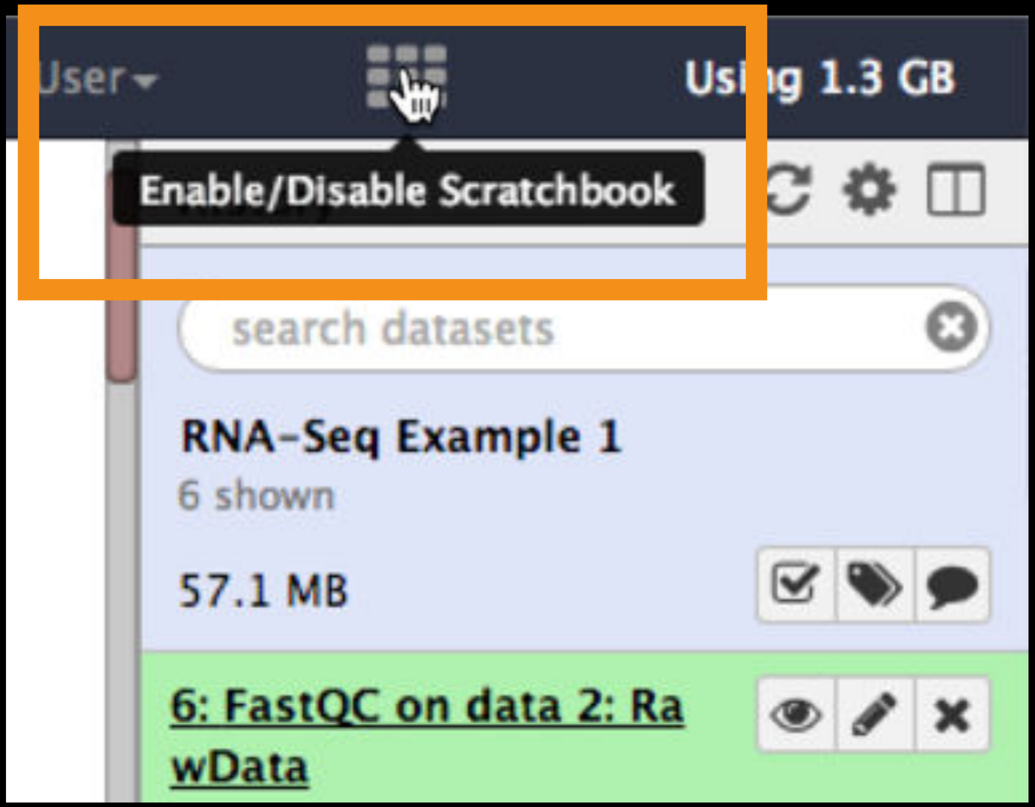
Run FastQC on post-Trimmatic Datasets

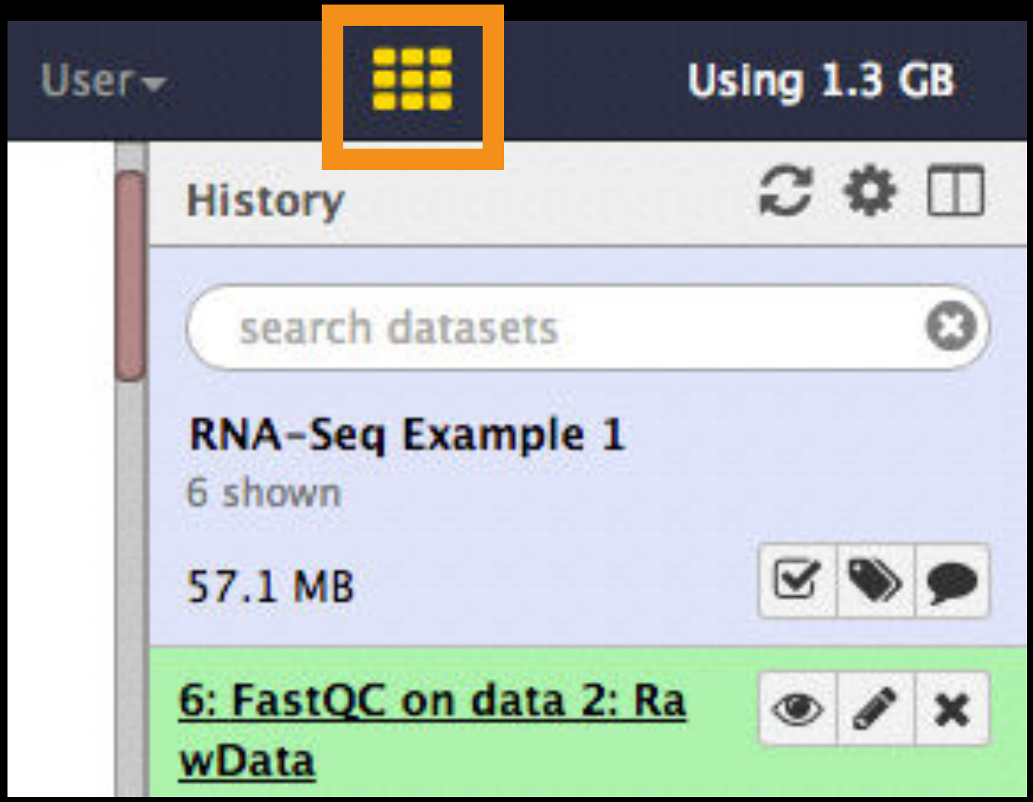
NGS QC and Manipulation → FastQC

Now, let's see what changed

Shared History: RNA-Seq MeOH_REP1 through QC

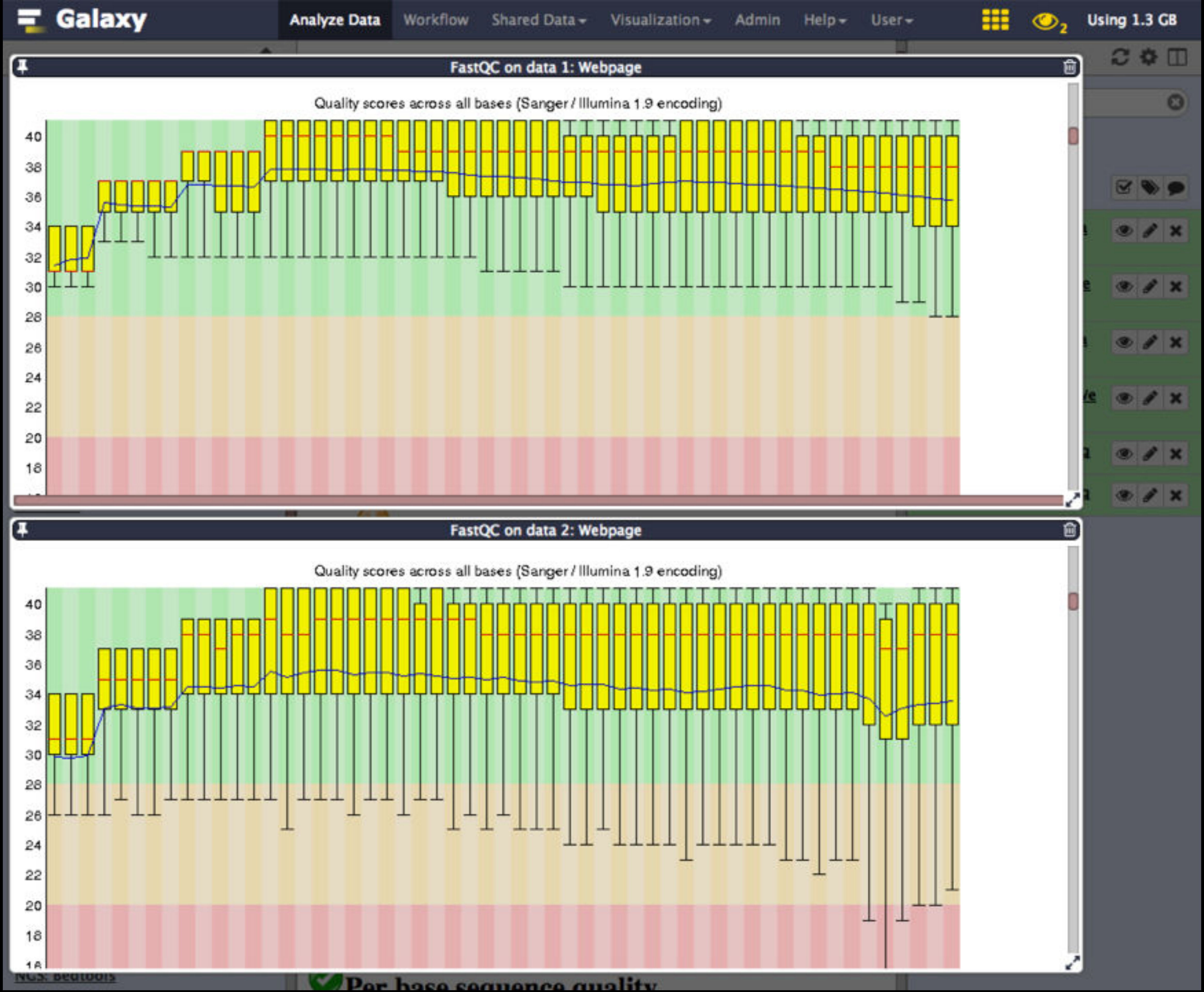
Scratchbook: View multiple datasets





And the icon turns **yellow**!

Poke the pre-Trimmomatic reverse read FastQC report in the eye, and then poke the post-Trimmomatic FastQC report in the eye.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 1.3 GB
FastQC on data 1: Webpage
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
FastQC on data 2: Webpage
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
NGS: Bedtools
Per base sequence quality

And after some resizing and scrolling you see this

NGS Data Quality Assessment: Done!

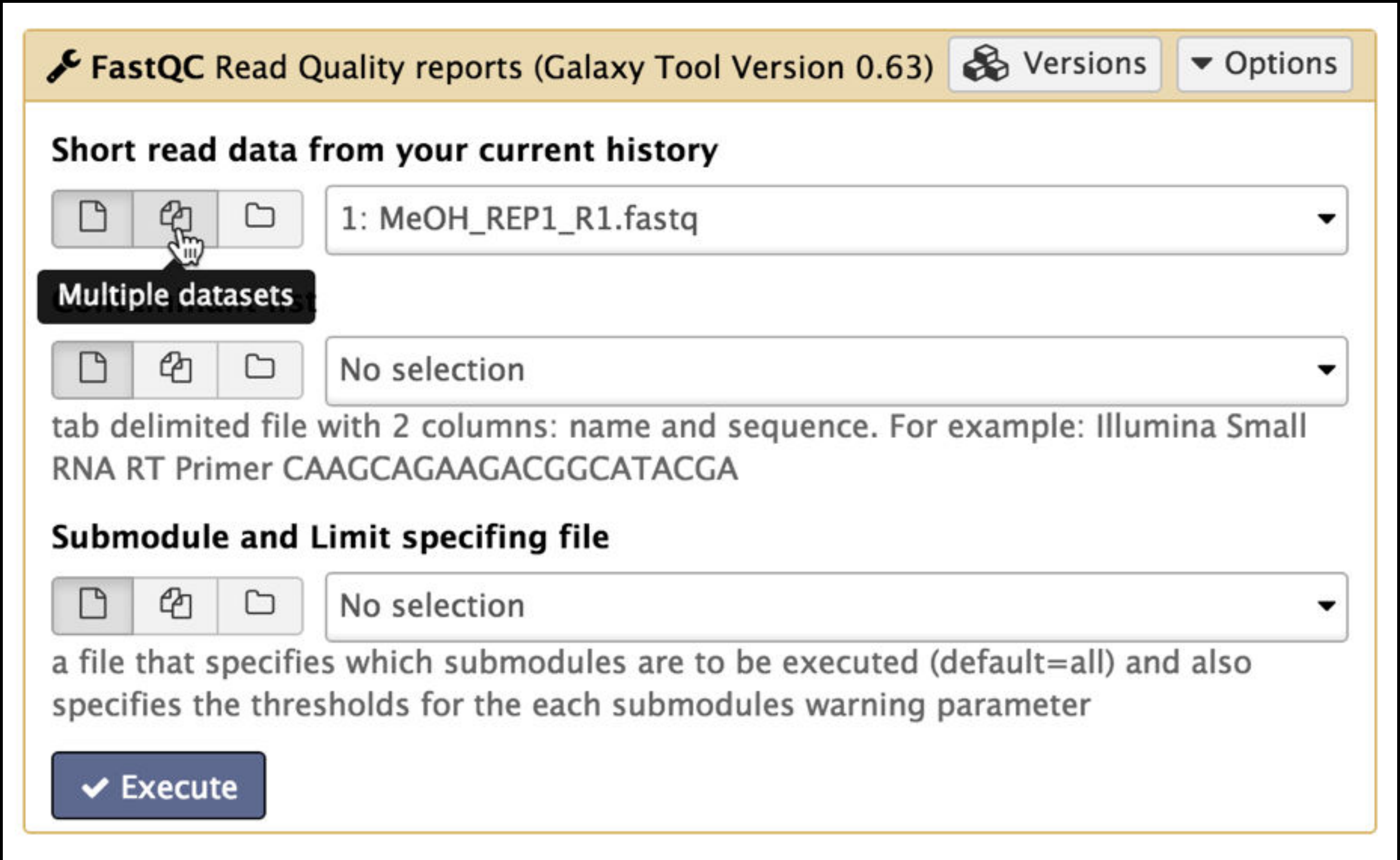
Now, just 10 more datasets to go!

Sit back and relax

This icon on a slide means please park your analysis skills for now. You may follow along in Galaxy, but there is **no need** to click Execute.

We will do the heavy lifting for you!

Your Friend: The Multiple datasets button



FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

1: MeOH_REP1_R1.fastq

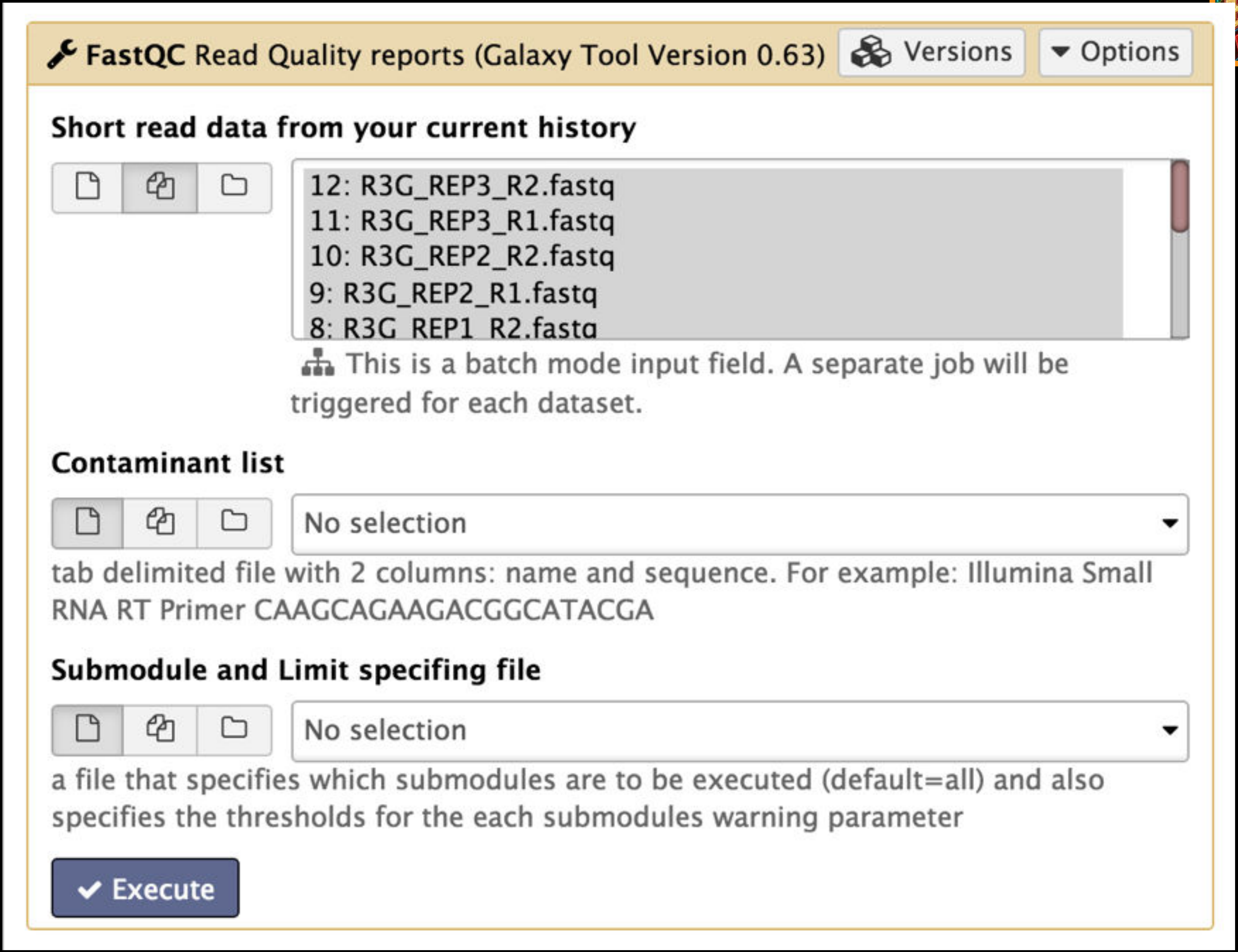
No selection

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file

No selection

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter



FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

12: R3G_REP3_R2.fastq
11: R3G_REP3_R1.fastq
10: R3G_REP2_R2.fastq
9: R3G_REP2_R1.fastq
8: R3G_REP1_R2.fastq

This is a batch mode input field. A separate job will be triggered for each dataset.

