http://bit.ly/GxyCam2015

Introduction to Galaxy: RNA-seq & ChIP-seq Data Analysis

University of Cambridge
11-12 June 2015

Dave Clements
Galaxy Project
Johns Hopkins University

http://bit.ly/GxyCam2015

Agenda: Day 1

9:30 Welcome

10:00 Basic Analysis with Galaxy
A worked example demonstrating Galaxy Basics

10:50 Break

11:10 Basic Analysis (continued)

12:30 Lunch (on your own)

13:30 RNA-Seq Analysis
Quality Control and Galaxy Workflows

15:00 Break

15:20 RNA-Seq Analysis
Mapping and Splice Junction Identification

17:00 Done

http://bit.ly/GxyCam2015

Goals

Provide a basic introduction to using Galaxy for bioinformatic analysis.

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

Not Goals

This workshop will *not* cover

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

This workshop does cover ChIP-Seq and RNA-Seq but you won't be an expert at either of these 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

Galaxy is available online, for free

http://usegalaxy.org

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

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

Galaxy is available as Open Source Software

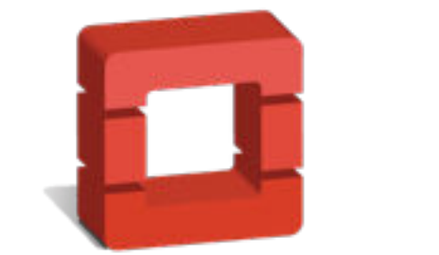
Galaxy is installed in locations around the world.

Some of them are free for anyone to use too.

http://getgalaxy.org

bit.ly/gxyServers

Galaxy is available on the Cloud



http://aws.amazon.com/education

http://globus.org/

http://wiki.galaxyproject.org/Cloud

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)

SLIPSTREAM APPLIANCE GALAXY EDITION
Galaxy made easy.

abgenomica

AIS ENTERPRISE-CLASS DATA CENTER SERVICES

GenomeCloud

BIOTEAM

OpenHelix™

Galaxy Project: Further reading & Resources

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://wiki.galaxyproject.org/Cloud

http://bit.ly/gxychoices

Agenda: Day 1

9:30 Welcome

10:00 **Basic Analysis with Galaxy**
A worked example demonstrating Galaxy Basics

10:50 Break

11:10 Basic Analysis (continued)

12:30 Lunch (on your own)

13:30 **RNA-Seq Analysis**
Quality Control and Galaxy Workflows

15:00 Break

15:20 **RNA-Seq Analysis**
Mapping and Splice Junction Identification

17:00 Done

http://bit.ly/GxyCam2015

Basic Analysis

Which exons have most overlapping Repeats?

