Galaxy Project Update

ISMB / ECCB 2015

Dublin, Ireland

Dave Clements
and the Galaxy Team
Johns Hopkins University

Multi-Sample Analysis

Need to scale to support current and coming experimental designs:

500 specimens

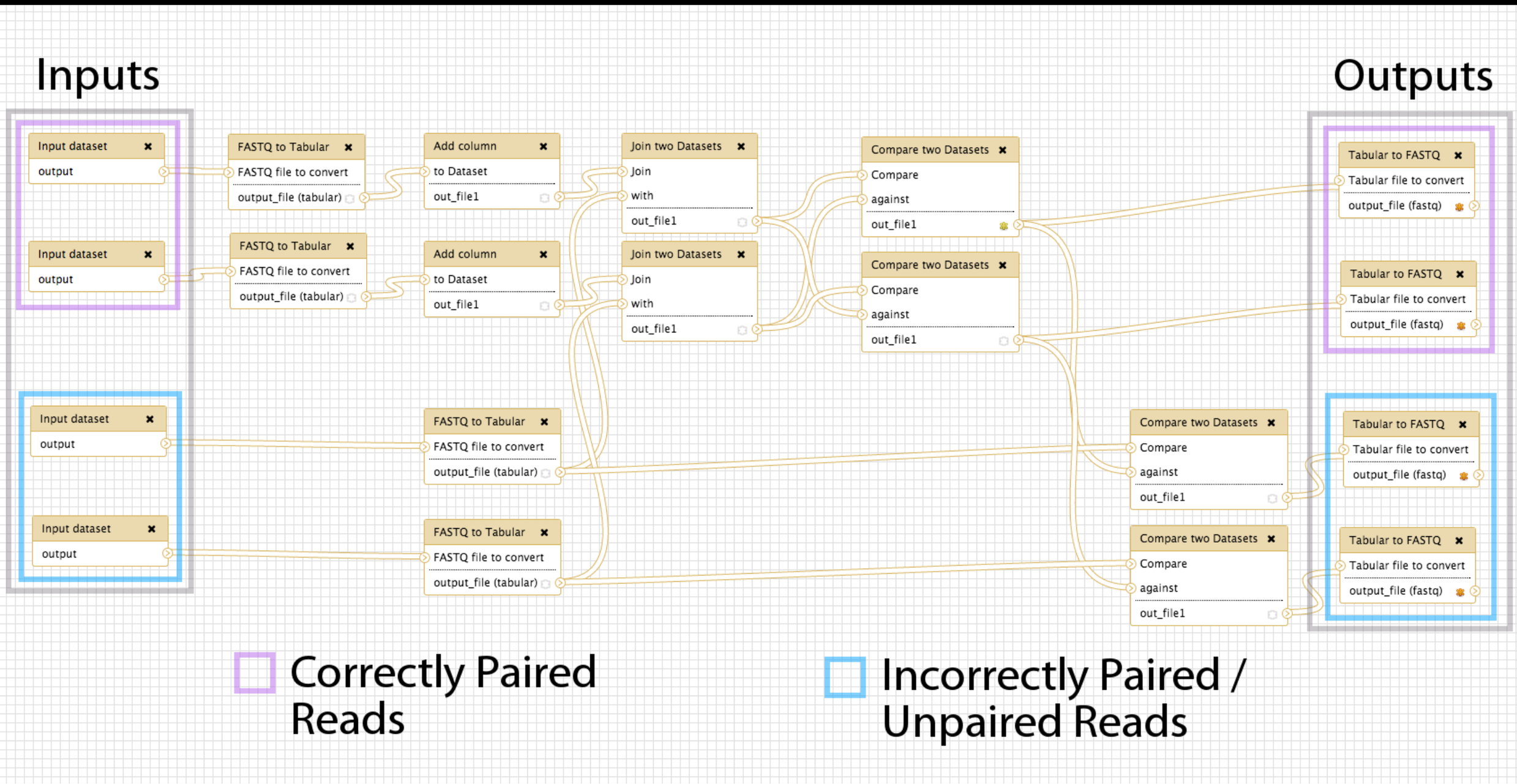
or

20 different conditions with 4 replicates each

or

...