Contaminant list

No selection

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file

No selection

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Leap Forward!

Import one of these shared histories:

Shared Data → Published Histories →

RNA-Seq, Post-QC, reduced or

RNA-Seq, Post-QC

RNA-seq Exercise: Mapping with Tophat2

- 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

- Supply your own junction Data? → Yes
  - Use Gene Annotation → Yes
  - Gene Model Annotation → genes_chr12.gtf

You can also restrict Tophat to known annotations

- 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

TopHat Manual

Mapping w/ Tophat: Max # of Alignments Allowed

Some reads align to more than one place equally well.

For such reads, how many should Tophat include?

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

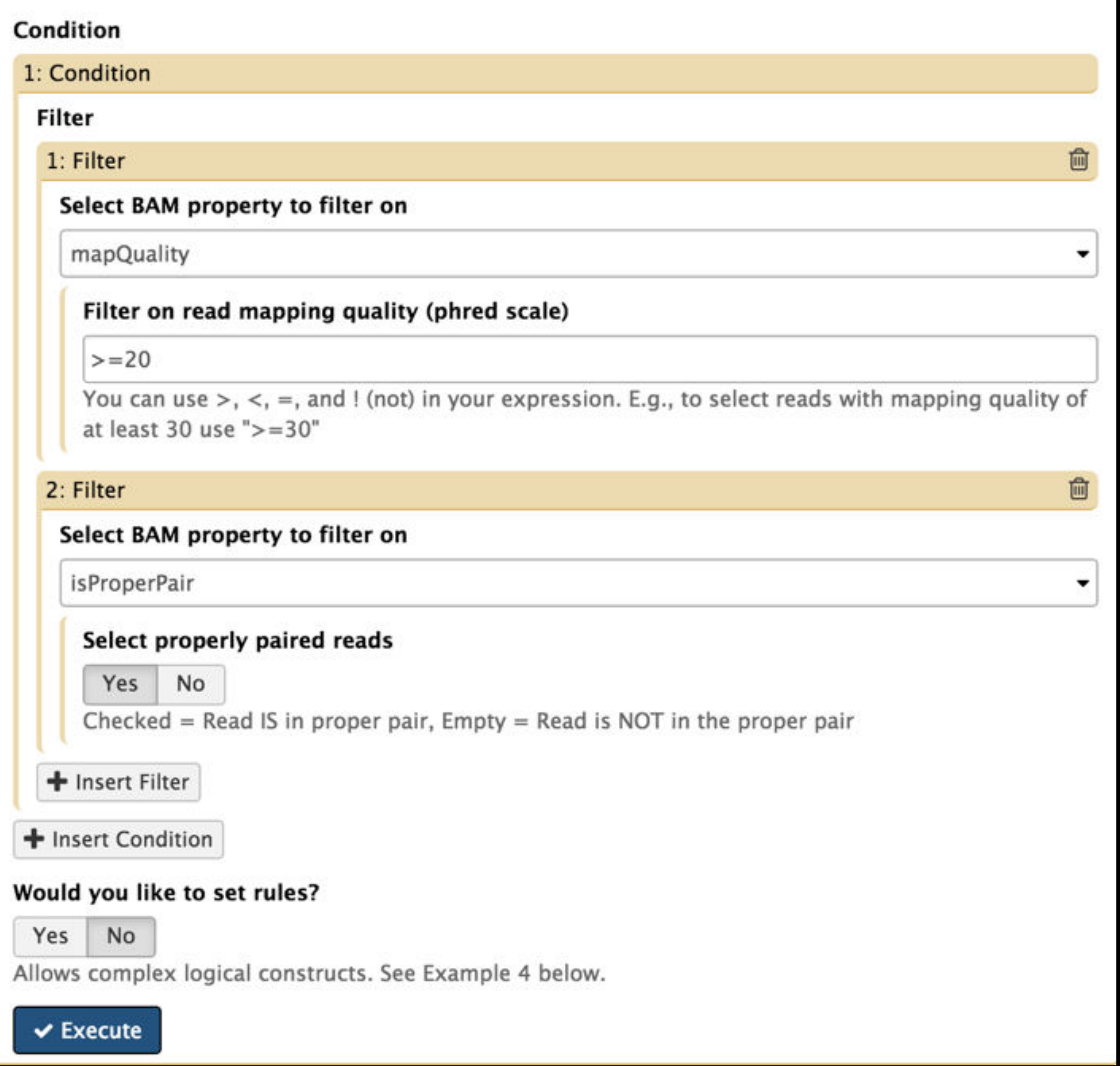
Tophat breaks ties randomly.

Tophat assigns equal fractional credit to all *n* mappings

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

TopHat Manual

Mapping With Tophat: What to keep?



Condition

1: Condition

Filter

1: Filter

Select BAM property to filter on

mapQuality

Filter on read mapping quality (phred scale)

>=20

You can use >, <, =, and ! (not) in your expression. E.g., to select reads with mapping quality of at least 30 use ">=30"

2: Filter

Select BAM property to filter on

isProperPair

Select properly paired reads

[✓ Yes] No

Checked = Read IS in proper pair, Empty = Read is NOT in the proper pair

Would you like to set rules?

Yes [✓ No]

Allows complex logical constructs. See Example 4 below.

NGS BAM Tools → Filter

Shared History: RNA-Seq through Mapping or RNA-Seq through Mapping, reduced

Mapping With Tophat: Only 5 more to do!

Hmmm.

Could use *Multiple Datasets* feature like we did with FastQC.
Could also construct *workflows*.

Another solution is

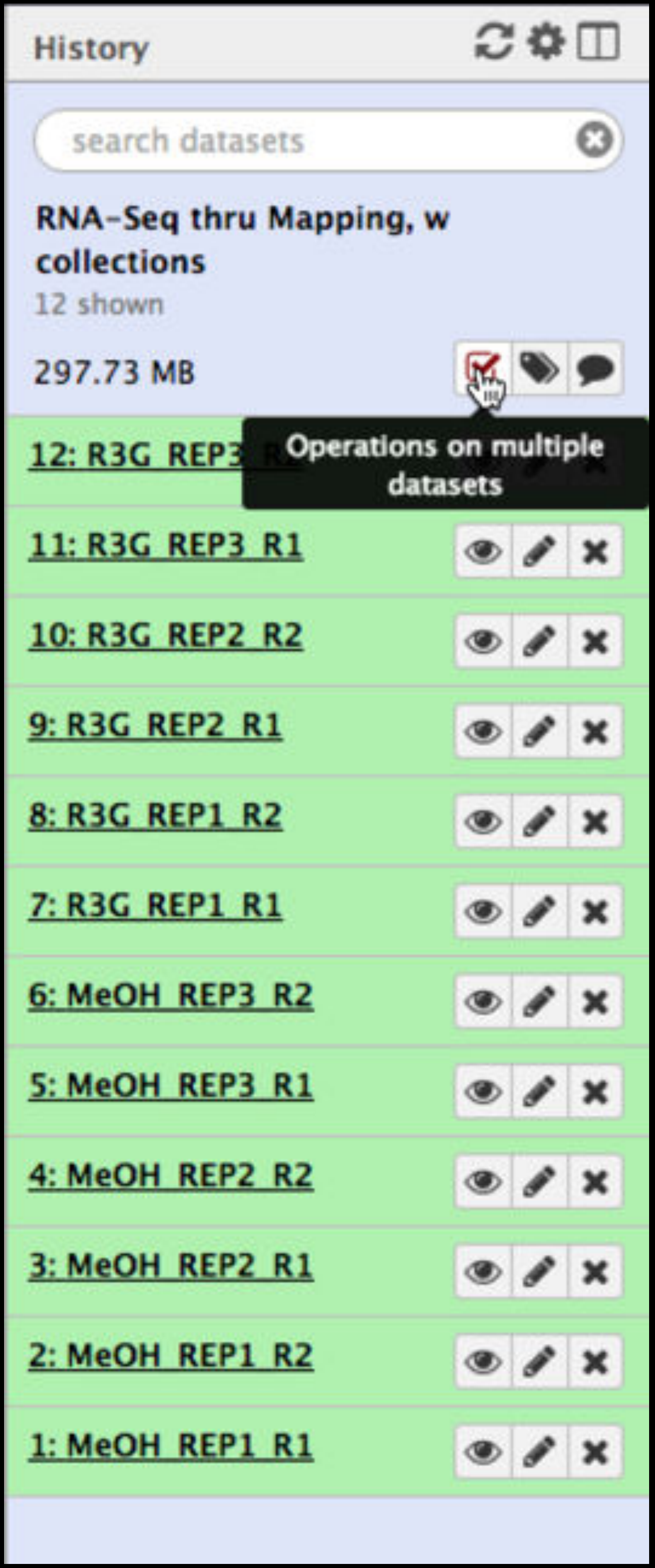
Collections

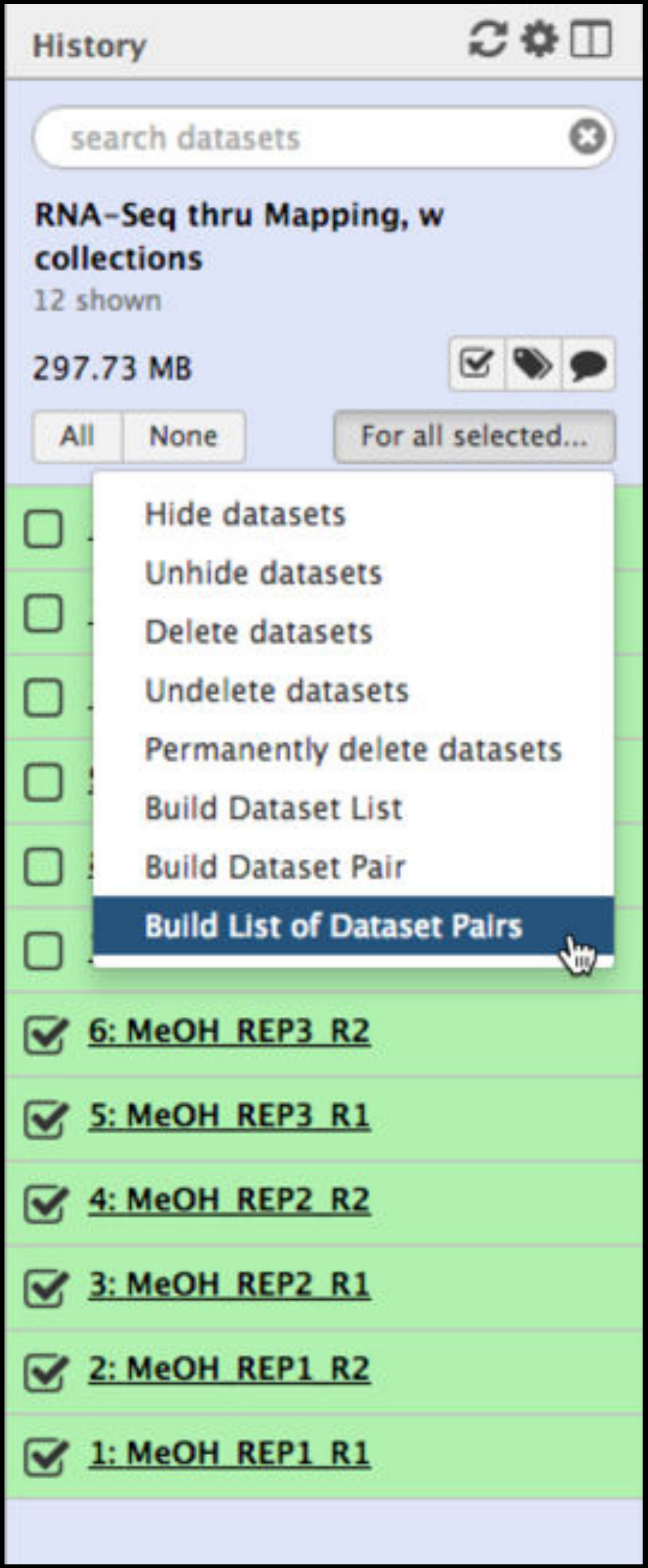
Dataset collections!

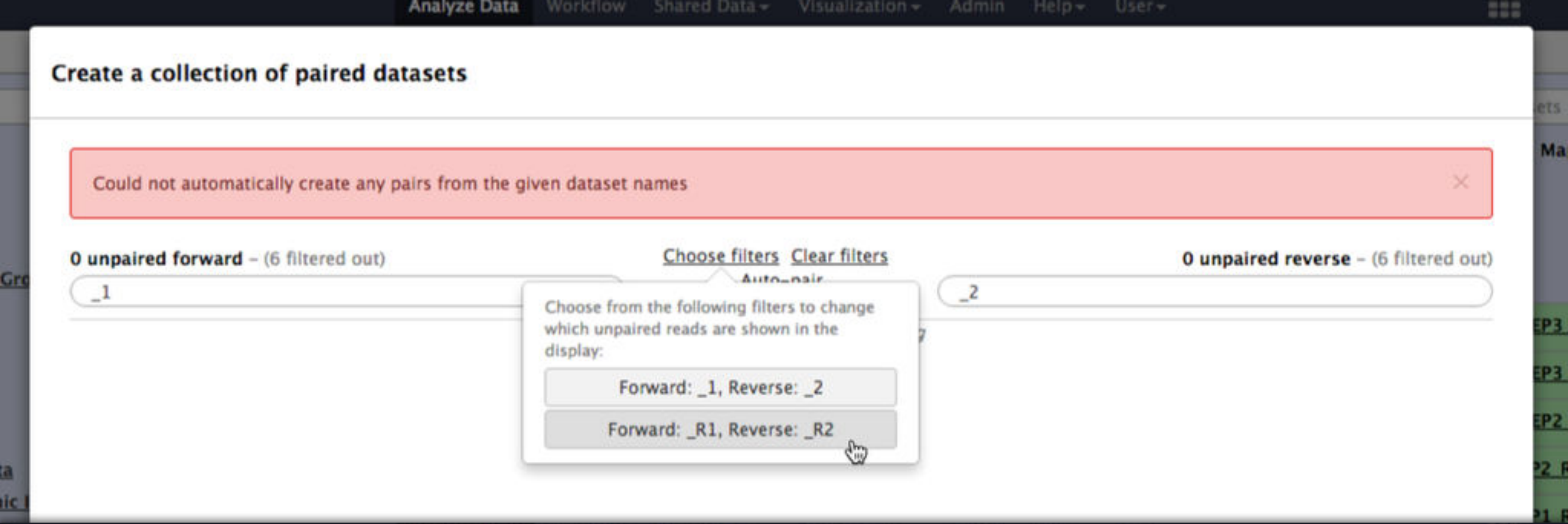
Dataset Collections give Galaxy semantic knowledge about dataset relationships.