Use Human, HG38, Chromosome 22

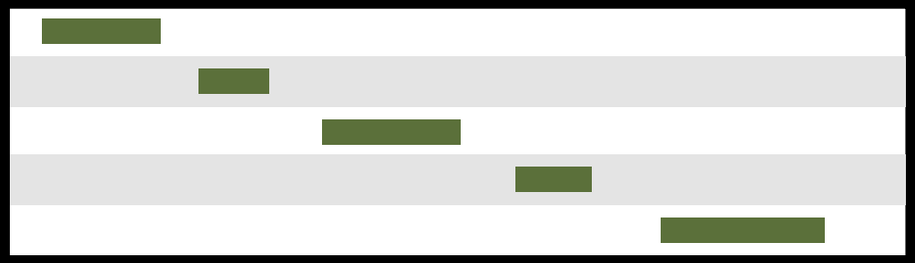
test.galaxyproject.org

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

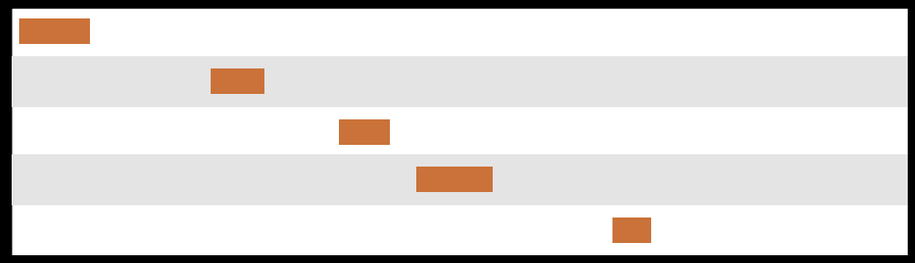
Exons & Repeats: A General Plan

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

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

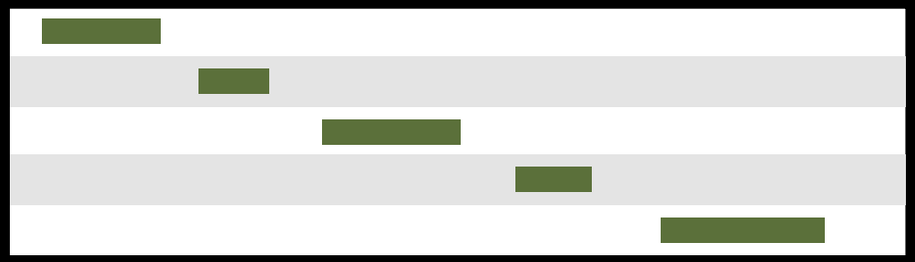


Exons

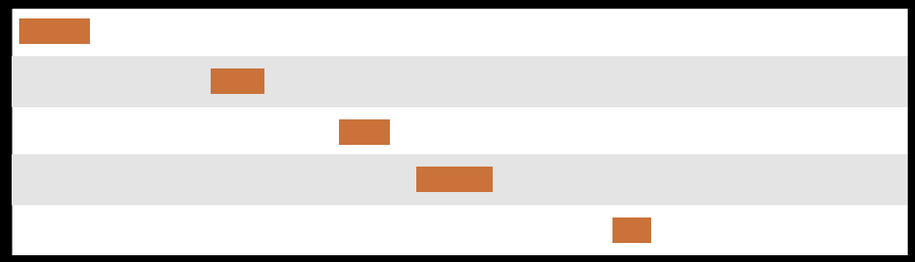


Repeats

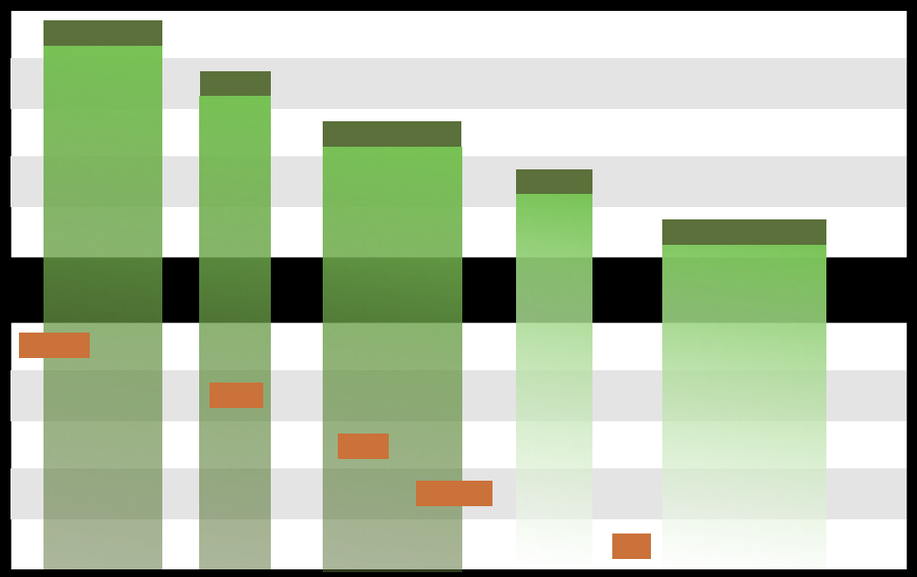
(Identify which exons have Repeats)