Multi-Sample Analysis: Workflows



Multi-Sample:
Multiple dataset selection within a tool

FastQC Read Quality reports (Galaxy Tool Version 0.63) Versions Options

Short read data from your current history

1: MeOH_REP1_R1.fastq

Multiple datasets

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

Execute

FastQC Read Quality reports (Galaxy Tool Version 0.63) Versions Options

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

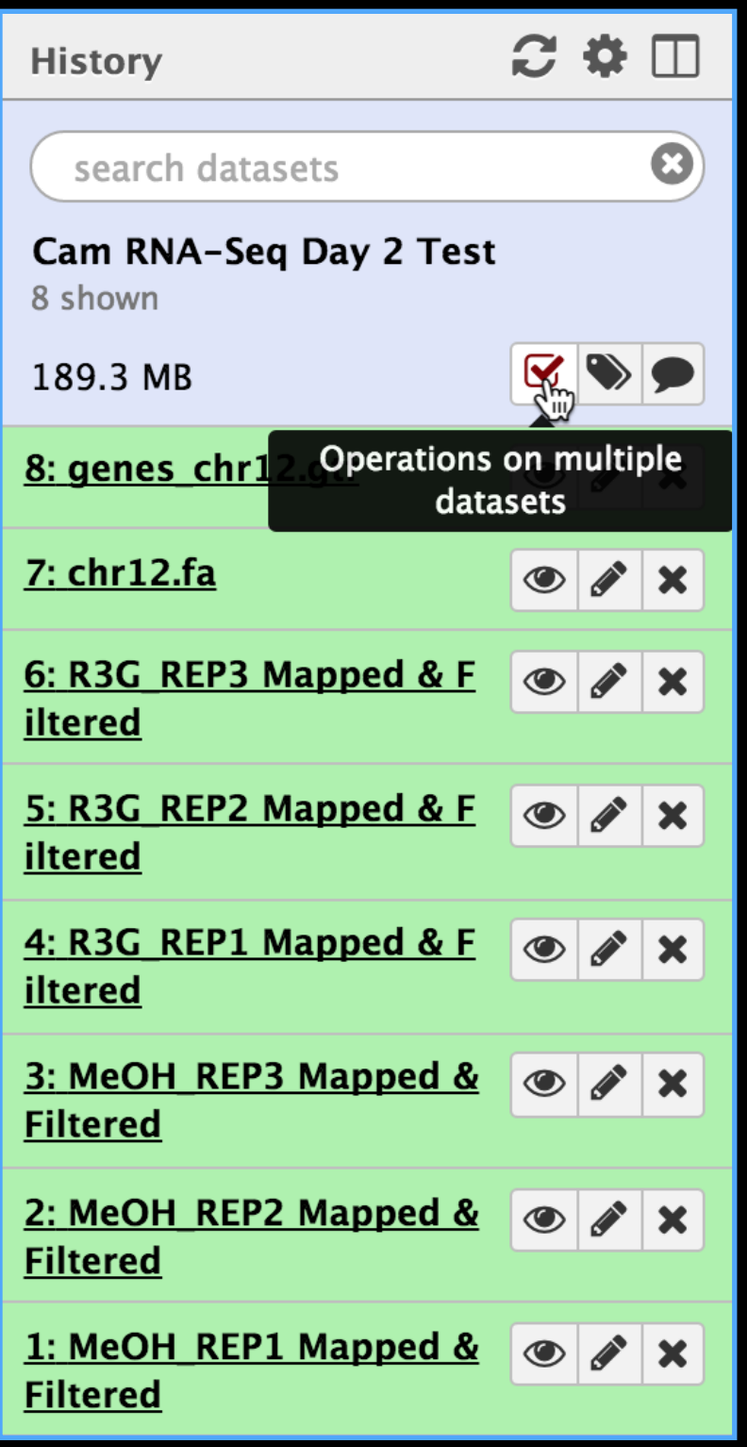
Submodule and Limit specifing file

No selection

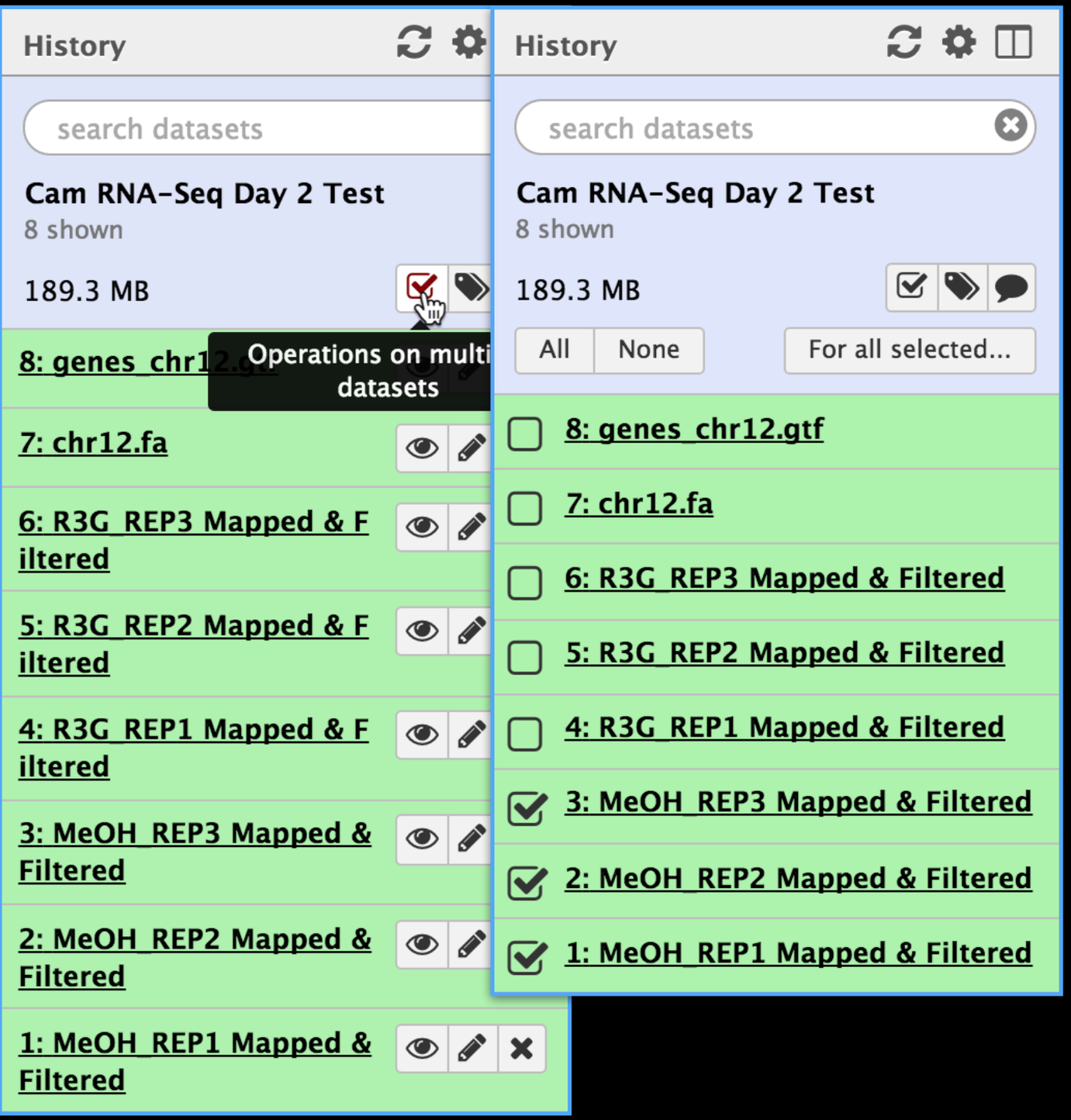
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Execute

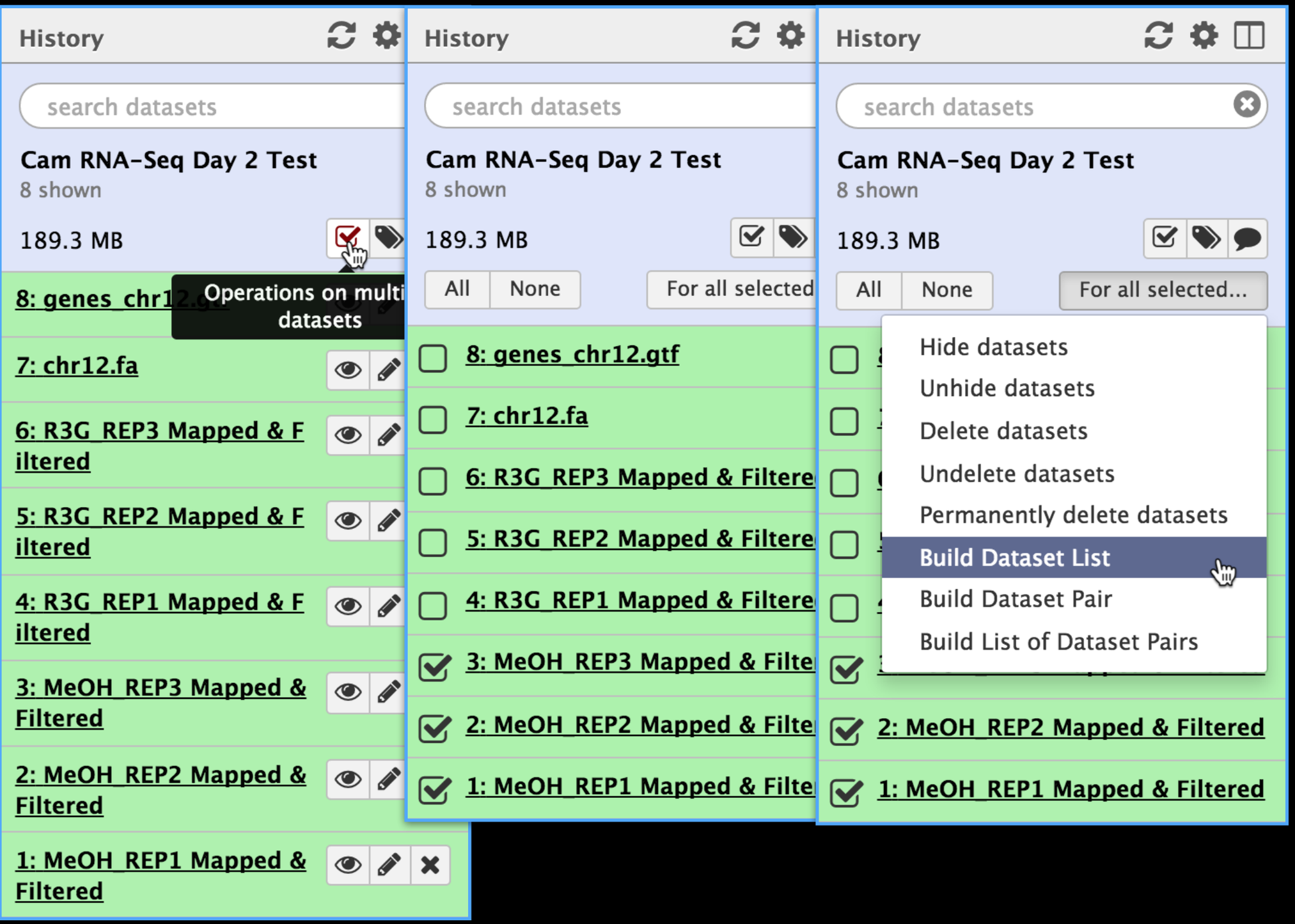
Multi-Sample Analysis: Collections



Multi-Sample Analysis: Collections



Multi-Sample Analysis: Collections



Multi-Sample Analysis: Collections

History

search datasets

Cam RNA-Seq Day 2 Test
8 shown

189.3 MB

Operations on multi datasets

8: genes_chr12.gtf

7: chr12.fa

6: R3G_REP3 Mapped & Filtered

5: R3G_REP2 Mapped & Filtered

4: R3G_REP1 Mapped & Filtered

3: MeOH_REP3 Mapped & Filtered

2: MeOH_REP2 Mapped & Filtered

1: MeOH_REP1 Mapped & Filtered

History

search datasets

Cam RNA-Seq Day 2 Test
8 shown

189.3 MB

All None For all selected

8: genes_chr12.gtf

7: chr12.fa

6: R3G_REP3 Mapped & Filtere

5: R3G_REP2 Mapped & Filtere

4: R3G_REP1 Mapped & Filtere

3: MeOH_REP3 Mapped & Filte

2: MeOH_REP2 Mapped & Filte

1: MeOH_REP1 Mapped & Filte

History

search datasets

Cam RNA-Seq Day 2 Test
8 shown

189.3 MB

All None For all sele

Hide datasets
Unhide datasets
Delete datasets
Undelete datasets
Permanently delete data
Build Dataset List
Build Dataset Pair
Build List of Dataset Pai

2: MeOH_REP2 Mapped &

1: MeOH_REP1 Mapped &

History

search datasets

Cam RNA-Seq Day 2 Test
10 shown

189.3 MB

10: R3G
a list of datasets

9: MeOH
a list of datasets

8: genes_chr12.gtf

7: chr12.fa

6: R3G_REP3 Mapped & Filtered

5: R3G_REP2 Mapped & Filtered

4: R3G_REP1 Mapped & Filtered

3: MeOH_REP3 Mapped & Filtered

2: MeOH_REP2 Mapped & Filtered

1: MeOH_REP1 Mapped & Filtered

Multi-Sample Analysis: Collections

StringTie transcript assembly and quantification (Galaxy Tool Version 1.0.3)