Tools can then take advantage of this knowledge.

Dataset collections







Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names

0 unpaired forward – (6 filtered out)

Choose filters Clear filters

0 unpaired reverse – (6 filtered out)

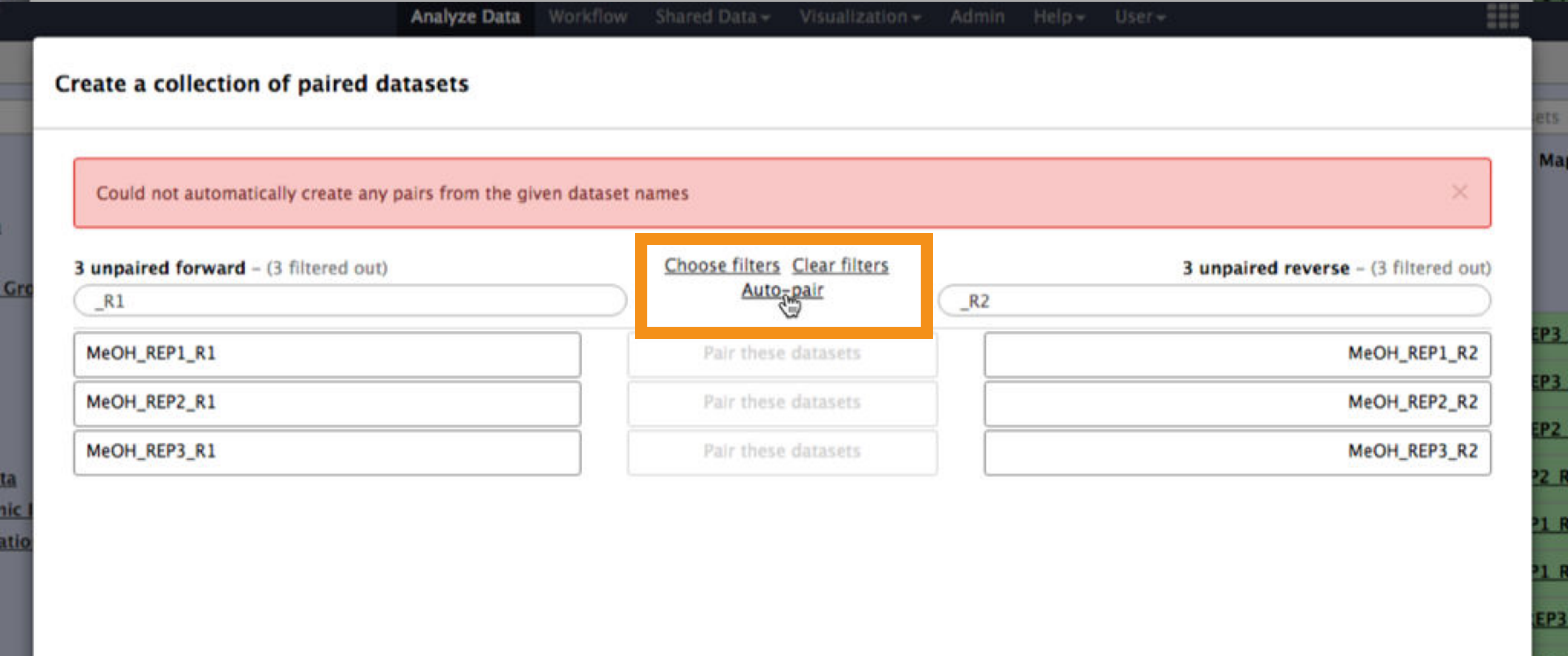
_1

_2

Choose from the following filters to change which unpaired reads are shown in the display:

Forward: _1, Reverse: _2

Forward: _R1, Reverse: _R2



Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names

3 unpaired forward – (3 filtered out)

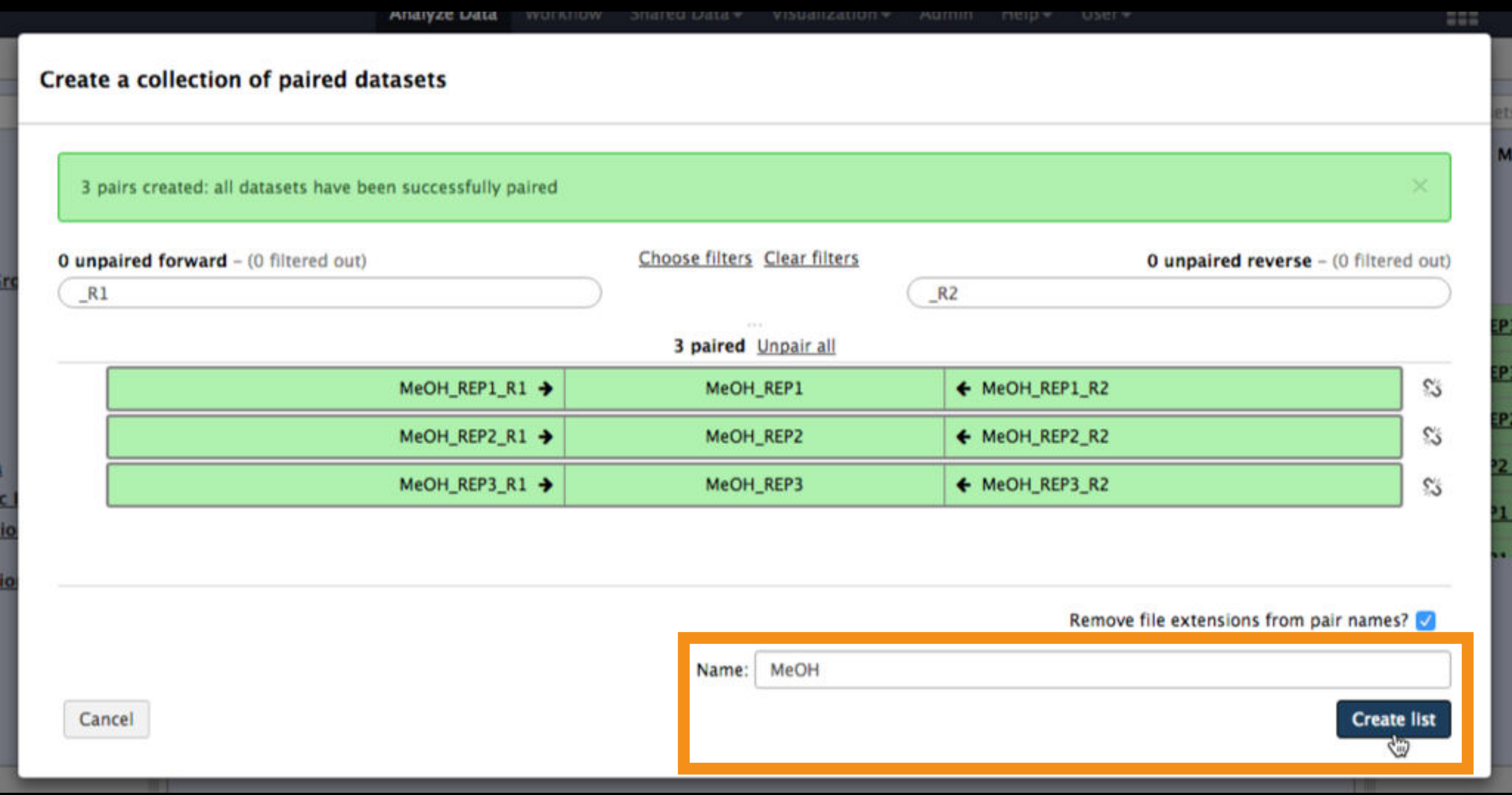
Choose filters Clear filters
Auto-pair

3 unpaired reverse – (3 filtered out)

_R1

_R2

MeOH_REP1_R1	Pair these datasets	MeOH_REP1_R2
MeOH_REP2_R1	Pair these datasets	MeOH_REP2_R2
MeOH_REP3_R1	Pair these datasets	MeOH_REP3_R2



Create a collection of paired datasets

3 pairs created: all datasets have been successfully paired

0 unpaired forward – (0 filtered out)

_R1

0 unpaired reverse – (0 filtered out)