Exons



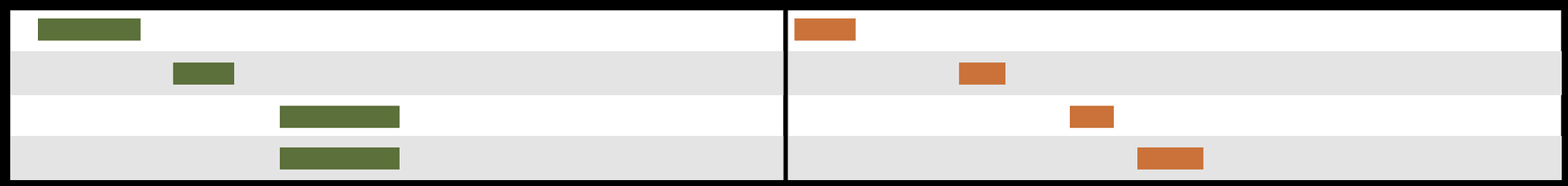
Repeats



Exons

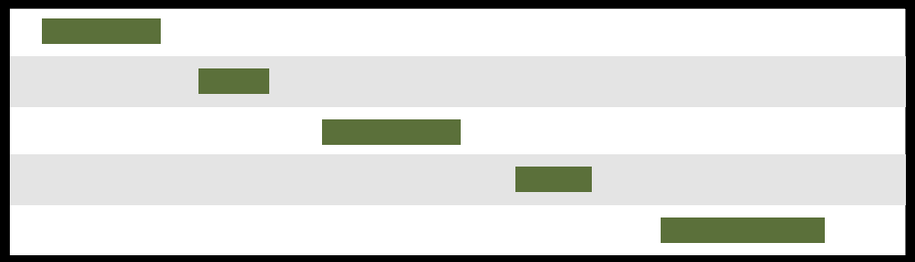
Repeats

Overlap pairings

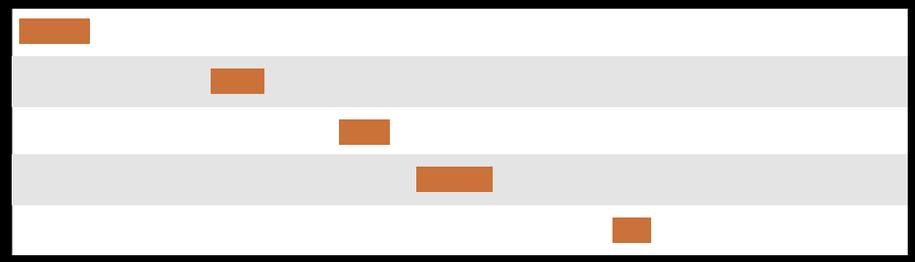


Operate on Genomic Intervals → **Join**

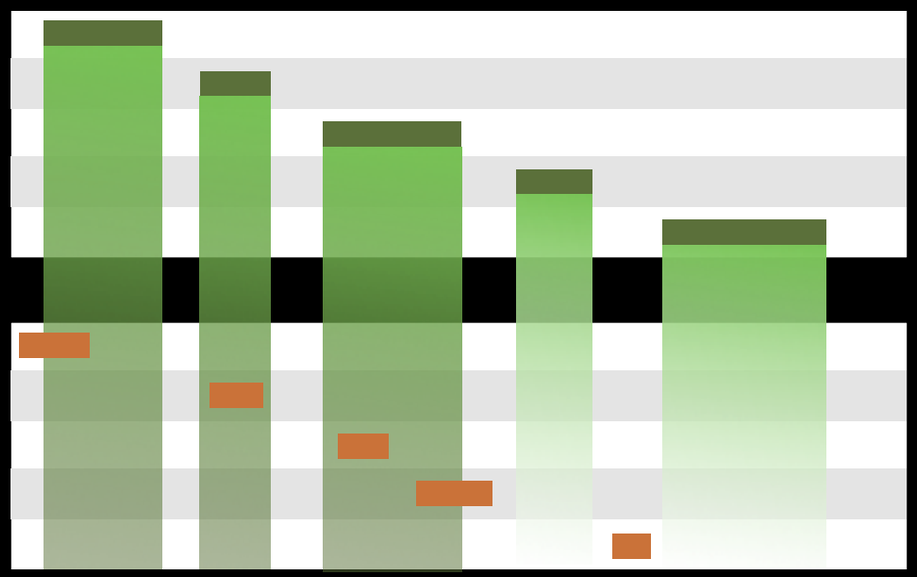
(Identify which exons have Repeats)