Options

Mapped reads to assemble transcripts from

3: MeOH_REP3 Mapped & Filtered

Dataset collection

Use ... assembly

Use GFF

Reference annotation to use for guiding the assembly

8: genes_chr12.gtf

-G

Perform abundance estimation only of input transcrip

Yes | No

-e

Output additional files for use in Ballgown

Yes | No

-b

Options

Use defaults

Job Resource Parameters

Use default job resource parameters

StringTie transcript assembly and quantification (Galaxy Tool Version 1.0.3)

Options

Mapped reads to assemble transcripts from

10: R3G

10: R3G

9: MeOH

Use GFF file to gu

Use GFF

Reference annotation to use for guiding the assembly process

8: genes_chr12.gtf

-G

Perform abundance estimation only of input transcripts

Yes | No

-e

Output additional files for use in Ballgown

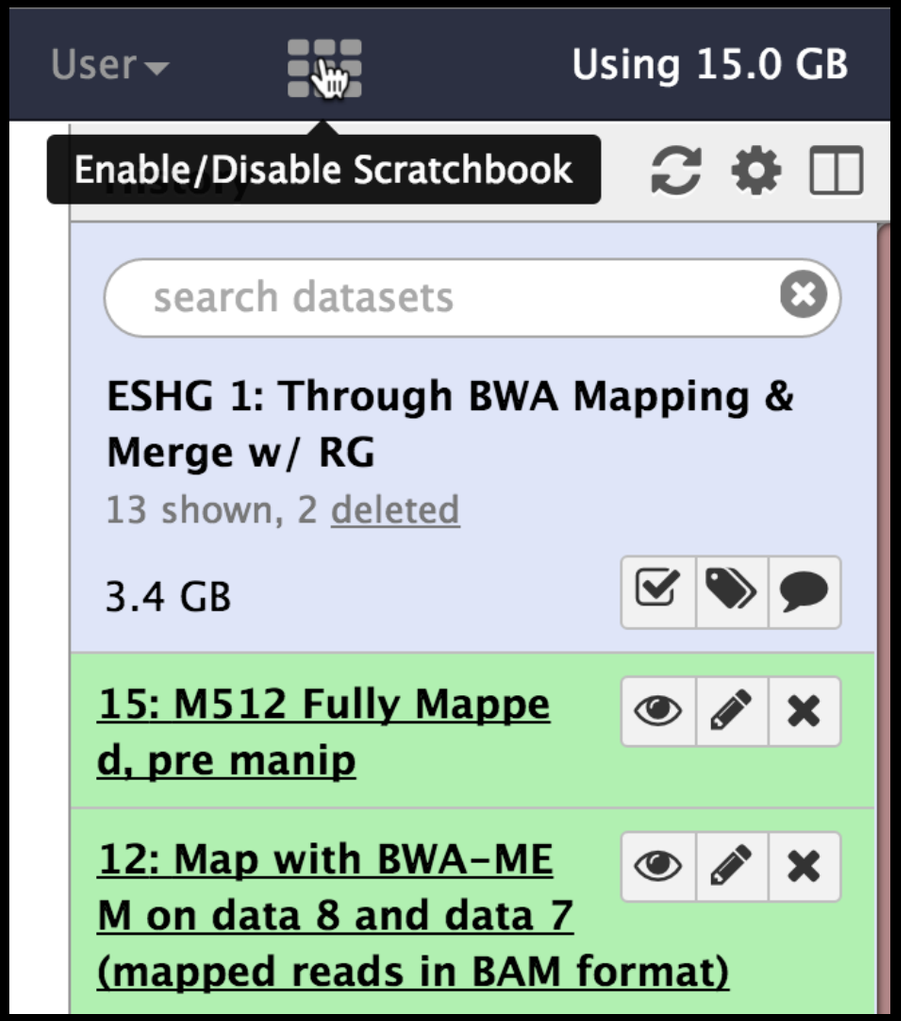
Yes | No

-b

Options

Use defaults

User Interface: Scratchbook



User Interface: Scratchbook

User Using 15.0 GB

Enable/Disable Scratchbook

search datasets

ESHG 1: Through BWA Mapping & Merge w/ RG

https://test.galaxyproject.org

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Admin | Help | wnk1 4 of 460

Significant

1	2	3	4	5
test_id	gene_id	gene	locus	samp
XLOC_000005	XLOC_000005	WNK1	chr12:862088-1020664	MeO
XLOC_000013	XLOC_000013	FKBP4	chr12:2904107-2914587	MeO
XLOC_000021	XLOC_000021	CCND2	chr12:4382901-4414522	MeO
XLOC_000023	XLOC_000023	RAD51AP1	chr12:4647949-4669213	MeO
XLOC_000035	XLOC_000035	TAPBPL	chr12:6561176-6579843	MeO
XLOC_000037	XLOC_000037	SCARNA10	chr12:6603297-6641132	MeO
XLOC_000038	XLOC_000038	GAPDH	chr12:6643584-6647537	MeO
XLOC_000039	XLOC_000039	COPS7A	chr12:6833149-6841041	MeO
XLOC_000042	XLOC_000042	CD4	chr12:6898637-6929976	MeO

Enough data to make a call

test_id	gene_id	gene	locus
XLOC_000001	XLOC_000001	LOC100288778	chr12:87983-91263
XLOC_000003	XLOC_000003	CCDC77	chr12:389222-551806
XLOC_000005	XLOC_000005	WNK1	chr12:862088-1020664
XLOC_000009	XLOC_000009	ADIPOR2	chr12:1800246-1897845
XLOC_000013	XLOC_000013	FKBP4	chr12:2904107-2914587
XLOC_000014	XLOC_000014	ITFG2	chr12:2921786-2934237
XLOC_000016	XLOC_000016	RHNO1	chr12:2945981-2998691
XLOC_000021	XLOC_000021	CCND2	chr12:4382901-4414522
XLOC_000023	XLOC_000023	RAD51AP1	chr12:4647949-4669213
XLOC_000025	XLOC_000025	NDUFA9	chr12:4758263-4796720

genes_chr12.gtf

-	.	gene_id "NINJ2"; ge
-	.	gene_id "NINJ2"; ge
-	.	gene_id "NINJ2"; ge
-	2	gene_id "NINJ2"; ge
-	0	gene_id "NINJ2"; ge
-	.	gene_id "NINJ2"; ge
-	0	gene_id "NINJ2"; ge
-	.	gene_id "NINJ2"; ge
-	.	gene_id "NINJ2"; ge
+	.	gene_id "WNK1"; g
+	.	gene_id "WNK1"; g
+	.	gene_id "WNK1"; g
+	.	gene_id "WNK1"; g
+	0	gene_id "WNK1"; g
+	0	gene_id "WNK1"; g
+	0	gene_id "WNK1"; g
+	0	gene_id "WNK1"; g
+	.	gene_id "WNK1"; g
+	.	gene_id "WNK1"; g
+	.	gene_id "WNK1"; g
+	.	gene_id "WNK1"; g
+	0	gene_id "WNK1"; g
+	0	gene_id "WNK1"; g