_R2

3 paired

MeOH_REP1_R1 → MeOH_REP1 ← MeOH_REP1_R2

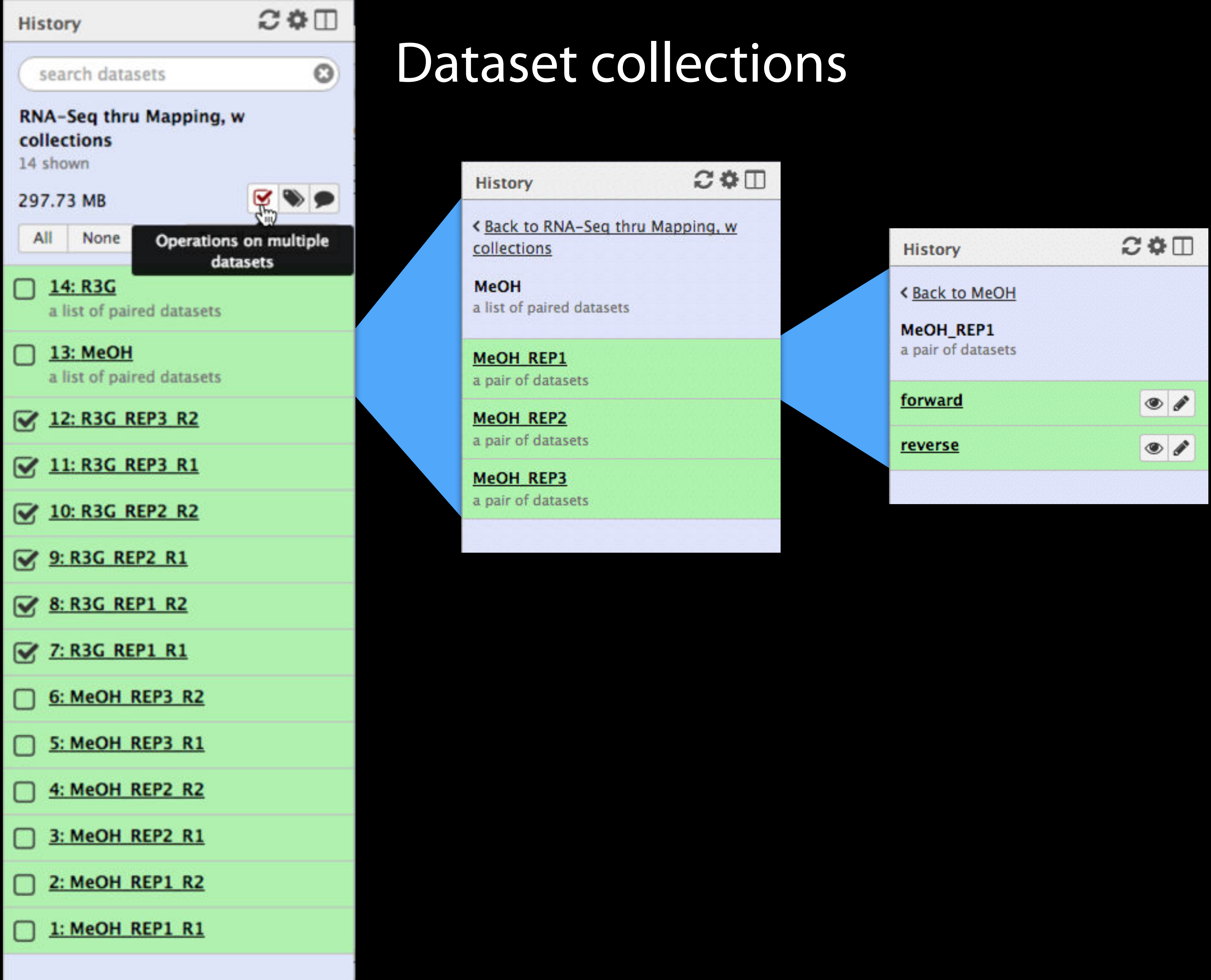
MeOH_REP2_R1 → MeOH_REP2 ← MeOH_REP2_R2

MeOH_REP3_R1 → MeOH_REP3 ← MeOH_REP3_R2

Remove file extensions from pair names? ☑

Name: MeOH

Dataset collections



History

search datasets

RNA-Seq thru Mapping, w collections

14 shown

297.73 MB

All None

Operations on multiple datasets

□ 14: R3G
a list of paired datasets

□ 13: MeOH
a list of paired datasets

☑ 12: R3G_REP3_R2

☑ 11: R3G_REP3_R1

☑ 10: R3G_REP2_R2

☑ 9: R3G_REP2_R1

☑ 8: R3G_REP1_R2

☑ 7: R3G_REP1_R1

□ 6: MeOH_REP3_R2

□ 5: MeOH_REP3_R1

□ 4: MeOH_REP2_R2

□ 3: MeOH_REP2_R1

□ 2: MeOH_REP1_R2

□ 1: MeOH_REP1_R1

History

< Back to RNA-Seq thru Mapping, w collections

MeOH
a list of paired datasets

MeOH_REP1
a pair of datasets

MeOH_REP2
a pair of datasets

MeOH_REP3
a pair of datasets

History

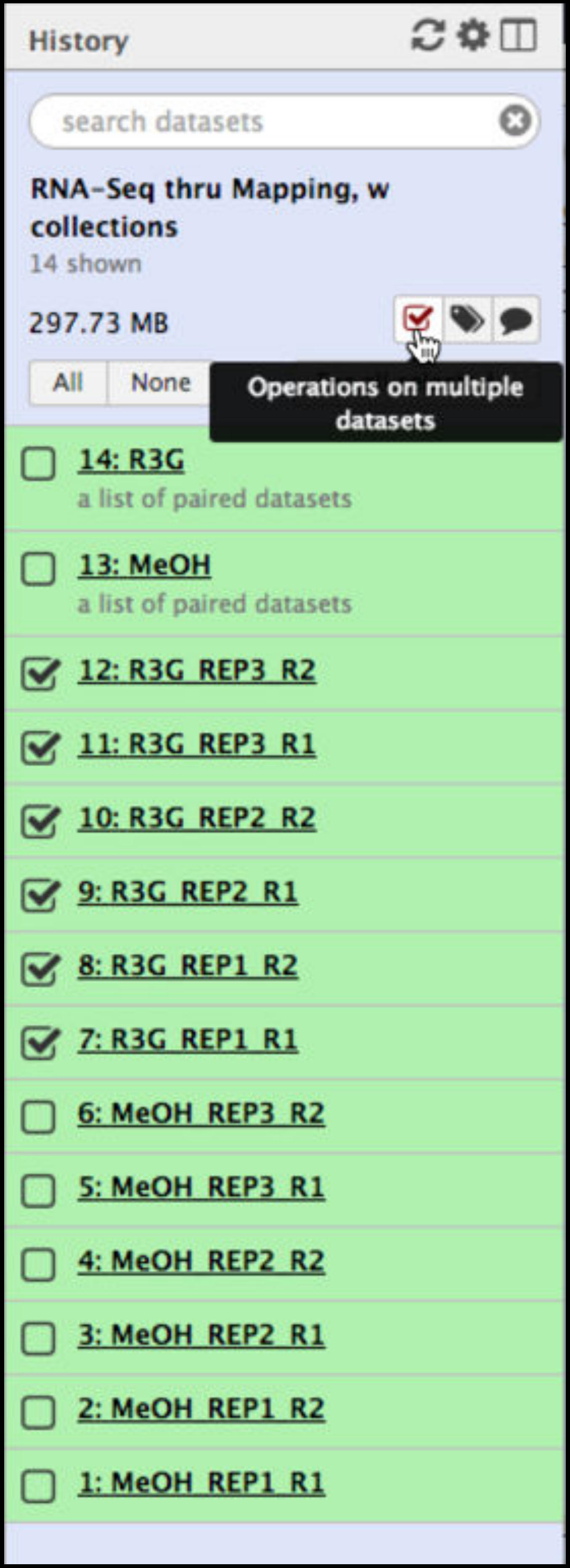
< Back to MeOH

MeOH_REP1
a pair of datasets

forward

reverse

Dataset collections Created



History

search datasets

RNA-Seq thru Mapping, w collections

14 shown

297.73 MB

All None

Operations on multiple datasets

□ 14: R3G
a list of paired datasets

□ 13: MeOH
a list of paired datasets

☑ 12: R3G_REP3_R2

☑ 11: R3G_REP3_R1

☑ 10: R3G_REP2_R2

☑ 9: R3G_REP2_R1

☑ 8: R3G_REP1_R2

☑ 7: R3G_REP1_R1

□ 6: MeOH_REP3_R2

□ 5: MeOH_REP3_R1

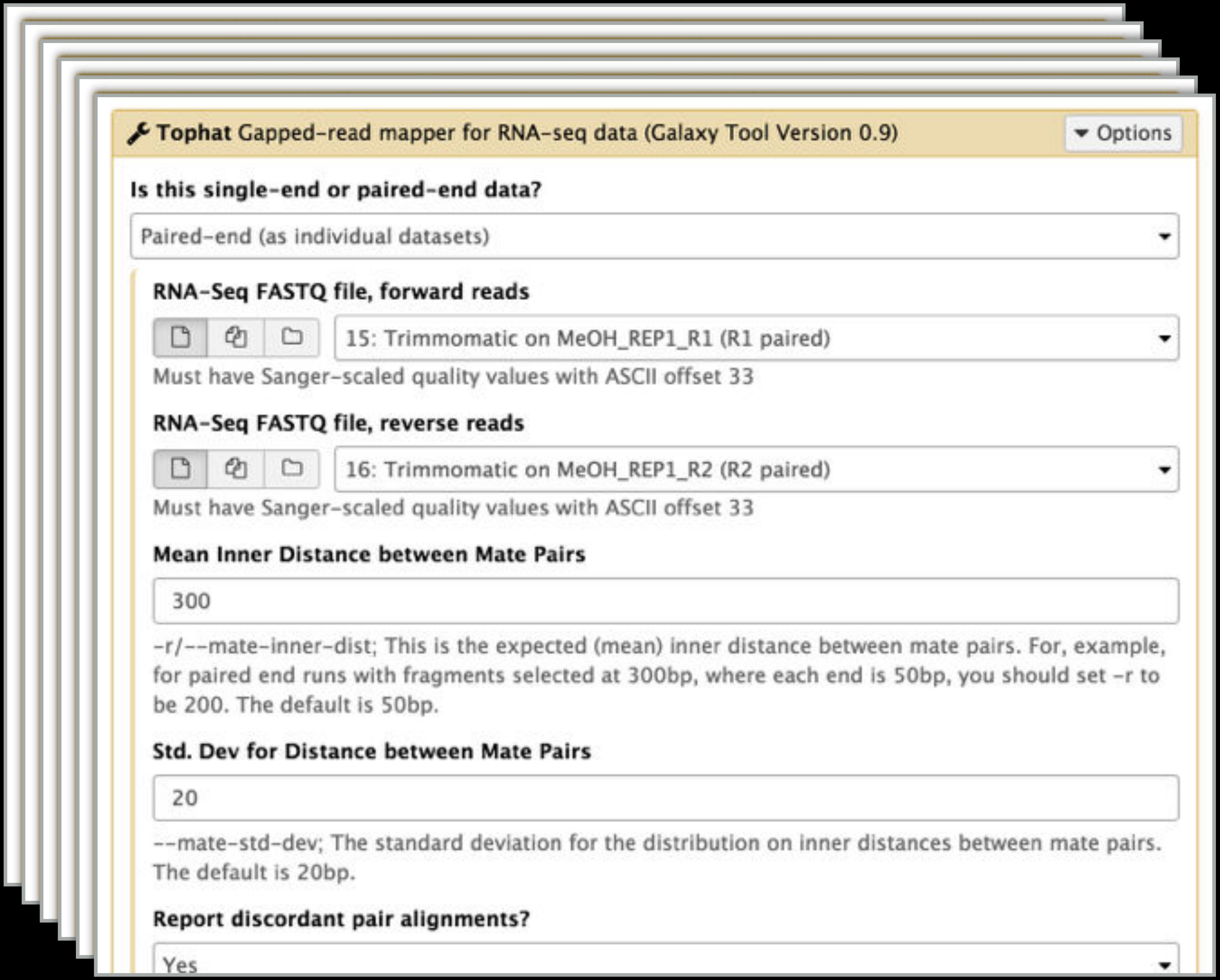
□ 4: MeOH_REP2_R2

□ 3: MeOH_REP2_R1

□ 2: MeOH_REP1_R2

□ 1: MeOH_REP1_R1

Before Dataset collections



Tophat Gapped-read mapper for RNA-seq data (Galaxy Tool Version 0.9)

Is this single-end or paired-end data?

Paired-end (as individual datasets)

RNA-Seq FASTQ file, forward reads

15: Trimmomatic on MeOH_REP1_R1 (R1 paired)

Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads

16: Trimmomatic on MeOH_REP1_R2 (R2 paired)

Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs

300

-r/--mate-inner-dist; This is the expected (mean) inner distance between mate pairs. For, example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set -r to be 200. The default is 50bp.

Std. Dev for Distance between Mate Pairs

20

--mate-std-dev; The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.

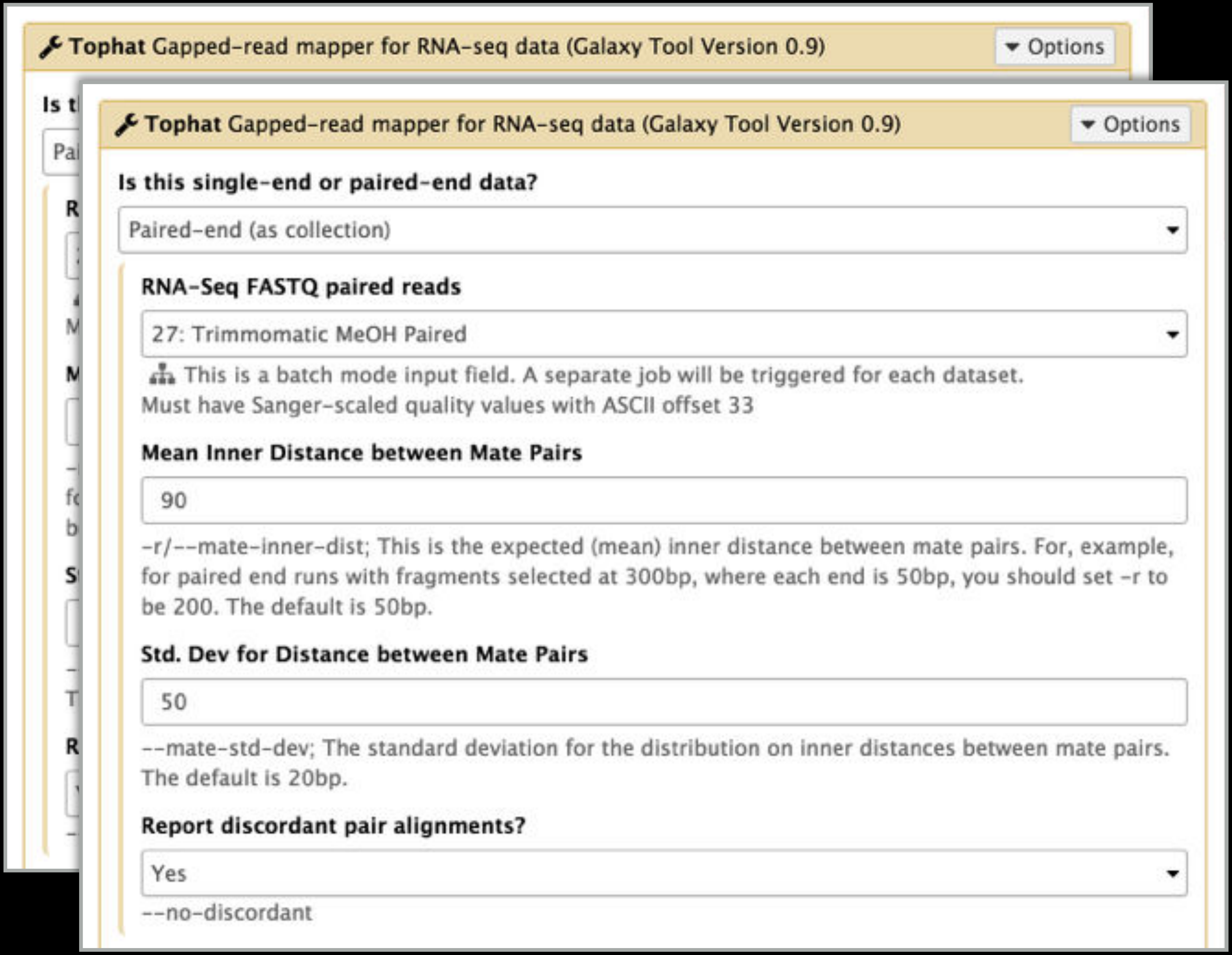
Report discordant pair alignments?

Yes

Old: x6

(once per pair - error prone; Trimmomatic was x12)

After Dataset collections



Tophat Gapped-read mapper for RNA-seq data (Galaxy Tool Version 0.9)

Tophat Gapped-read mapper for RNA-seq data (Galaxy Tool Version 0.9)

Is this single-end or paired-end data?

Paired-end (as collection)

RNA-Seq FASTQ paired reads

27: Trimmomatic MeOH Paired

This is a batch mode input field. A separate job will be triggered for each dataset.
Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs

90

-r/--mate-inner-dist; This is the expected (mean) inner distance between mate pairs. For, example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set -r to be 200. The default is 50bp.

Std. Dev for Distance between Mate Pairs

50

--mate-std-dev; The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.

Report discordant pair alignments?

Yes

--no-discordant

New: x2

(once per condition)

Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

http://bit.ly/gxyismms2016

Agenda

9:00 Welcome

9:20 Basic Analysis with Galaxy
A worked example demonstrating Galaxy Basics

10:45 Break

11:00 Basic Analysis into Reusable Workflows

12:20 Lunch (on your own)

1:20 RNA-Seq Analysis, Part I

2:50 Break

3:05 RNA-Seq Analysis, Part II

17:00 Done

http://bit.ly/gxyismms2016

All our data is mapped! Leap Forward!

Import one of these shared histories:

Shared Data → Published Histories →

RNA-Seq, Post-Mapping, reduced or

RNA-Seq, Post-Mapping

Differential expression with CuffDiff

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

NGS: RNA Analysis → Cuffdiff

Cuffdiff

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

Now supports geometric normalization, the same model used by DESeq (and in fact, it's now the default). Less prone to distortion from highly expressed genes.

Cuffdiff: Which transcript definitions to use?

We'll use the official genome annotations

(We told Tophat to only use these)

But there are a world of options out there for discovering and using novel transcripts.

StringTie, Cufflinks, Cuffmerge, ...

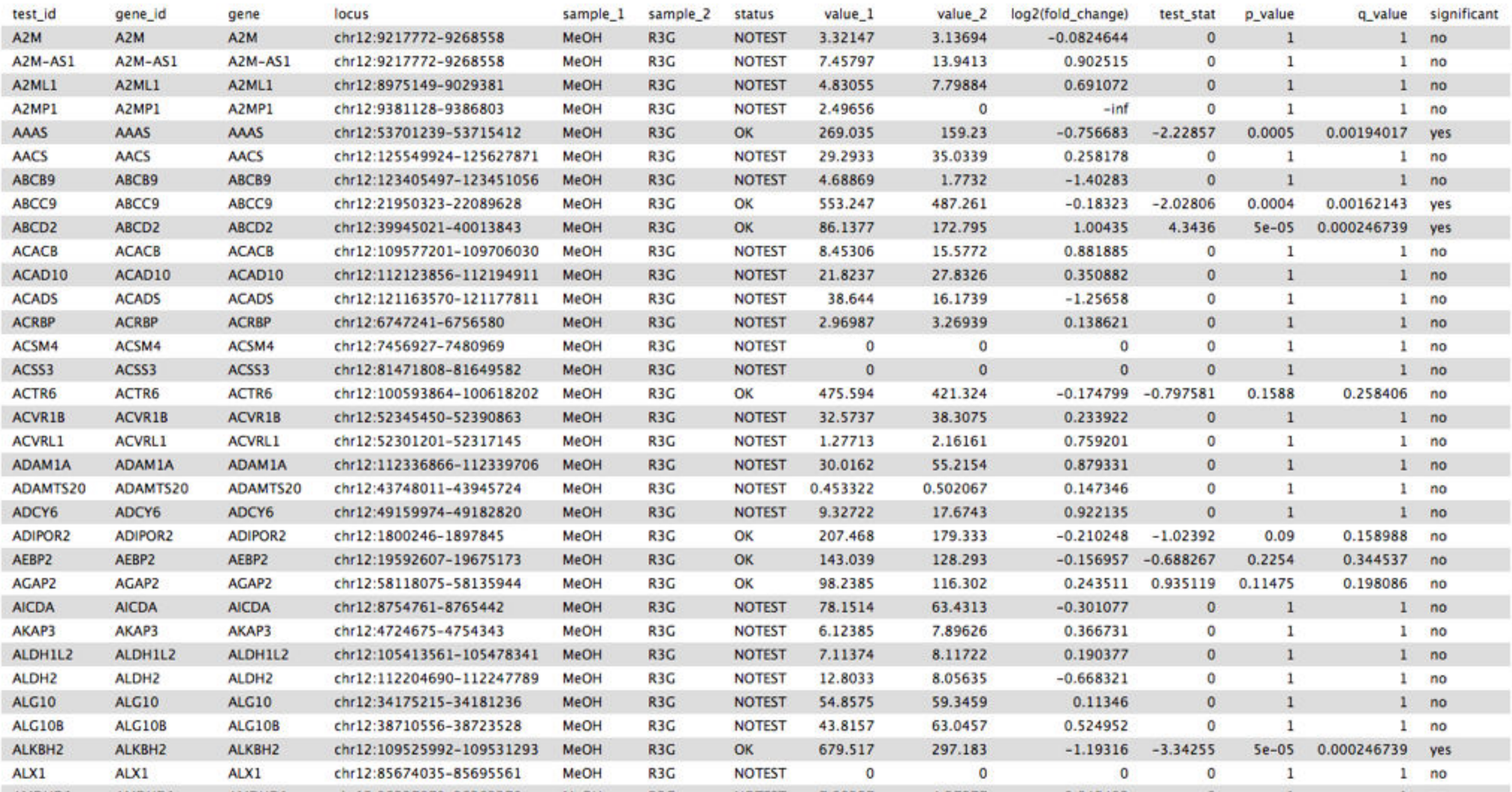
Cuffdiff

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

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
p_value	Uncorrected P value for test statistic
q_value	FDR-adjusted p-value for the test statistic
status	Was there enough data to run the test?
significant	and, was the gene differentially expressed?

Cuffdiff

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

Returns the list of genes with

1) enough data to make a call, and

2) that are called as differentially expressed.

Cuffdiff: Next Steps

Try running Cuffdiff with different normalization and dispersion estimation methods.