Exons



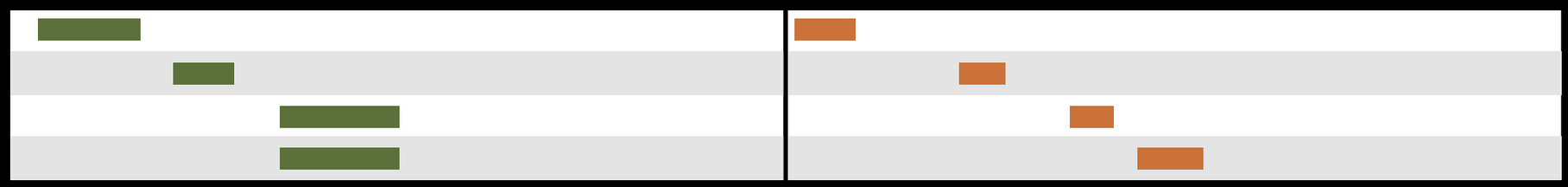
Repeats



Exons

Repeats

Overlap pairings

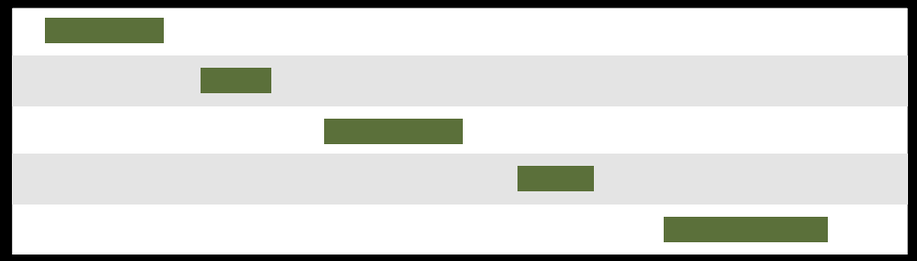


Exon overlap counts

Join, Subtract, and Group → Group

(Count Repeats per exon)