abase: hg19

'OK',

0 valid lines

3

gene

0001 LOC100288778

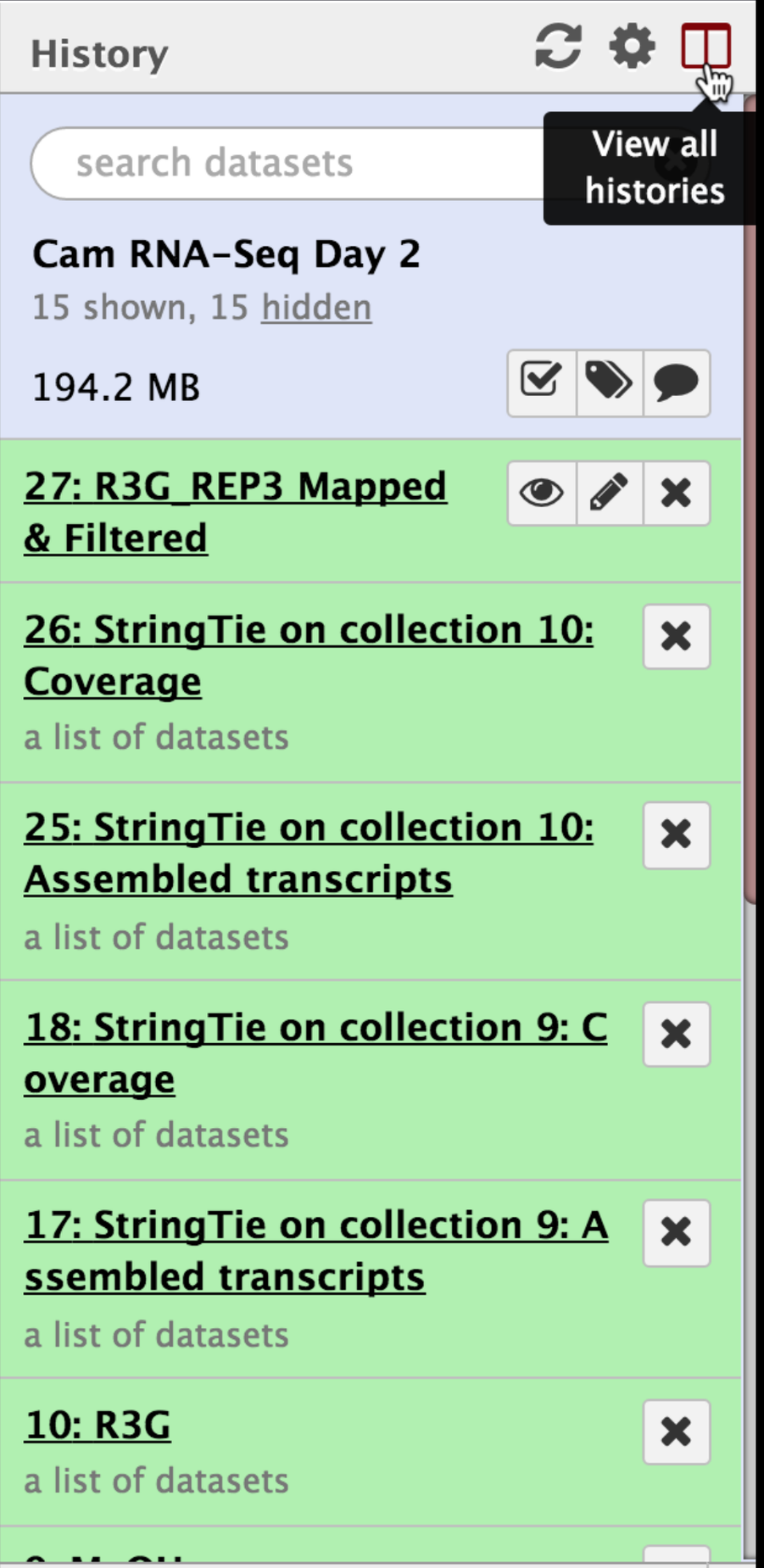
0003 CCDC77

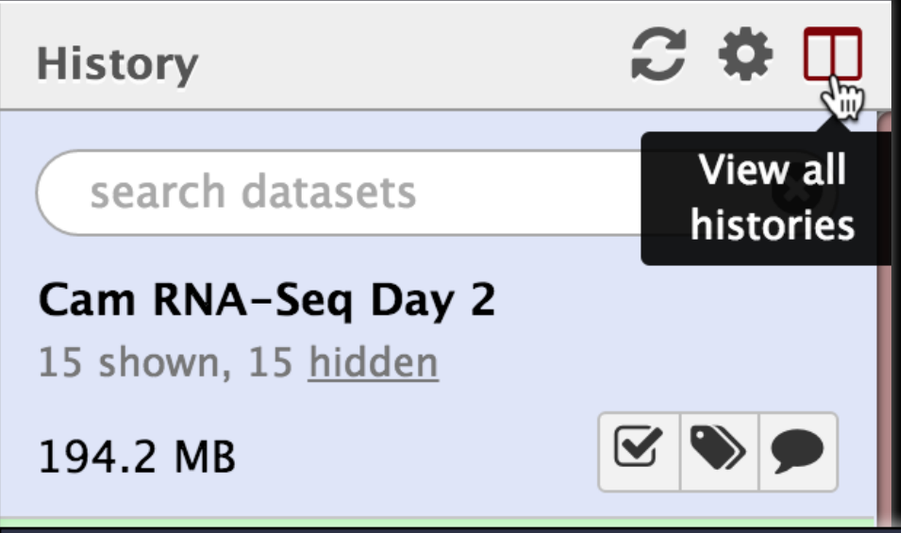
0005 WNK1

0009 ADIPOR2

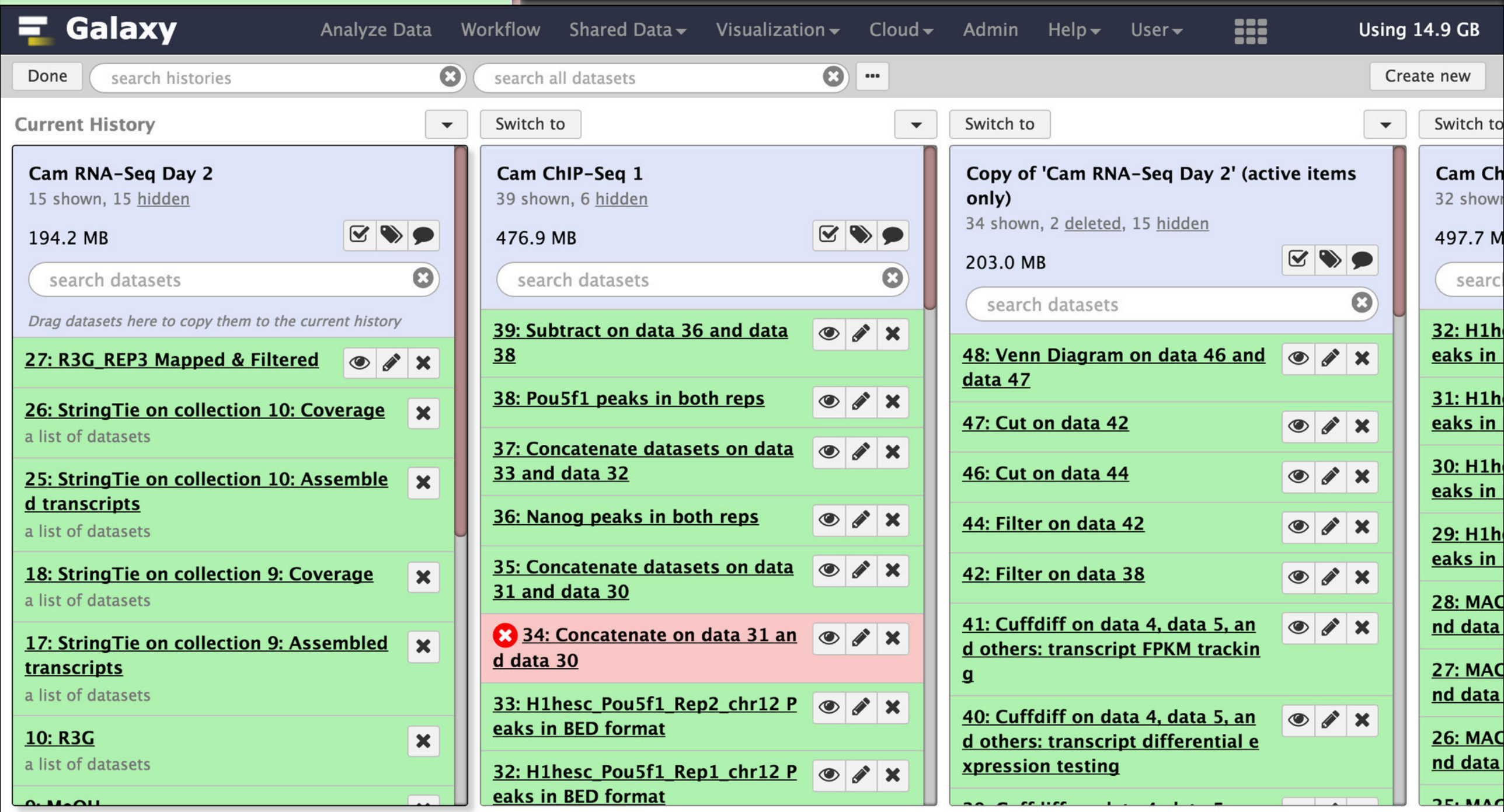
0013 FKBP4

User Interface: Histories

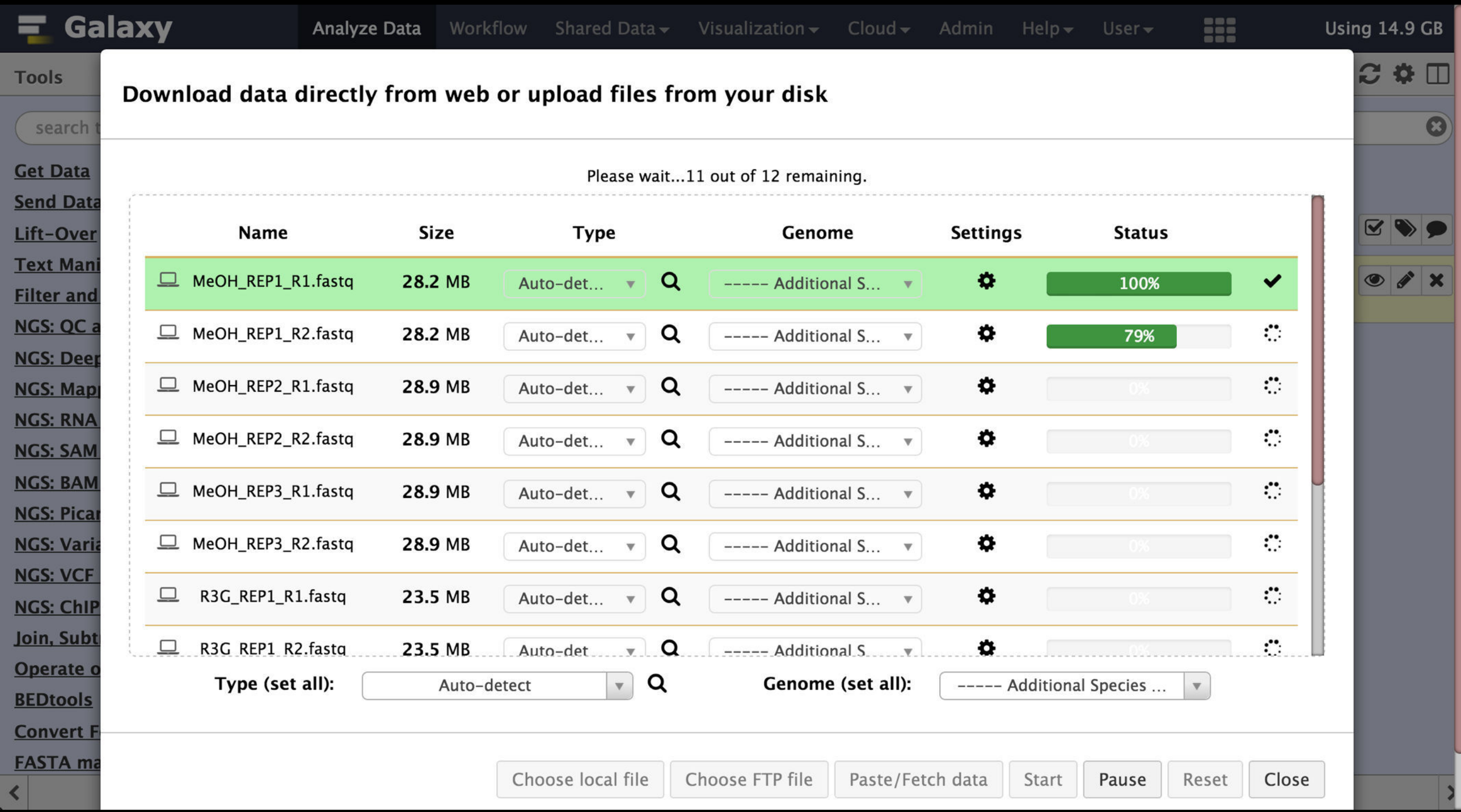




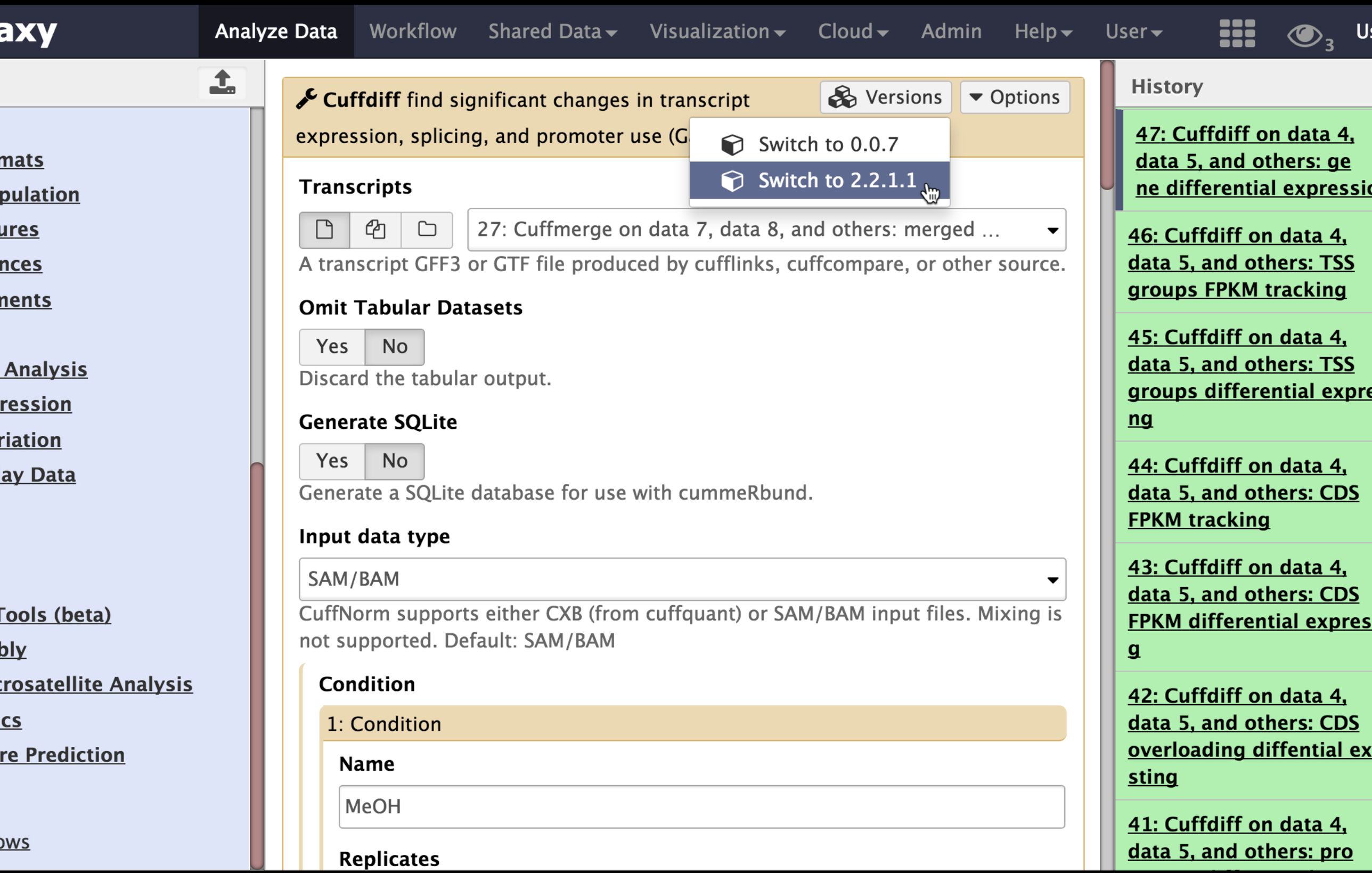
User Interface: Histories



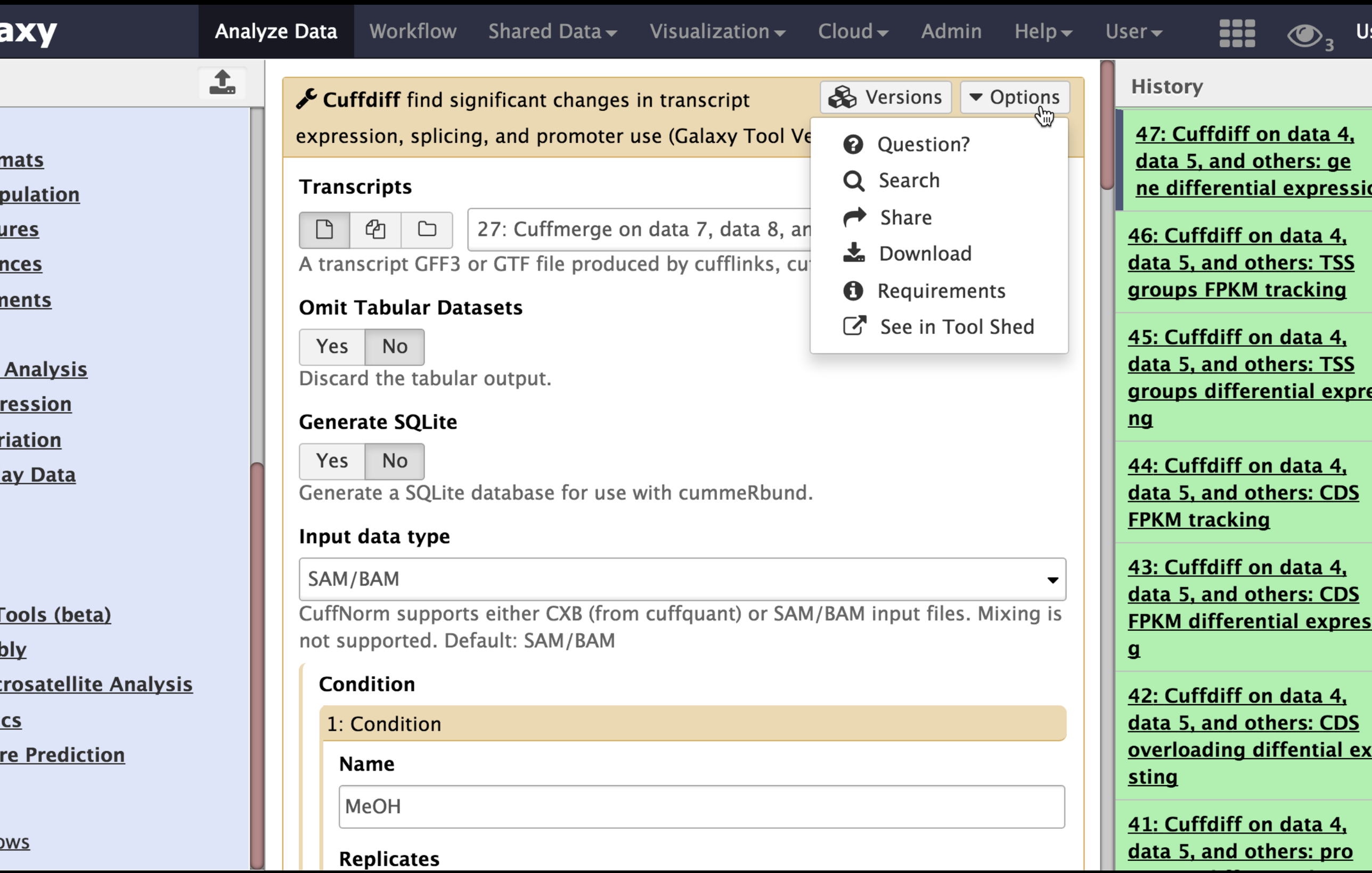
User Interface: Data Load



Tools: Multiple version support



Tool: Options



Tools: New Forms Engine

CuffNorm supports either CXB (from cuffquant) or SAM/BAM input files. Mixing is not supported. Default: SAM/BAM

Condition

1: Condition

Name

MeOH

Replicates

5: R3G_REP2 Mapped & Filtered
4: R3G_REP1 Mapped & Filtered
3: MeOH_REP3 Mapped & Filtered
2: MeOH_REP2 Mapped & Filtered
1: MeOH_REP1 Mapped & Filtered

2: Condition

Name

R3G

Replicates

6: R3G_REP3 Mapped & Filtered
5: R3G_REP2 Mapped & Filtered
4: R3G_REP1 Mapped & Filtered
3: MeOH_REP3 Mapped & Filtered
2: MeOH_REP2 Mapped & Filtered

Insert Condition

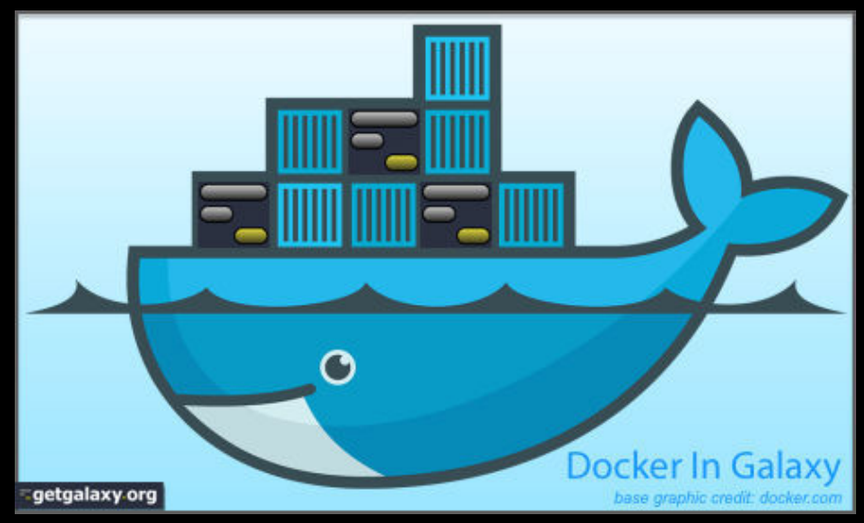
Library normalization method

History

47: Cuffdiff on data 4, data 5, and others: gene differential expressi
46: Cuffdiff on data 4, data 5, and others: TSS groups FPKM tracking
45: Cuffdiff on data 4, data 5, and others: TSS groups differential expr ng
44: Cuffdiff on data 4, data 5, and others: CDS FPKM tracking
43: Cuffdiff on data 4, data 5, and others: CDS FPKM differential expres g
42: Cuffdiff on data 4, data 5, and others: CDS overloading diffential ex sting
41: Cuffdiff on data 4, data 5, and others: pro

Using Docker for Tool Integration

Tools can now be run in Docker containers.