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

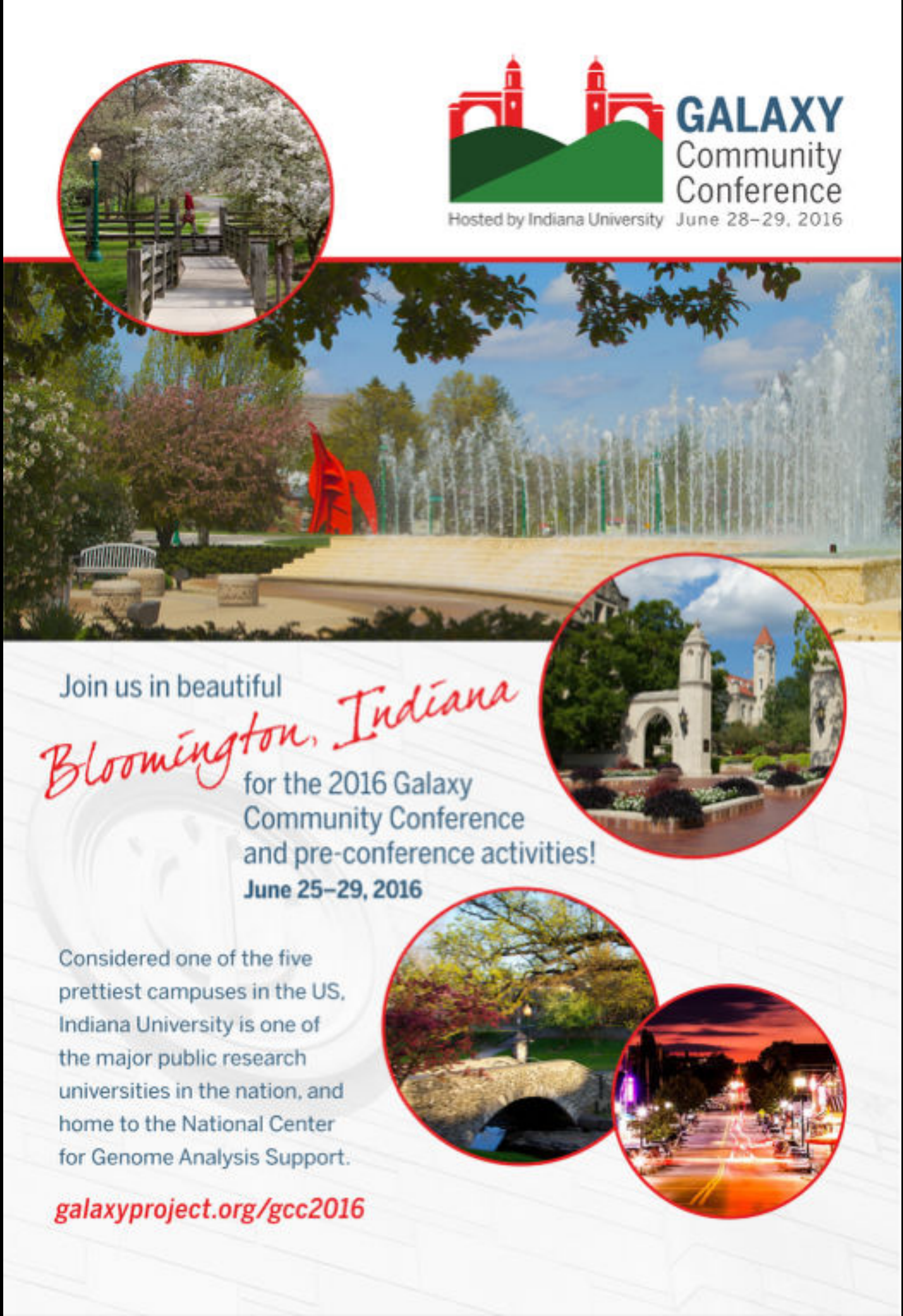
Are there any patterns to the identified genes?

Shared History: RNA-Seq trimmed reads to diff gene

2016 Galaxy Community Conference (GCC2016)

June 25-29, 2016
Bloomington, Indiana

galaxyproject.org/GCC2016



GALAXY
Community
Conference

Hosted by Indiana University
June 28–29, 2016

Join us in beautiful *Bloomington, Indiana* for the 2016 Galaxy Community Conference and pre-conference activities!

June 25–29, 2016

Considered one of the five prettiest campuses in the US, Indiana University is one of the major public research universities in the nation, and home to the National Center for Genome Analysis Support.

galaxyproject.org/gcc2016

Galaxy Resources and Community

Mailing Lists (very active)
Unified Search
Issues Board
Events Calendar, News Feed
Community Wiki
GalaxyAdmins
Screencasts
Tool Shed
Public Installs
CiteULike group, Mendeley mirror
Annual Community Meting

http://wiki.galaxyproject.org

Galaxy Community Resources: Galaxy Biostar

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

Absolutely have to encourage community support.

Project traditionally used mailing list

Moved the user support list to Galaxy Biostar, an online forum, that uses the Biostar platform

https://biostar.usegalaxy.org/

Galaxy Community Resources: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (2336 posts in 2015, 1000+ members)

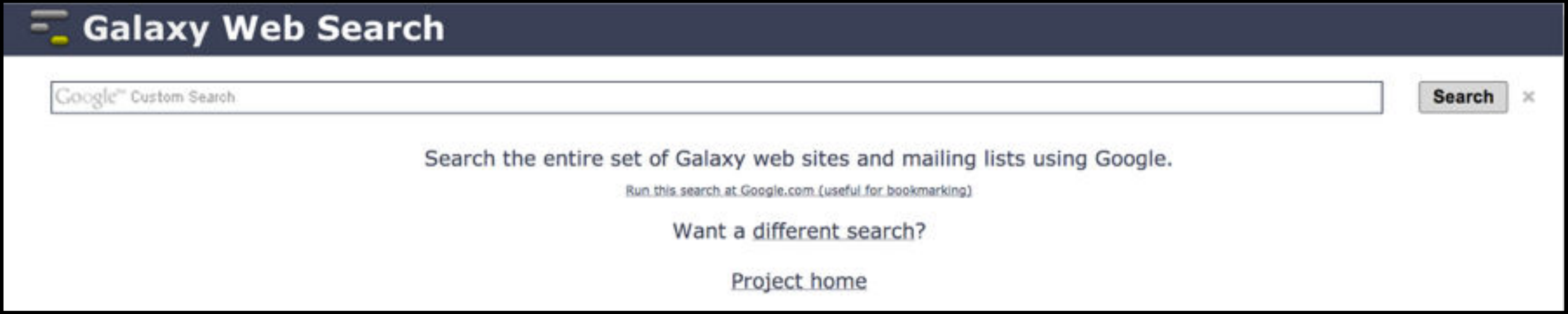
Galaxy-Announce

Project announcements, low volume, moderated

Low volume (36 posts in 2015, 6500+ members)

Also Galaxy-UK, -France, -Proteomics, -Training, ...

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



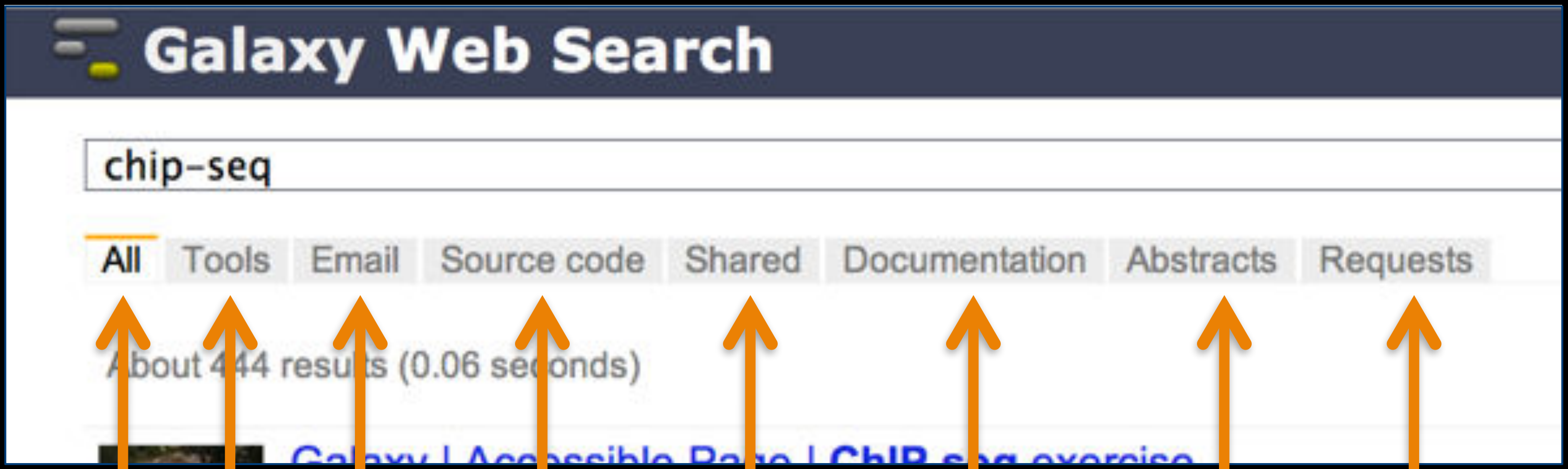
Galaxy Web Search

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

Run this search at Google.com (useful for bookmarking)

Want a different search?

Project home



Galaxy Web Search

chip-seq

About 444 results (0.06 seconds)

Galaxy | Accessible Page | ChIP-seq exercise

Find

Everything on …

Tools for …

Email about …

Source code for …

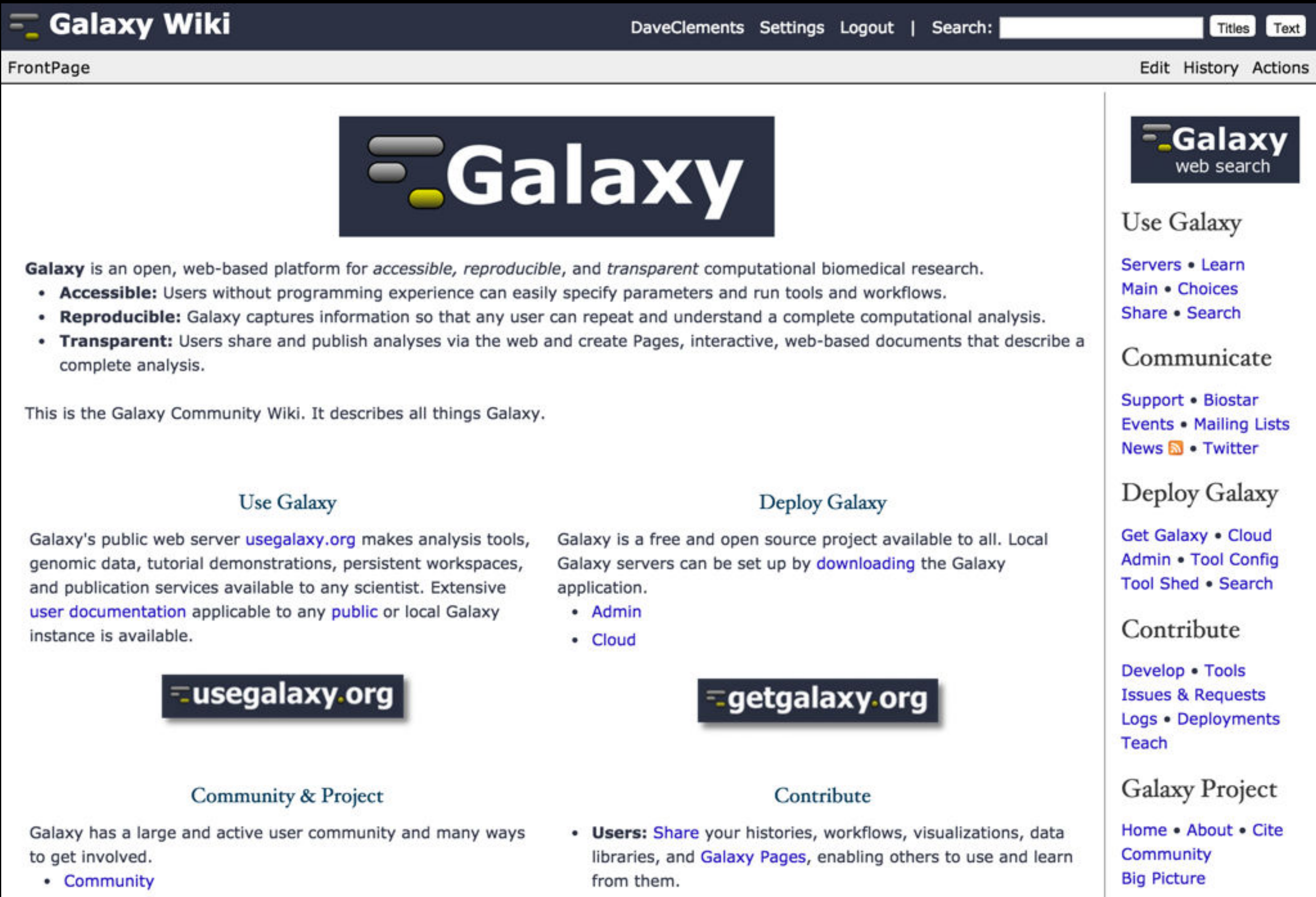
Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

http://wiki.galaxyproject.org



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

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

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

Use Galaxy

Galaxy's public web server usegalaxy.org makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive user documentation applicable to any public or local Galaxy instance is available.

Deploy Galaxy

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

- Admin
- Cloud

Community & Project

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

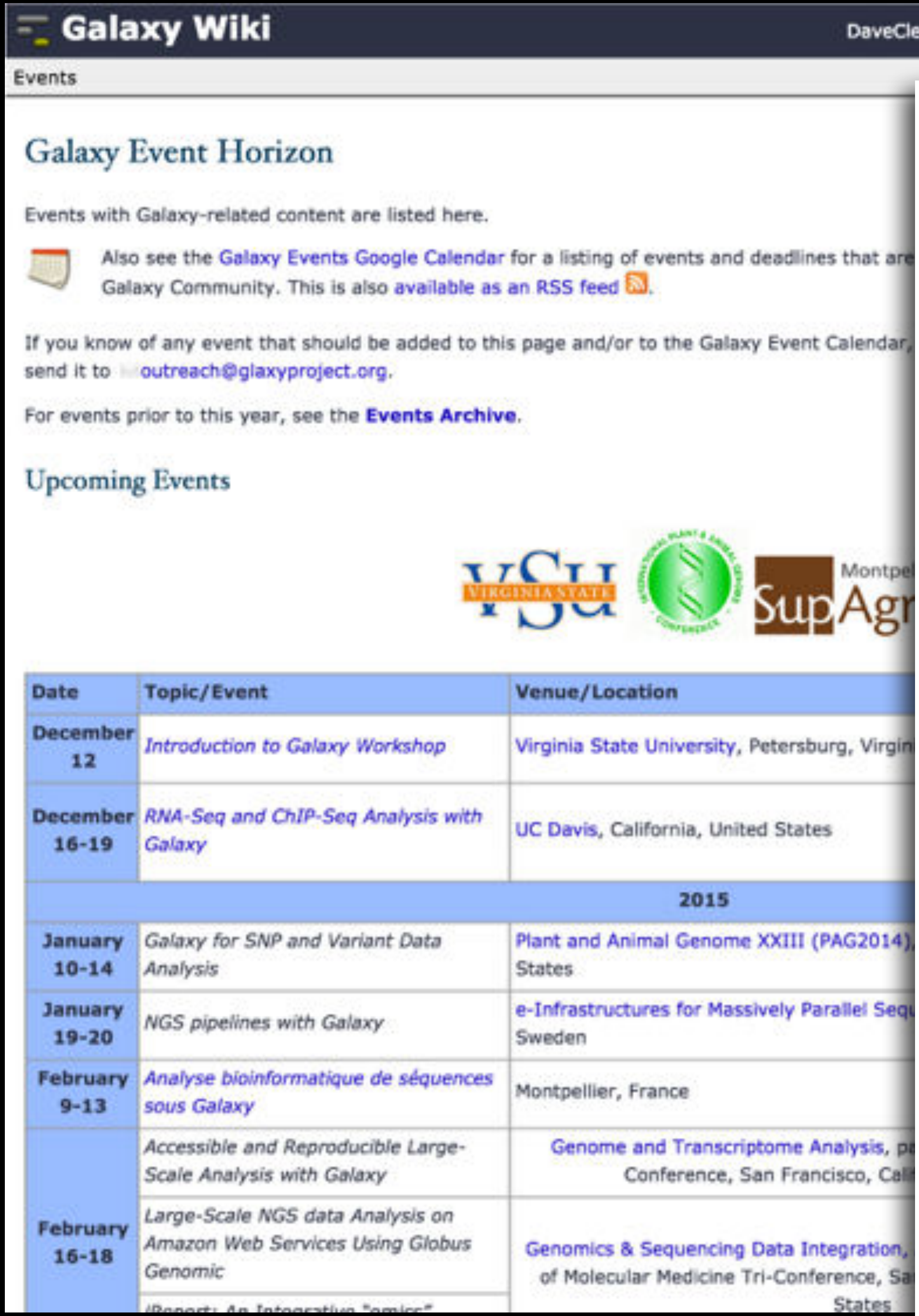
- Community

Contribute

- **Users:** Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.