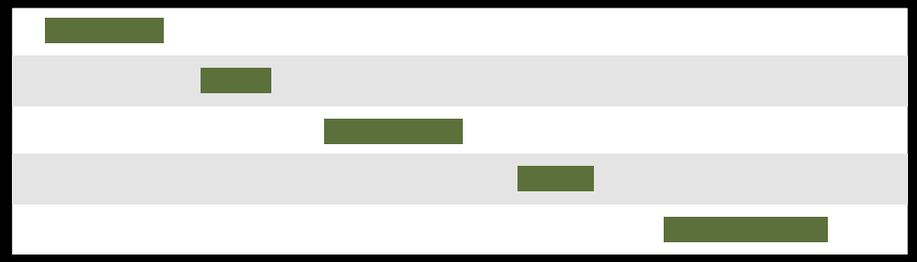
Exon overlap counts



Exons

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

Exon overlap counts



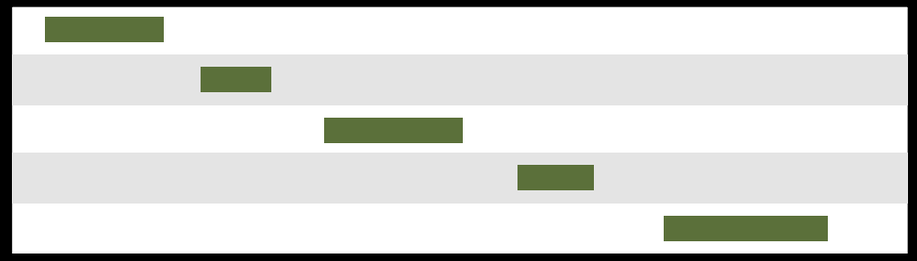
Exons

Join on exon name

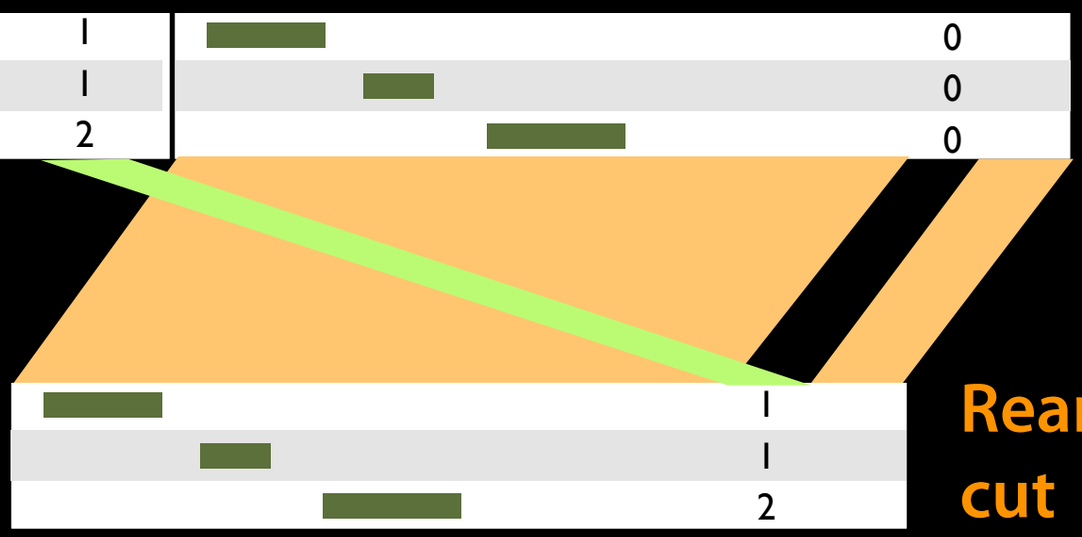
Join, Subtract, and Group → Join

(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)

Agenda: Day 1

9:30 Welcome

10:00 Basic Analysis with Galaxy
A worked example demonstrating Galaxy Basics

10:50 Break

11:10 Basic Analysis (continued)

12:30 Lunch (on your own)