Tool authors may annotate Docker container ids the tool works with in the Tool XML file

Deployers may specify Docker container ids on a per destination basis.

https://wiki.galaxyproject.org/Admin/Tools/Docker

Interactive Environments

Interactive environments such as IPython and RStudio can now be invoked from Galaxy.

Work contributed by

Björn Grüning
University of Freiburg

Eric Rashe
Texas A&M University

Interactive Environments

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 93.7 MB

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
chr2L	dm3.ensGene.tbl	gene	20385293	20399394	.	-	.	Name=FBgn0032857;ID=FBgn0032
chr2L	dm3.ensGene.tbl	mRNA	20385293	20399394	.	-	.	ID=FBtr0081370;Parent=FBgn0032
chr2L	dm3.ensGene.tbl	exon	20399206	20399394	.	-	.	ID=exon:FBtr0081370:1;Parent=FBt
chr2L	dm3.ensGene.tbl	exon	20385293	20387592	.	-	.	Parent=FBtr0081370;ID=exon:FBtr0
chr2L	dm3.ensGene.tbl	five_prime_UTR	20399206	20399394	.	-	.	Parent=FBtr0081370;ID=five_prime_
chr2L	dm3.ensGene.tbl	five_prime_UTR	20387532	20387592	.	-	.	Parent=FBtr0081370;ID=five_prime_
chr2L	dm3.ensGene.tbl	start_codon	20387529	20387531	.	-	0	Parent=FBtr0081370;ID=start_codo
chr2L	dm3.ensGene.tbl	CDS	20386317	20387531	.	-	0	ID=CDS:FBtr0081370:1;Parent=FBt
chr2L	dm3.ensGene.tbl	stop_codon	20386317	20386319	.	-	0	ID=stop_codon:FBtr0081370:1;Pare
chr2L	dm3.ensGene.tbl	mRNA	20385293	20399394	.	-	.	Parent=FBgn0032857;ID=FBtr0081
chr2L	dm3.ensGene.tbl	exon	20399092	20399394	.	-	.	ID=exon:FBtr0081369:1;Parent=FBt
chr2L	dm3.ensGene.tbl	exon	20385293	20387592	.	-	.	ID=exon:FBtr0081369:2;Parent=FBt
chr2L	dm3.ensGene.tbl	five_prime_UTR	20399092	20399394	.	-	.	Parent=FBtr0081369;ID=five_prime_