Events

News



Galaxy Event Horizon

Events with Galaxy-related content are listed here.

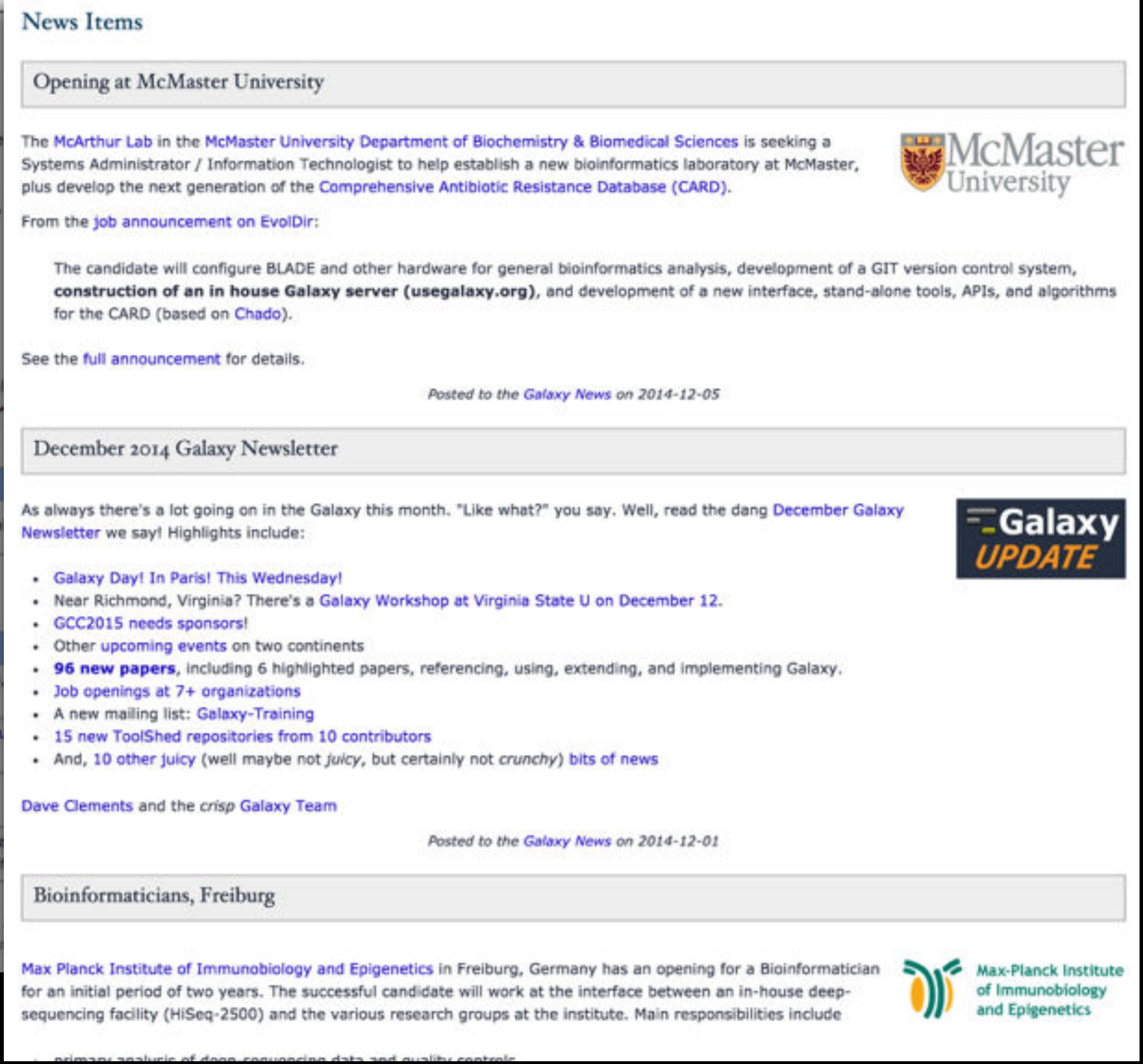
Also see the Galaxy Events Google Calendar for a listing of events and deadlines that are
Galaxy Community. This is also available as an RSS feed.

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

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

Upcoming Events

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



News Items

Opening at McMaster University

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

From the job announcement on EvolDir:

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

See the full announcement for details.

Posted to the Galaxy News on 2014-12-05

December 2014 Galaxy Newsletter

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

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

Dave Clements and the *crisp* Galaxy Team

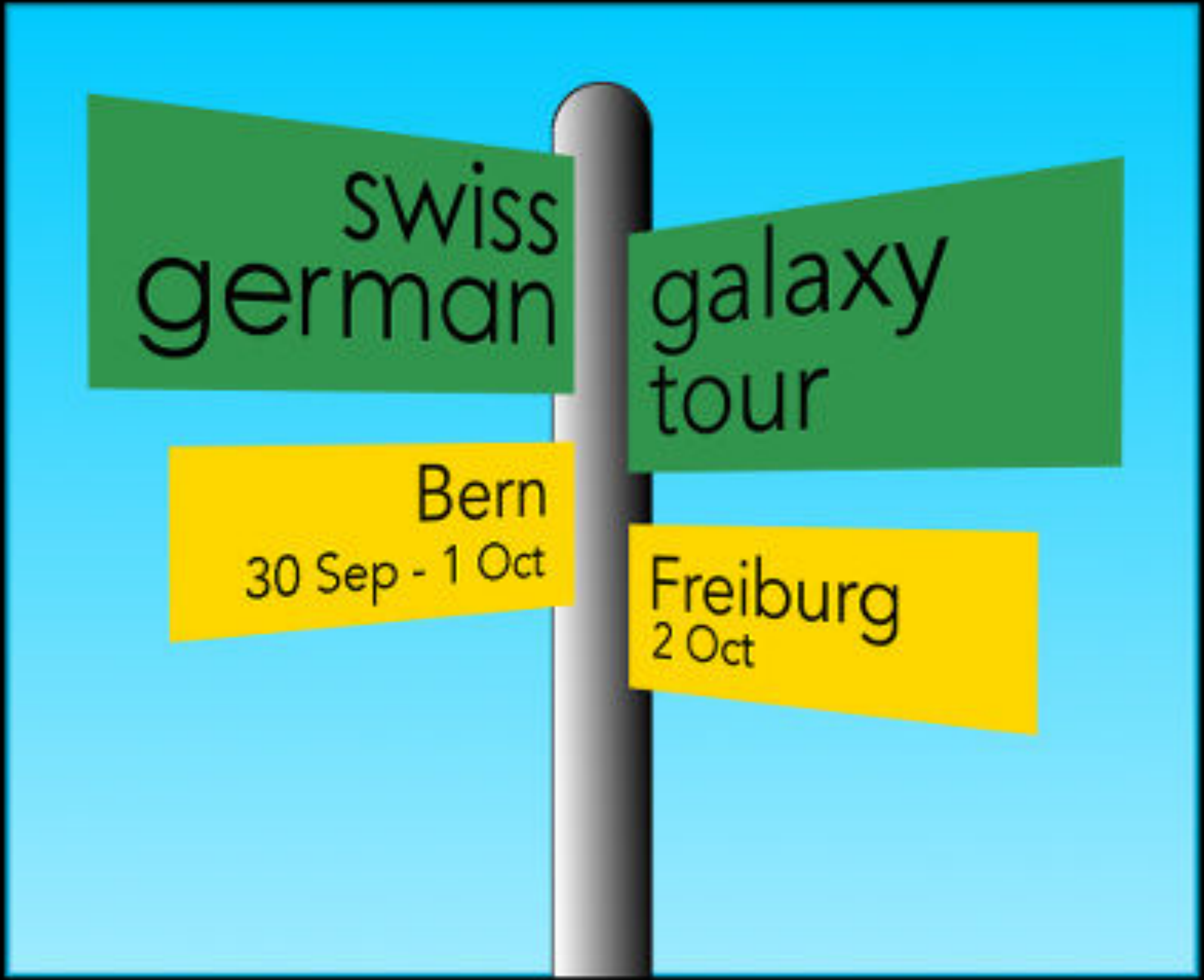
Posted to the Galaxy News on 2014-12-01

Bioinformaticians, Freiburg

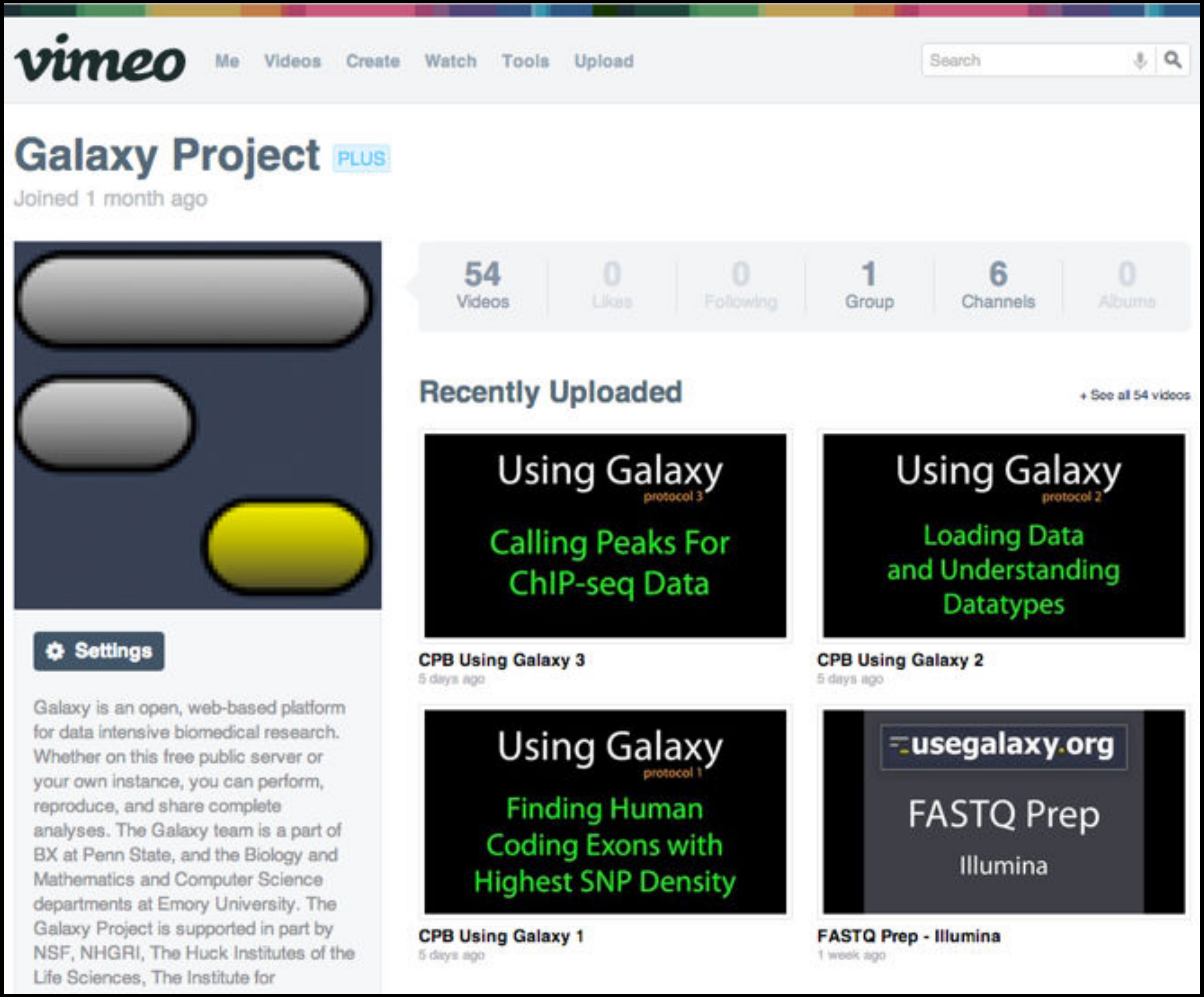
Max Planck Institute of Immunobiology and Epigenetics in Freiburg, Germany has an opening for a Bioinformatician for an initial period of two years. The successful candidate will work at the interface between an in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the institute.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

- primary analysis of deep-sequencing data and quality controls

We also support community organized efforts and events.



Galaxy Resources & Community: Videos



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

usegalaxy.org FASTQ Prep Illumina

FASTQ Prep - Illumina

1 week ago

Settings

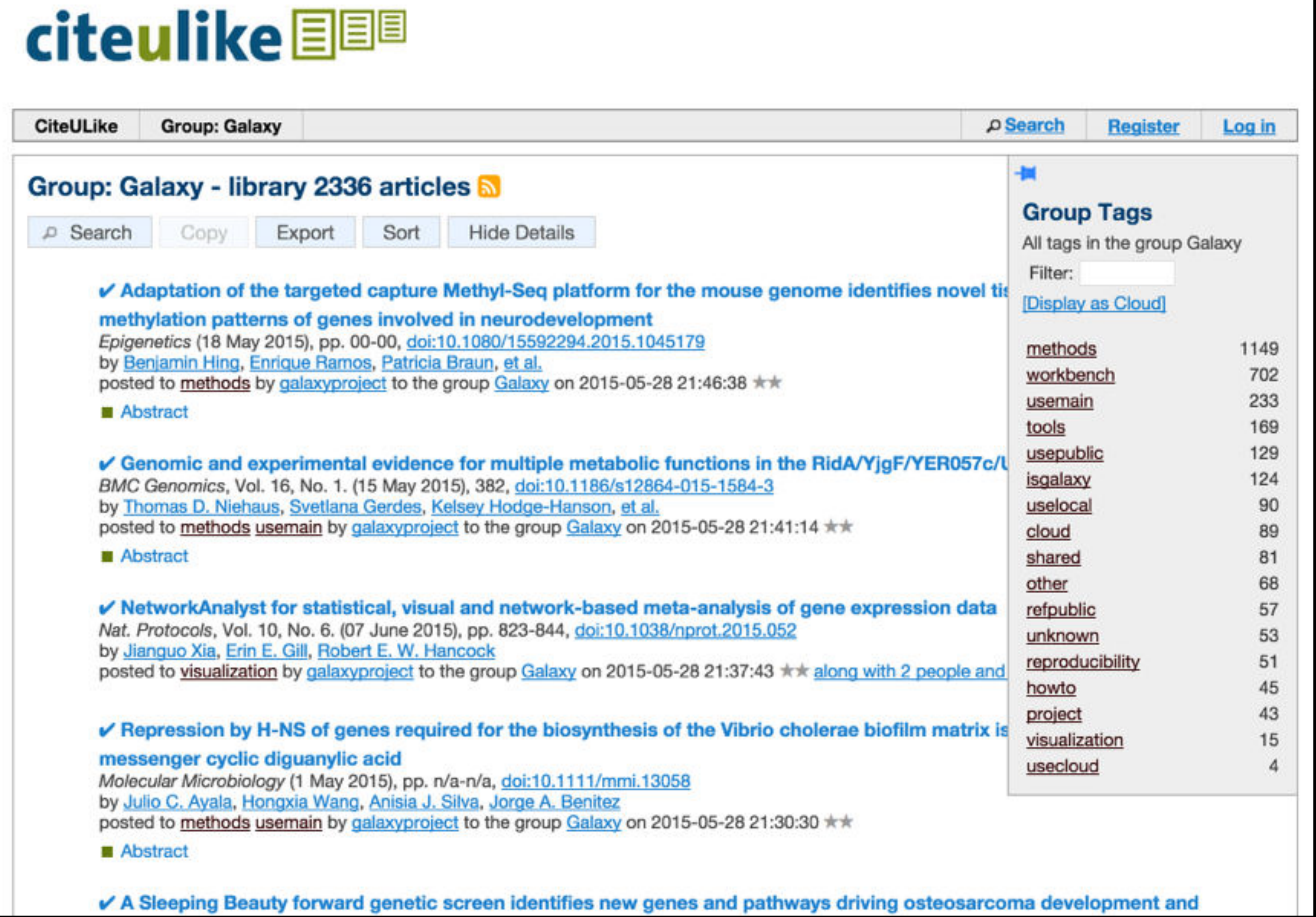
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments 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