chr2L	dm3.ensGene.tbl	five_prime_UTR	20387532	20387592	.	-	.	ID=five_prime_UTR:FBtr0081369:2;
chr2L	dm3.ensGene.tbl	start_codon	20387529	20387531	.	-	0	ID=start_codon:FBtr0081369:1;Pare
chr2L	dm3.ensGene.tbl	CDS	20386317	20387531	.	-	0	Parent=FBtr0081369;ID=CDS:FBtr0
chr2L	dm3.ensGene.tbl	stop_codon	20386317	20386319	.	-	0	ID=stop_codon:FBtr0081369:1;Pare
chr2L	dm3.ensGene.tbl	gene	21418852	21419260	.	+	.	Name=FBgn0053802;ID=FBgn0053
chr2L	dm3.ensGene.tbl	transcript	21418852	21419260	.	+	.	Parent=FBgn0053802;ID=FBtr0091
chr2L	dm3.ensGene.tbl	exon	21418852	21419260	.	+	.	ID=exon:FBtr0091806:1;Parent=FBt
chr2L	dm3.ensGene.tbl	gene	6036997	6037902	.	-	.	ID=FBgn0031766;Name=FBgn0031
chr2L	dm3.ensGene.tbl	mRNA	6036997	6037902	.	-	.	ID=FBtr0079222;Parent=FBgn0031
chr2L	dm3.ensGene.tbl	five_prime_UTR	6037816	6037902	.	-	.	Parent=FBtr0079222;ID=five_prime_
chr2L	dm3.ensGene.tbl	start_codon	6037813	6037815	.	-	0	ID=start_codon:FBtr0079222:1;Pare
chr2L	dm3.ensGene.tbl	CDS	6037129	6037815	.	-	0	Parent=FBtr0079222;ID=CDS:FBtr0
chr2L	dm3.ensGene.tbl	stop_codon	6037129	6037131	.	-	0	ID=stop_codon:FBtr0079222:1;Pare
chr2L	dm3.ensGene.tbl	exon	6036997	6037902	.	-	.	ID=exon:FBtr0079222:1;Parent=FBt
chr2L	dm3.ensGene.tbl	gene	11083459	11085243	.	-	.	Name=FBgn0032337;ID=FBgn0032
chr2L	dm3.ensGene.tbl	mRNA	11083459	11085243	.	-	.	Parent=FBgn0032337;ID=FBtr0114
chr2L	dm3.ensGene.tbl	five_prime_UTR	11085171	11085243	.	-	.	Parent=FBtr0114455;ID=five_prime_
chr2L	dm3.ensGene.tbl	start_codon	11085168	11085170	.	-	0	ID=start_codon:FBtr0114455:1;Pare
chr2L	dm3.ensGene.tbl	exon	11084811	11085243	.	-	.	ID=exon:FBtr0114455:1;Parent=FBt
chr2L	dm3.ensGene.tbl	exon	11084167	11084347	.	-	.	ID=exon:FBtr0114455:2;Parent=FBt

History

search datasets

Screencast

4 shown, 2 deleted

46.9 MB

4: misopy.ipynb

IPython Notebook

format: **ipynb**, database: ?

uploaded ipynb file

IPython

RStudio

```
{
 "metadata": {
  "kernelspec": {
   "display_name": "Python 2",
   "language": "python",
   "name": "python2"
```

3: dm3.ensGene.gff3

~460,000 lines

format: **gff3**, database: ?

uploaded gff3 file

display with IGV local

1.Seqid	2.Source	3.Type
##gff-version 3		
chr2L	dm3.ensGene.tbl	gene
chr2L	dm3.ensGene.tbl	mRNA

Interactive Environments

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 93.7 MB

IP[y]: Notebook ipython_galaxy_notebook (autosaved)

File Edit View Insert Cell Kernel Help | Saving every 120s

Markdown | Cell Toolbar: None

Install Misopy

In [1]:
```
!pip install --user --quiet misopy
```

Restart the kernel (Kernel -> Restart), so that IPython finds Misopy

Get Samtools

In [1]:
```
from urllib import urlretrieve
urlretrieve("http://depot.galaxyproject.org/package/linux/x86_64/samtools
!tar -xf samtools.tgz
```

The samtools binary should now be in bin/samtools

Generate Indices of the bam Files

- Now the bam files from the Galaxy history can be imported with get(history_id)
- Then index files of the bam files can be generated via samtools

In [2]:
```
get(5)
get(6)
!mv 5 461177.bam
!mv 6 461178.bam

!bin/samtools index 461177.bam
!bin/samtools index 461178.bam
```

History

search datasets

Screencast

4 shown, 2 deleted

46.9 MB

4: misopy.ipynb

IPython Notebook

format: **ipynb**, database: ?

uploaded ipynb file


```
{
"metadata": {
 "kernelspec": {
  "display_name": "Python 2",
  "language": "python",
  "name": "python2"
```


3: dm3.ensGene.gff3

~460,000 lines

format: **gff3**, database: ?

uploaded gff3 file

display with IGV local


```
1.Seqid 2.Source          3.Type
##gff-version 3
chr2L   dm3.ensGene.tbl gene
chr2L   dm3.ensGene.tbl mRNA
```

Interactive Environments

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 93.7 MB

IP[y]: Notebook ipython_galaxy_notebook (unsaved changes)

File Edit View Insert Cell Kernel Help

Code | Cell Toolbar: None

The samtools binary should now be in bin/samtools

Generate Indices of the bam Files

- Now the bam files from the Galaxy history can be imported with get(history_id)
- Then index files of the bam files can be generated via samtools

In [2]:
```
get(1)
get(2)
!mv 1 461177.bam
!mv 2 461178.bam

!bin/samtools index 461177.bam
!bin/samtools index 461178.bam
```

If everything went well there should now be index files (*.bai) next to the bam files.

In [3]:
```
!ls *.ba*
```

```
461177.bam  461177.bam.bai  461178.bam  461178.bam.bai
```

Generate Index of an Annotation File

An annotation file is needed.

In [4]:
```
get(10)
!mv 10 dm3.ensGene.gff3
```

History

search datasets

Screencast

4 shown, 2 deleted

46.9 MB

4: misopy.ipynb

3: dm3.ensGene.gff3

2: 461178.bam

2.4 MB

format: **bam**, database: ?

uploaded bam file

display with IGV local

display in IGB View

Binary bam alignments file

1: 461177.bam

Interactive Environments

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 93.7 MB

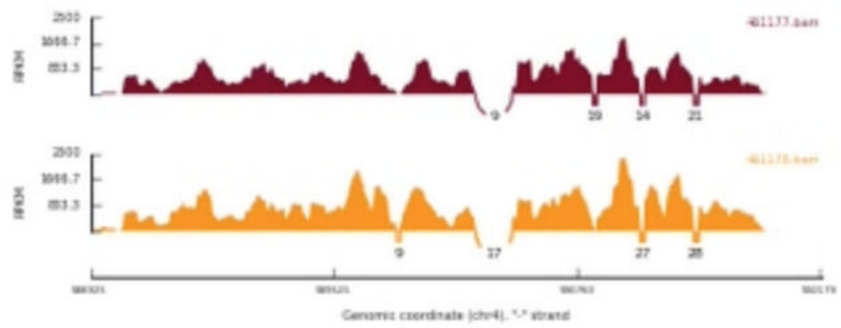
IP[y]: Notebook ipython_galaxy_notebook (unsaved changes)

File Edit View Insert Cell Kernel Help

Code | Cell Toolbar: None

```
Skipping read with CIGAR 35M65N2M
Skipping read with CIGAR 35M65N2M
Skipping read with CIGAR 2M70N35M
Saving plot to: /import/output/FBgn0039909.pdf
```

FBgn0039909



Finally save the pdf back to the Galaxy history

```
In [9]: put("output/FBgn0039909.pdf")
```

History

search datasets

Screencast

5 shown, 2 deleted

46.9 MB

7: FBgn0039909.pdf

This job is currently running

4: misopy.ipynb

3: dm3.ensGene.gff3

2: 461178.bam

2.4 MB

format: **bam**, database: ?

uploaded bam file

display with IGV local

display in IGB View

Binary bam alignments file

1: 461177.bam

Interactive Environments

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 93.7 MB

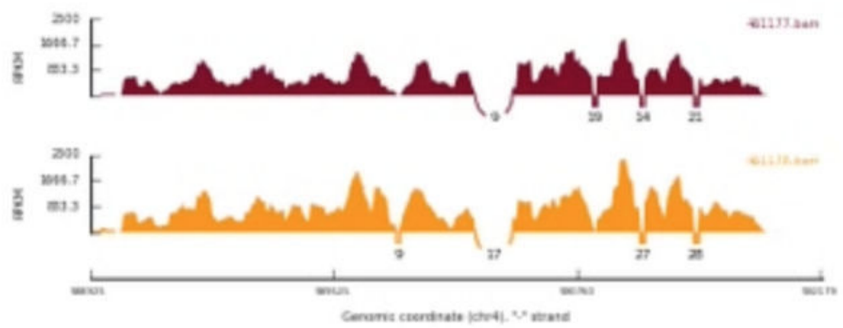
IP[y]: Notebook ipython_galaxy_notebook (autosaved)

File Edit View Insert Cell Kernel Help

Checkpoint created: 13:04:48

Code | Cell Toolbar: None

```
Skipping read with CIGAR 35M65N2M
Skipping read with CIGAR 35M65N2M
Skipping read with CIGAR 2M70N35M
Saving plot to: /import/output/FBgn0039909.pdf
```



Finally save the pdf back to the Galaxy history

```
In [9]: put("output/FBgn0039909.pdf")
```

History

search datasets

Screencast

5 shown, 2 deleted

47.0 MB

7: FBgn0039909.pdf

36.0 KB

format: **pdf**, database: ?