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

Group: Galaxy - library 2336 articles

⌕ Search | Copy | Export | Sort | Hide Details

✔ **Adaptation of the targeted capture Methyl-Seq platform for the mouse genome identifies novel tis methylation patterns of genes involved in neurodevelopment**
Epigenetics (18 May 2015), pp. 00-00, doi:10.1080/15592294.2015.1045179
by Benjamin Hing, Enrique Ramos, Patricia Braun, et al.
posted to methods by galaxyproject to the group Galaxy on 2015-05-28 21:46:38 ★★
■ Abstract

✔ **Genomic and experimental evidence for multiple metabolic functions in the RidA/YjgF/YER057c/U**
BMC Genomics, Vol. 16, No. 1. (15 May 2015), 382, doi:10.1186/s12864-015-1584-3
by Thomas D. Niehaus, Svetlana Gerdes, Kelsey Hodge-Hanson, et al.
posted to methods usemain by galaxyproject to the group Galaxy on 2015-05-28 21:41:14 ★★
■ Abstract

✔ **NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data**
Nat. Protocols, Vol. 10, No. 6. (07 June 2015), pp. 823-844, doi:10.1038/nprot.2015.052
by Jianguo Xia, Erin E. Gill, Robert E. W. Hancock
posted to visualization by galaxyproject to the group Galaxy on 2015-05-28 21:37:43 ★★ along with 2 people and

✔ **Repression by H-NS of genes required for the biosynthesis of the Vibrio cholerae biofilm matrix is messenger cyclic diguanylic acid**
Molecular Microbiology (1 May 2015), pp. n/a-n/a, doi:10.1111/mmi.13058
by Julio C. Ayala, Hongxia Wang, Anisia J. Silva, Jorge A. Benitez
posted to methods usemain by galaxyproject to the group Galaxy on 2015-05-28 21:30:30 ★★
■ Abstract

✔ **A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and**

Group Tags

All tags in the group Galaxy

Filter:

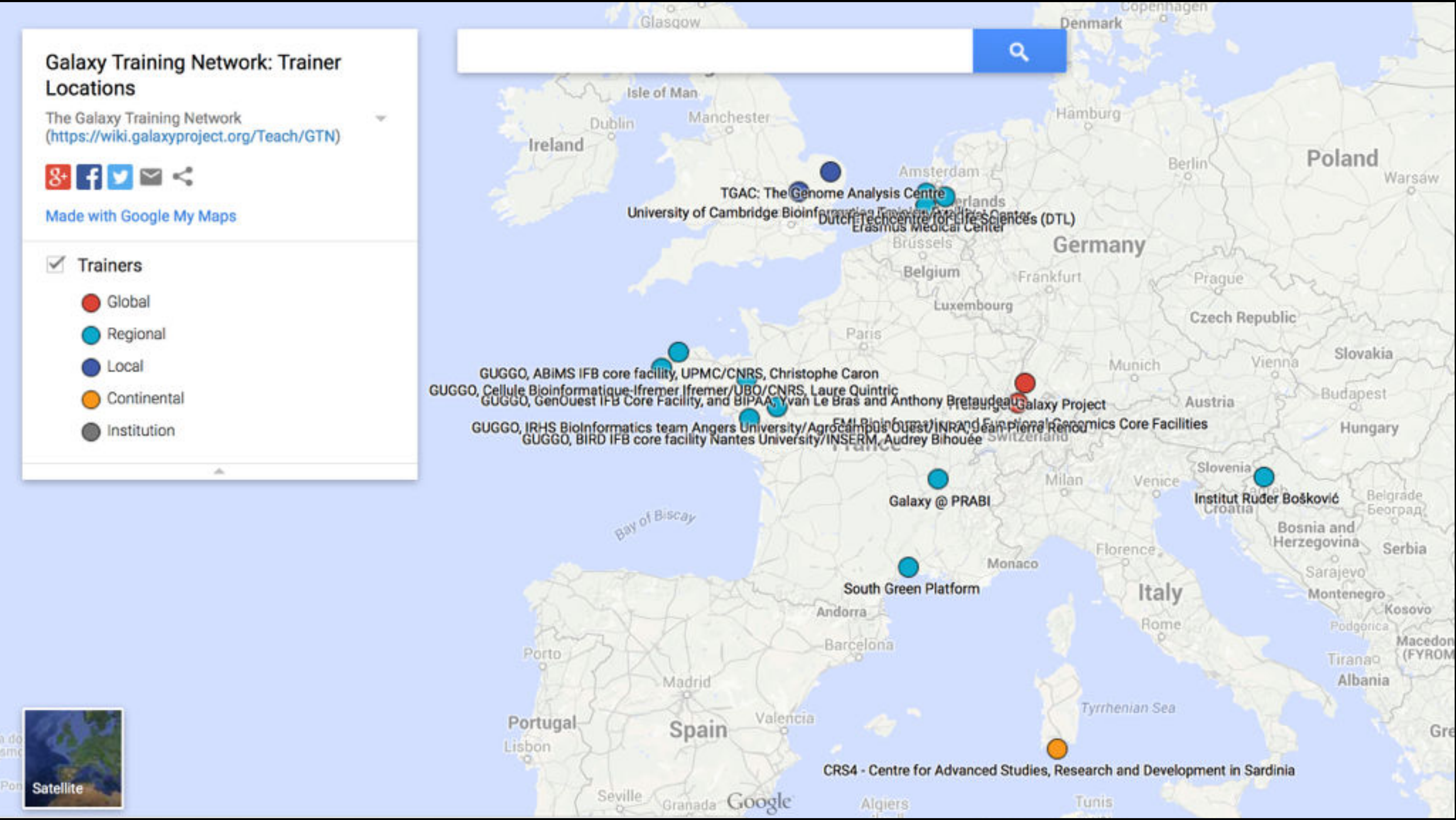
[Display as Cloud]

methods	1149
workbench	702
usemain	233
tools	169
usepublic	129
isgalaxy	124
uselocal	90
cloud	89
shared	81
other	68
refpublic	57
unknown	53
reproducibility	51
howto	45
project	43
visualization	15
usecloud	4

Now almost 3000 papers

http://bit.ly/gxycul

Scaling Training



Galaxy Training Network: Trainer Locations

The Galaxy Training Network (https://wiki.galaxyproject.org/Teach/GTN)

Made with Google My Maps

☑ Trainers

- ● Global
- ● Regional
- ● Local
- ● Continental
- ● Institution

TGAC: The Genome Analysis Centre
University of Cambridge Bioinformatics Training Facility
Dutch Techcentre for Life Sciences (DTL)
Erasmus Medical Center
GUGGO, ABiMS IFB core facility, UPMC/CNRS, Christophe Caron
GUGGO, Cellule Bioinformatique-Ifremer Ifremer/UBO/CNRS, Laure Quintric
GUGGO, GenOuest IFB Core Facility, and BIPAA, Yvan Le Bras and Anthony Bretaudeau
Freiburg Galaxy Project
GUGGO, IRHS Bioinformatics team Angers University/Agrocampus Ouest/INRA, Jean-Pierre Renou
FMI Bioinformatics and Functional Genomics Core Facilities
GUGGO, BIRD IFB core facility Nantes University/INSERM, Audrey Bihouée
Galaxy @ PRABI
Institut Ruđer Bošković
South Green Platform
CRS4 - Centre for Advanced Studies, Research and Development in Sardinia
Satellite
Google

Galaxy Training Network launched In October 2014.

bit.ly/gxygtn

Galaxy Project: Further reading & Resources

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://wiki.galaxyproject.org/Cloud

http://bit.ly/gxychoices

Feedback: We need it!

bit.ly/ISMMS16

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton

Dave Clements

Nate Coraor

Carl Eberhard

Jeremy Goecks

Sam Guerler

Jen Jackson

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Nitesh Turaga

http://wiki.galaxyproject.org/GalaxyTeam

bit.ly/ISMMS16

Acknowledgements

You
Andrew Sharp
Stuart Scott

ISMMS
AWS

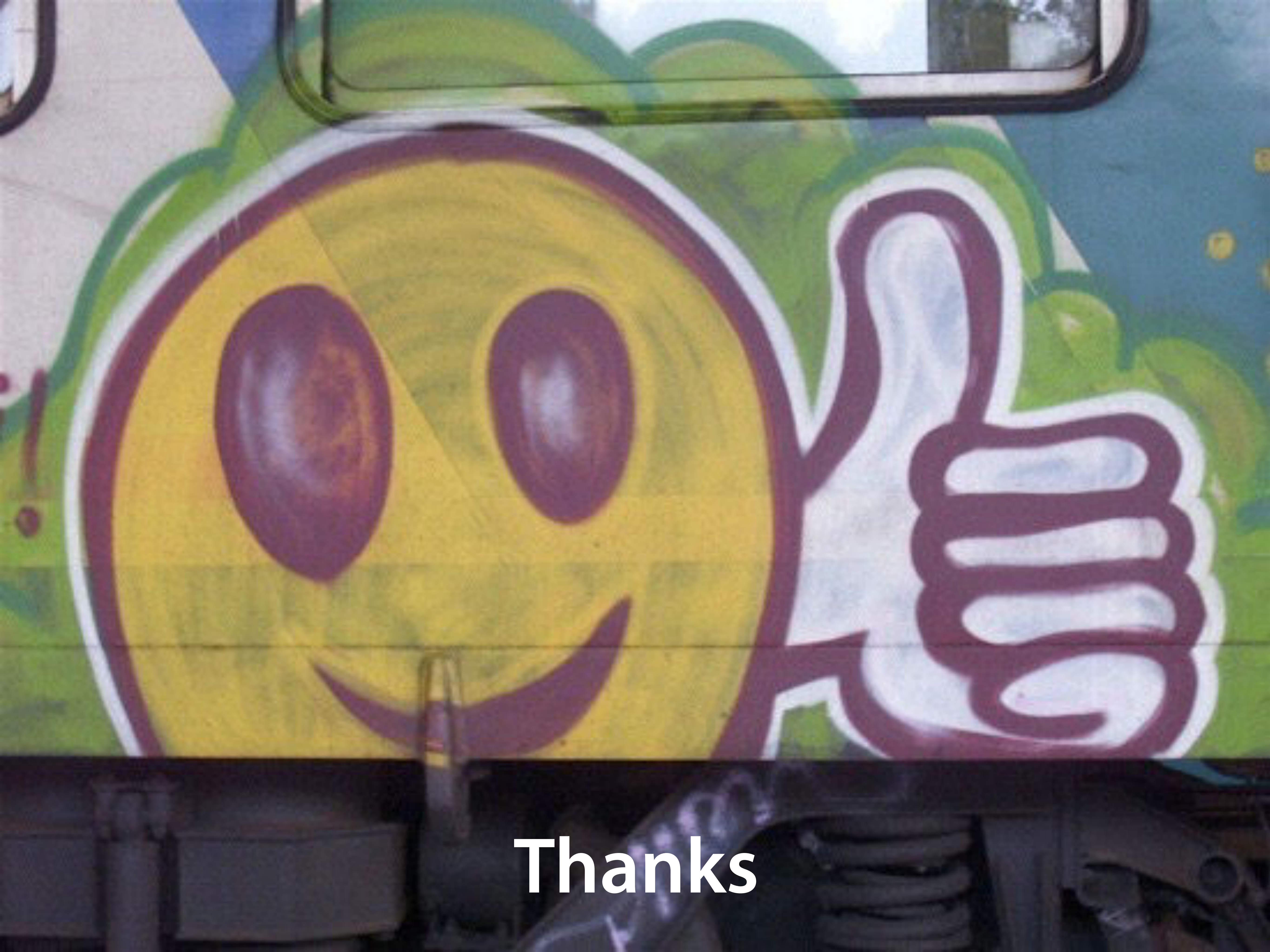
NIH
Johns Hopkins University
Penn State University

bit.ly/ISMMS16

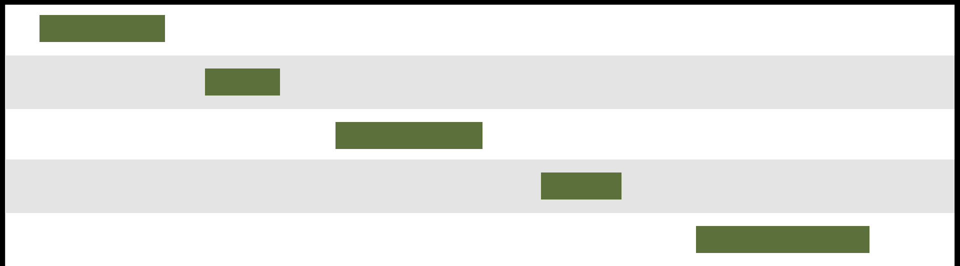
Agenda

9:00	Welcome
9:20	Basic Analysis with Galaxy A worked example demonstrating Galaxy Basics
10:45	Break
11:00	Basic Analysis into Reusable Workflows
12:20	Lunch (on your own)
1:20	RNA-Seq Analysis, Part I
2:50	Break
3:05	RNA-Seq Analysis, Part II
17:00	Done

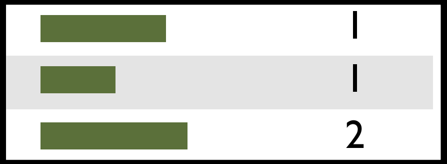
bit.ly/ISMMS16



Thanks



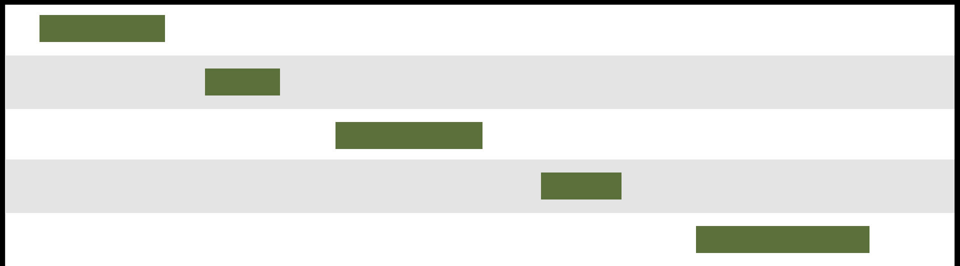
Exons



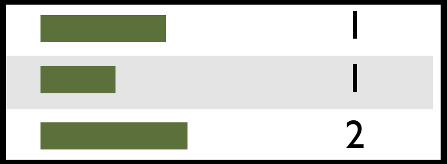
Exon overlap counts

We've answered our question, but we can do better.