uploaded pdf file

Image in pdf format

4: misopy.ipynb

3: dm3.ensGene.gff3

2: 461178.bam

2.4 MB

format: **bam**, database: ?

uploaded bam file

display with IGV local

display in IGB View

Binary bam alignments file

1: 461177.bam

Interactive Environments

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 93.7 MB

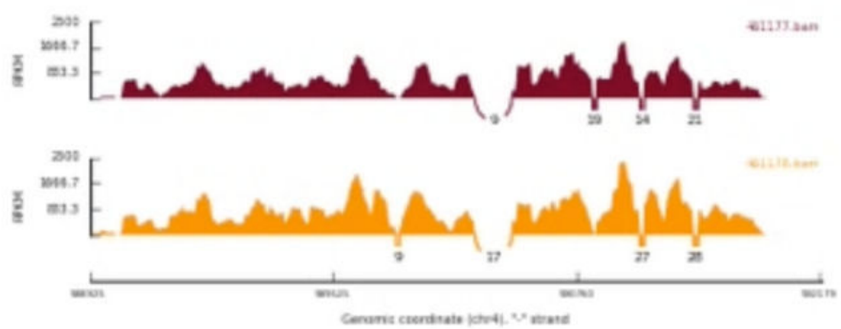
IP[y]: Notebook ipython_galaxy_notebook (autosaved)

File Edit View Insert Cell Kernel Help

Code | Cell Toolbar: None

```
Skipping read with CIGAR 35M65N2M
Skipping read with CIGAR 35M65N2M
Skipping read with CIGAR 2M70N35M
Saving plot to: /import/output/FBgn0039909.pdf
```

FBgn0039909



Genomic coordinate (chr4), "-" strand

Finally save the pdf back to the Galaxy history

```
In [9]: put("output/FBgn0039909.pdf")
```

History

search datasets

Screencast

6 shown, 2 deleted

47.0 MB

8: ipython_galaxy_notebook.ipynb

View data

IPython Notebook

format: ipynb, database: ?

uploaded ipynb file

```
{
 "metadata": {
  "kernelspec": {
   "display_name": "Python 2",
   "language": "python",
   "name": "python2"
```

7: FBgn0039909.pdf

4: misopy.ipynb

3: dm3.ensGene.gff3

2: 461178.bam

1: 461177.bam

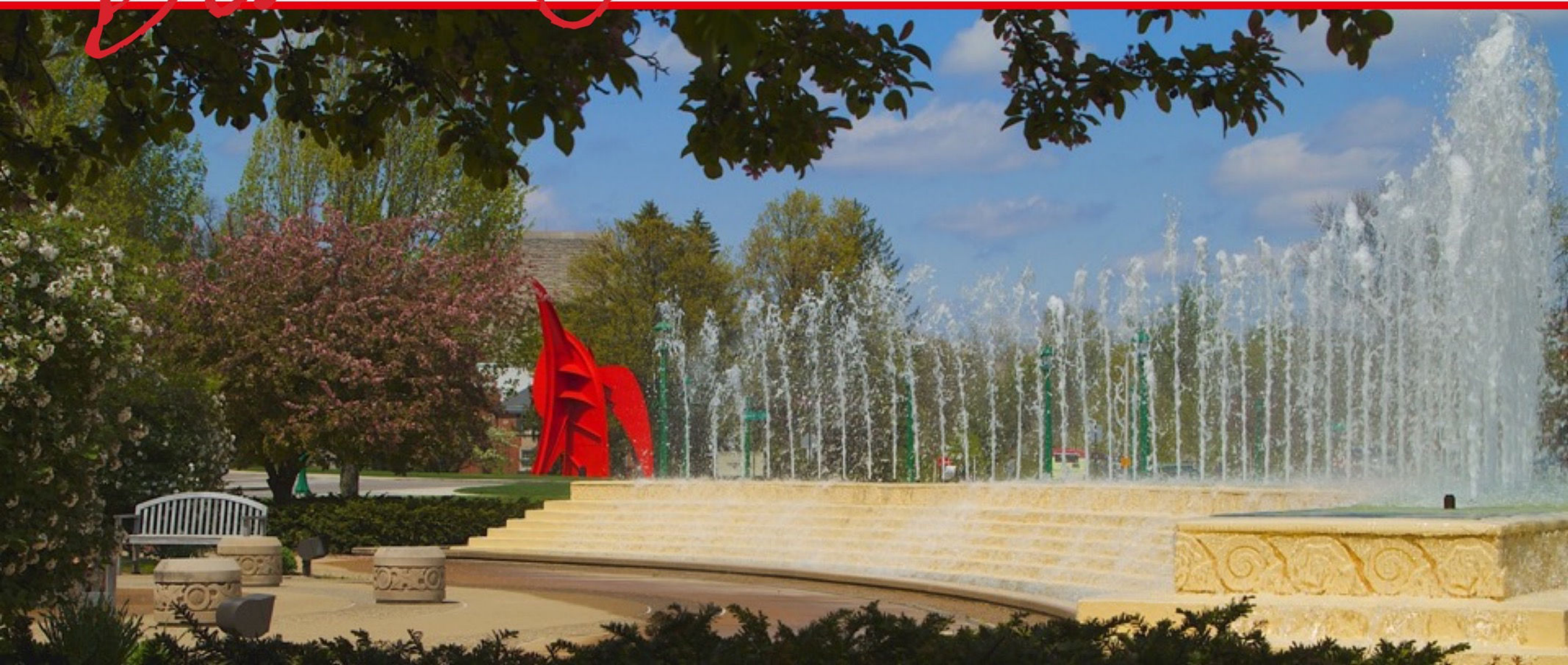
GALAXY Community Conference

Hosted by Indiana University June 28–29, 2016

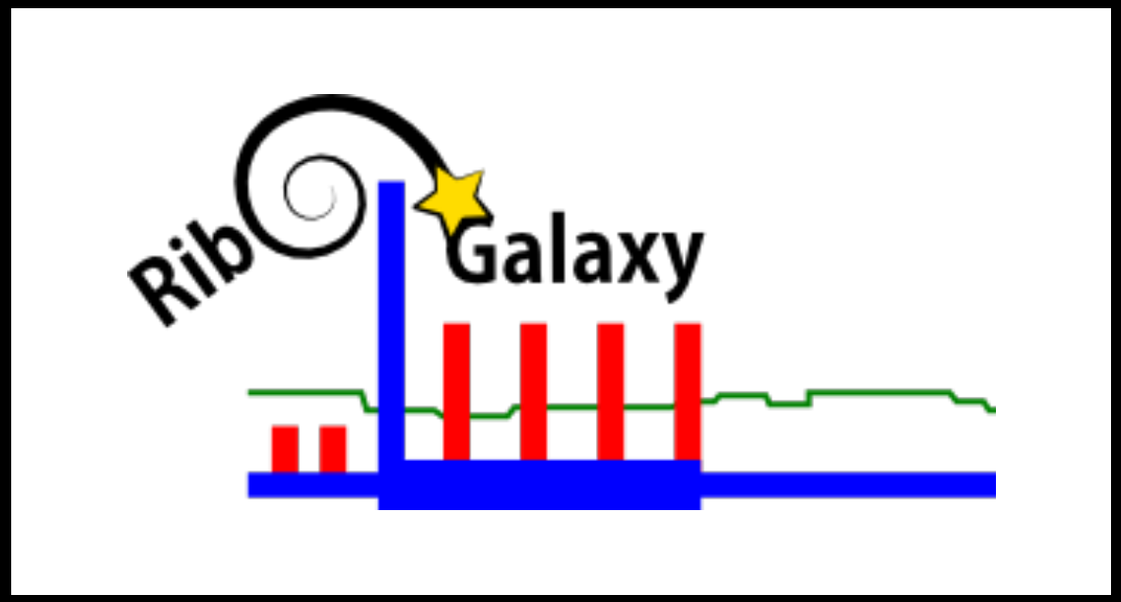
Join us in beautiful

Bloomington, Indiana

June 28-29, 2016



What's coming: *Right here! Right now!*



TT16: RiboGalaxy: a platform for the alignment, analysis and visualization of ribo-seq data

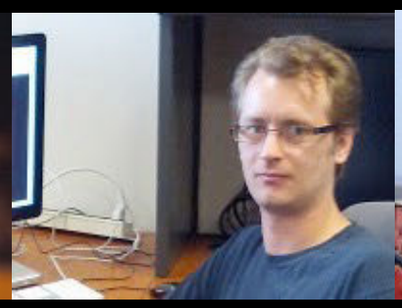
Audrey Michel

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

http://bit.ly/GxyCamFeedback

Dave Clements

Galaxy Project

Johns Hopkins University

clements@galaxyproject.org

Thanks

Scalability ...