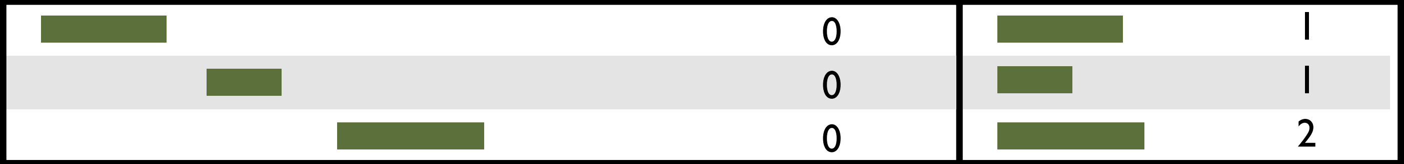
Incorporate the overlap count with rest of Exon information



Exons



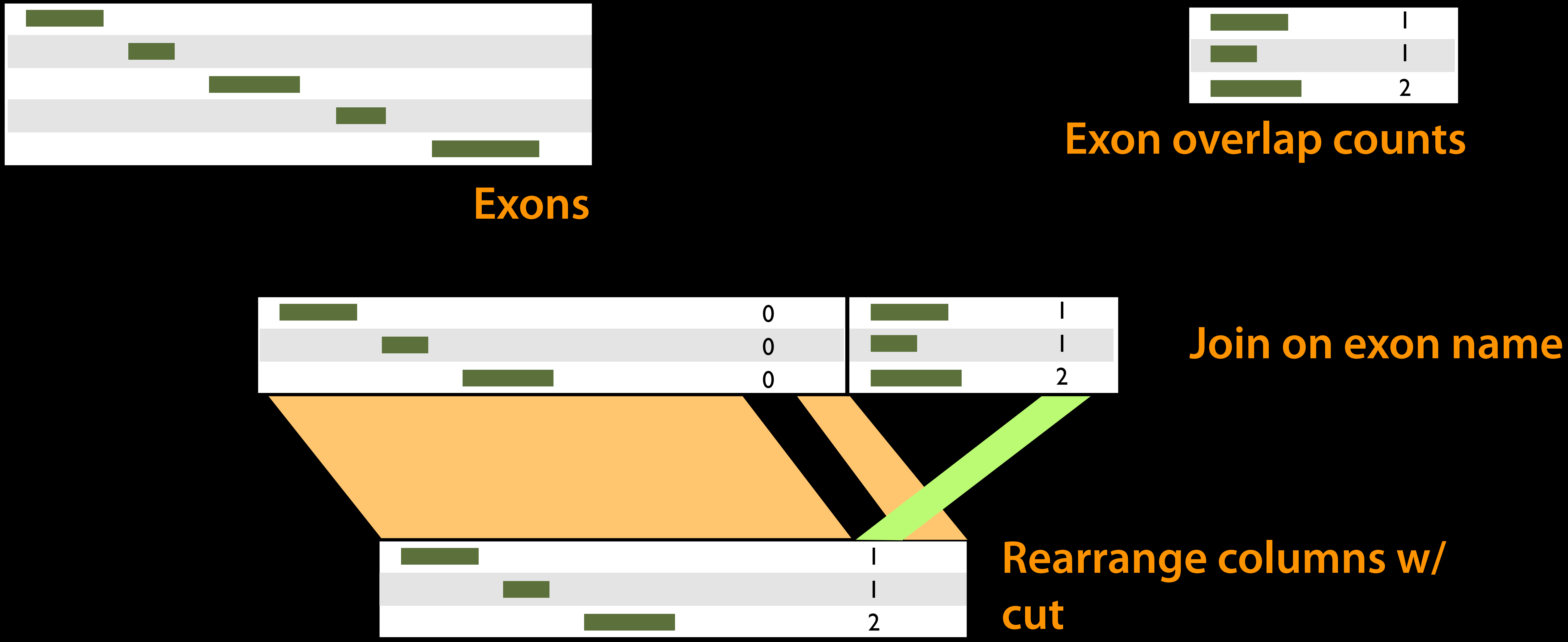
Exon overlap counts



Join on exon name

Join, Subtract, and Group → Join

(Incorporate the overlap count with rest of Exon information)



Exon overlap counts

Exons

Join on exon name

Rearrange columns w/ cut

Text Manipulation → Cut

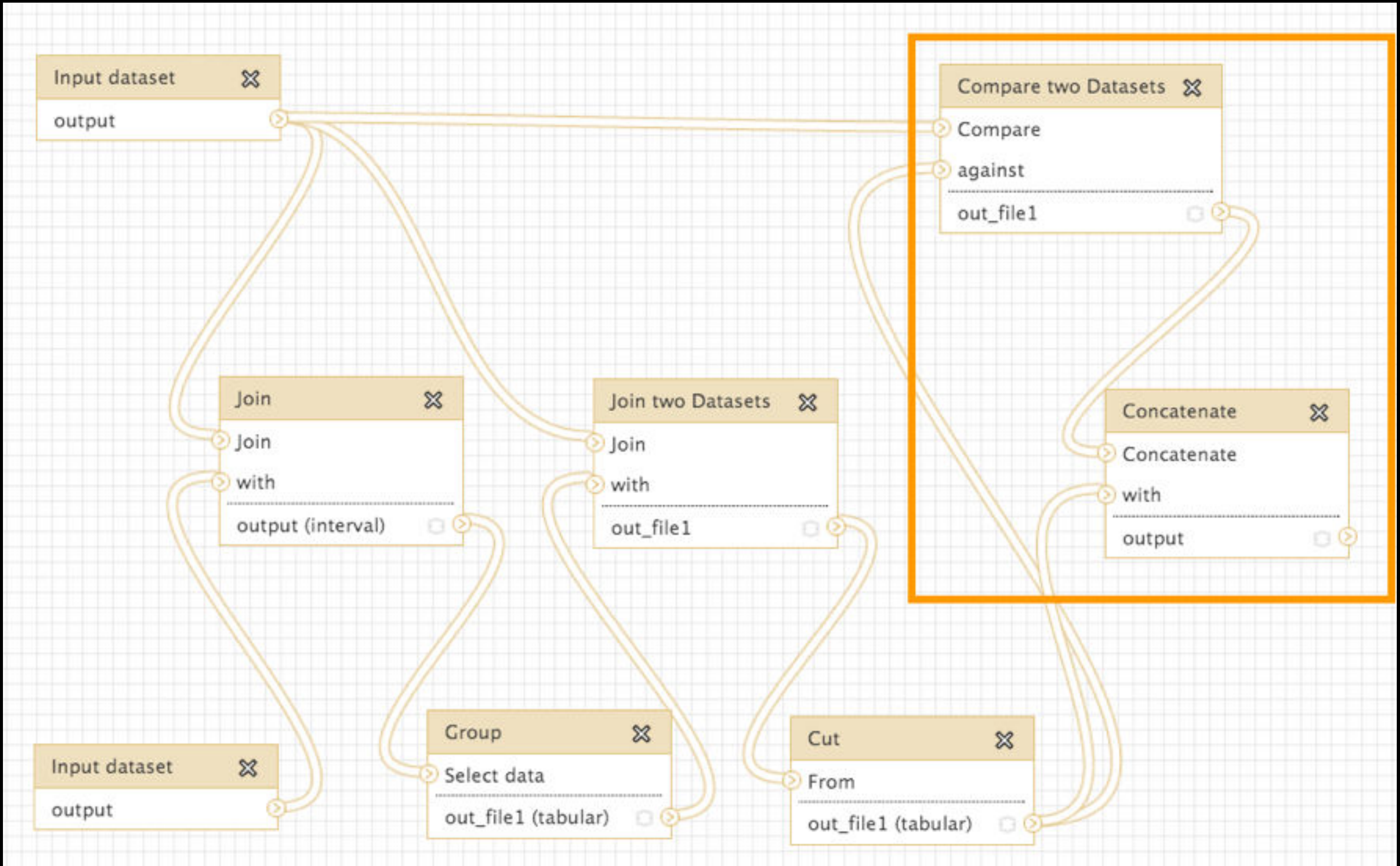
(Incorporate the overlap count with rest of Exon information)

Exons & Repeats: Exercise

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

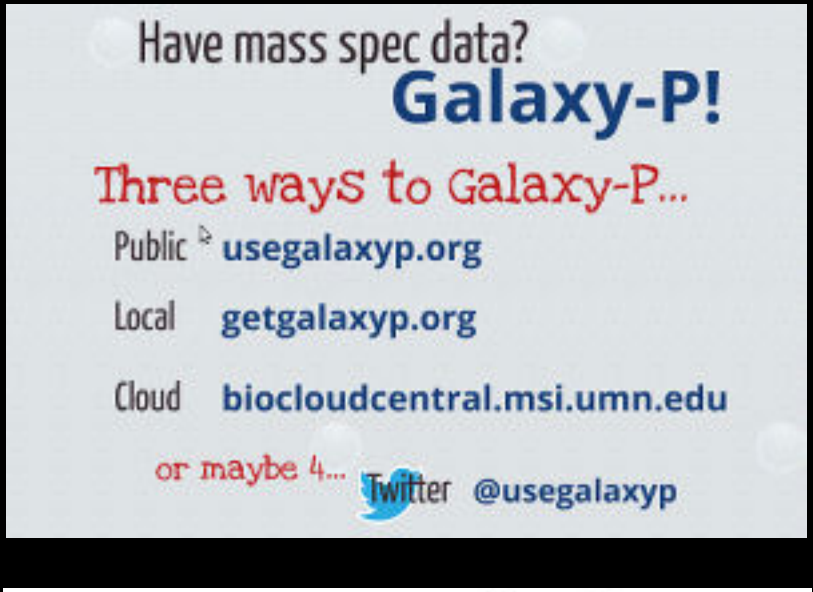
Everything you need will be in the toolboxes we used in the Exon-Repeats exercise.

One Possible Solution



Solution from Stanford Kwenda and Caron Griffiths, Pretoria.

Takes advantage of the fact that Exons already have 0 scores.



Have mass spec data?
Galaxy-P!

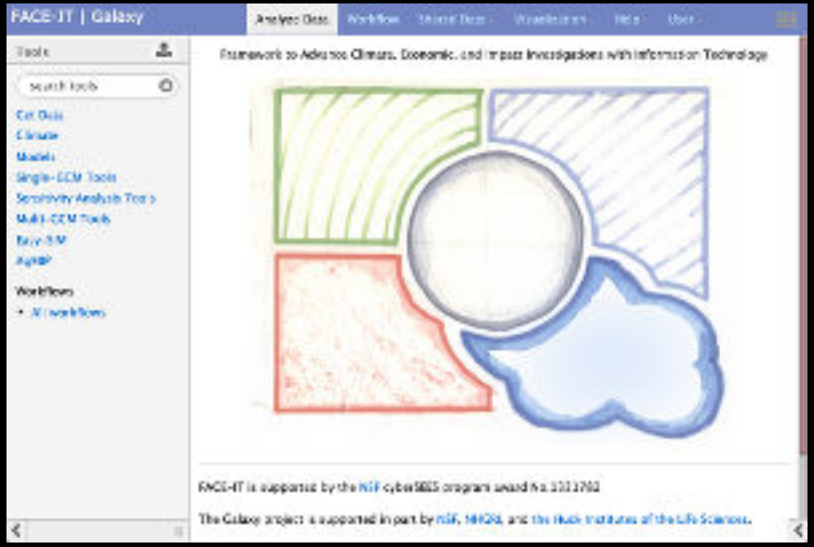
Three ways to Galaxy-P...

Public usegalaxyp.org

Local getgalaxyp.org

Cloud biocloudcentral.msi.umn.edu

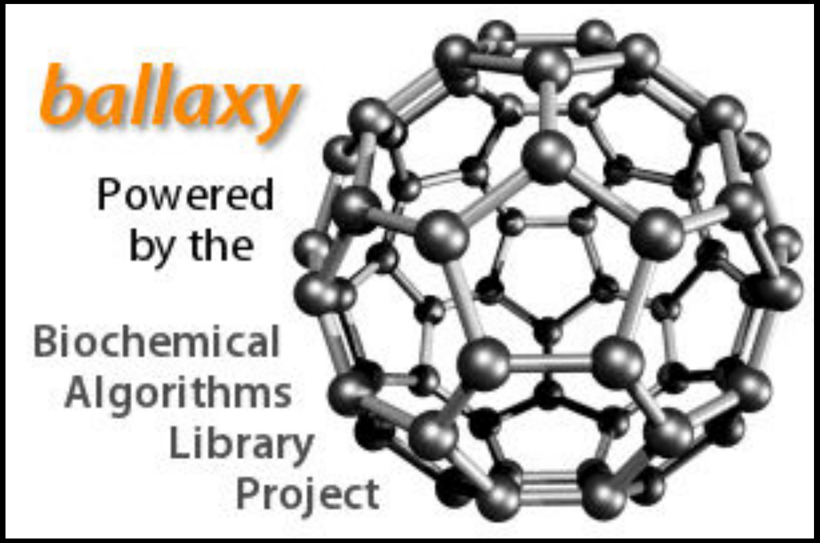
or maybe 4... Twitter @usegalaxyp

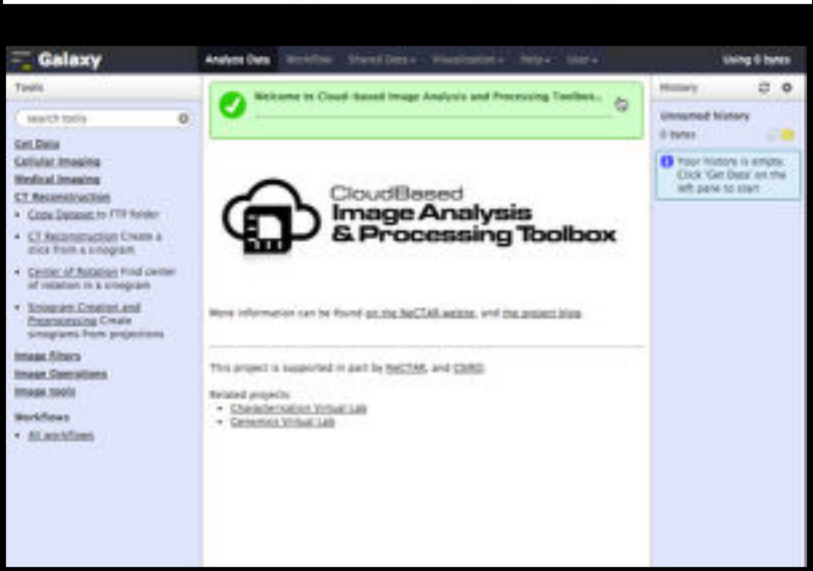


Framework to Advance Climate, Economic, and Impact Investigations with Information Technology

FACE-IT is supported by the NSF cyberSEES program award No. 1331782

The Galaxy project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.





Welcome to Cloud-based Image Analysis and Processing Toolbox...

- Copy Dataset to FTP folder
- CT Reconstruction Create a slice from a sinogram
- Center of Rotation Find center of rotation in a sinogram
- Sinogram Creation and Preprocessing Create sinograms from projections

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

More information can be found on the NeCTAR website, and the project blog.

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

Related projects:

- Characterisation Virtual Lab
- Genomics Virtual Lab

Proteomics
Metabolomics
Drug Discovery
Cosmology
Image Analysis
Social Science

Climate Change

Natural Language

Galaxy is hiring post-docs and software engineers at both Emory and Penn State.

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Local Galaxy Installs require a computational resource on which to be deployed

Control where tool execution happens

Galaxy works with DRMAA compliant cluster job schedulers (which is most of them).

Galaxy is just another client to your scheduler.

CLUSTER RESOURCES INC.
TORQUE

GRID ENGINE

Platform Computing

PBS Works™

DRMAA
Distributed Resource Management Application API — www.drmaa.org

Galaxy is available with Commercial Support

A ready-to-use appliance
(BioTeam)

Cloud-based solutions
(ABgenomica, AIS, GenomeCloud)

Consulting & Customization
(BioTeam, Deena Bioinformatics)

Training
(OpenHelix)