13:30 RNA-Seq Analysis
Quality Control and Galaxy Workflows

15:30 Break

15:50 RNA-Seq Analysis
Mapping and Splice Junction Identification

17:00 Done

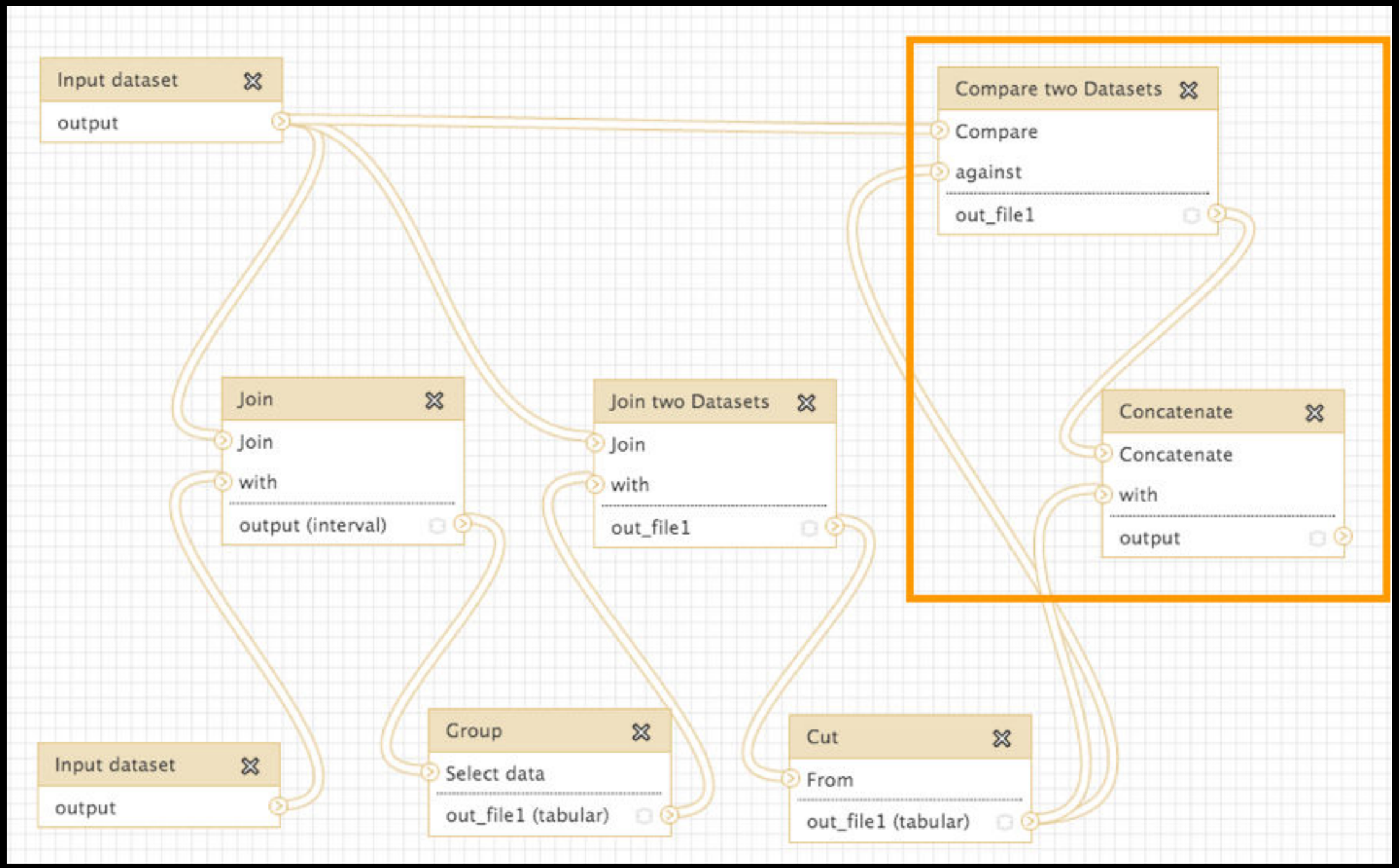
http://bit.ly/GxyCam2015

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.

Agenda: Day 1

9:30 Welcome

10:00 Basic Analysis with Galaxy
A worked example demonstrating Galaxy Basics

10:50 Break

11:10 Basic Analysis (continued)

12:30 Lunch (on your own)

13:30 RNA-Seq Analysis
Quality Control and Galaxy Workflows

15:30 Break

15:50 RNA-Seq Analysis
Mapping and Splice Junction Identification

17:00 Done

http://bit.ly/GxyCam2015

Exons & Repeats: Done?

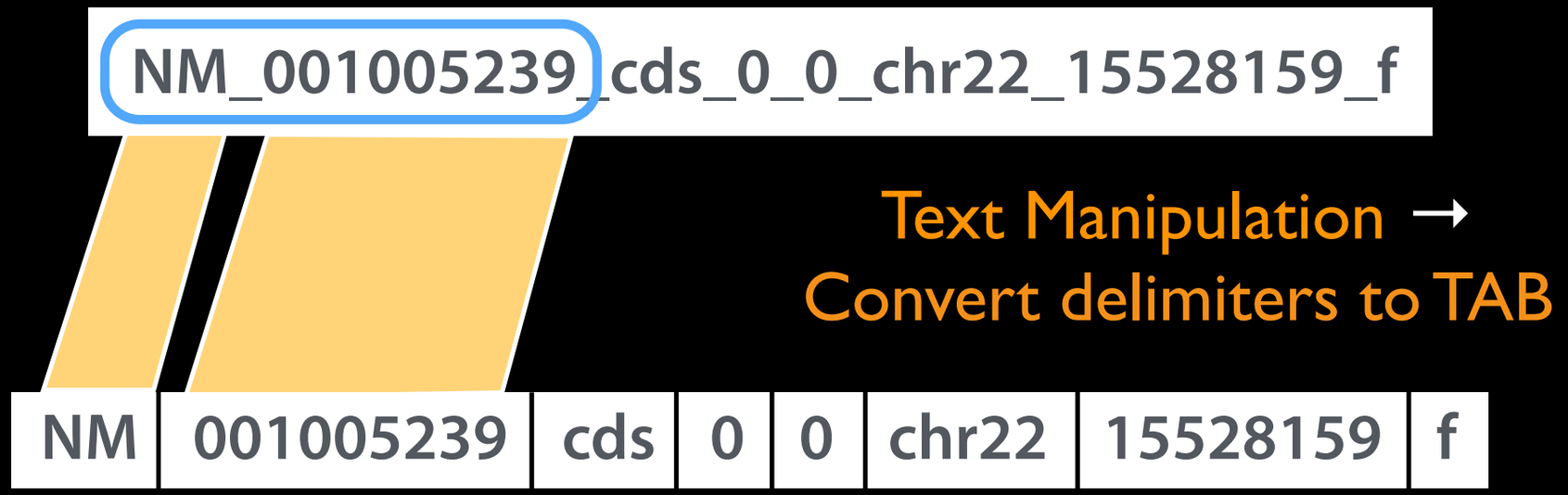
We now know which exons have repeats, and we have that information in a format that can be understood by many tools.

Let's see what those genes do.

Get the gene

NM_001005239_cds_0_0_chr22_15528159_f

Get the gene


NM_001005239_cds_0_0_chr22_15528159_f
Text Manipulation →
Convert delimiters to TAB
NM
001005239
cds
0
0
chr22
15528159
f

Get the gene

NM_001005239_cds_0_0_chr22_15528159_f

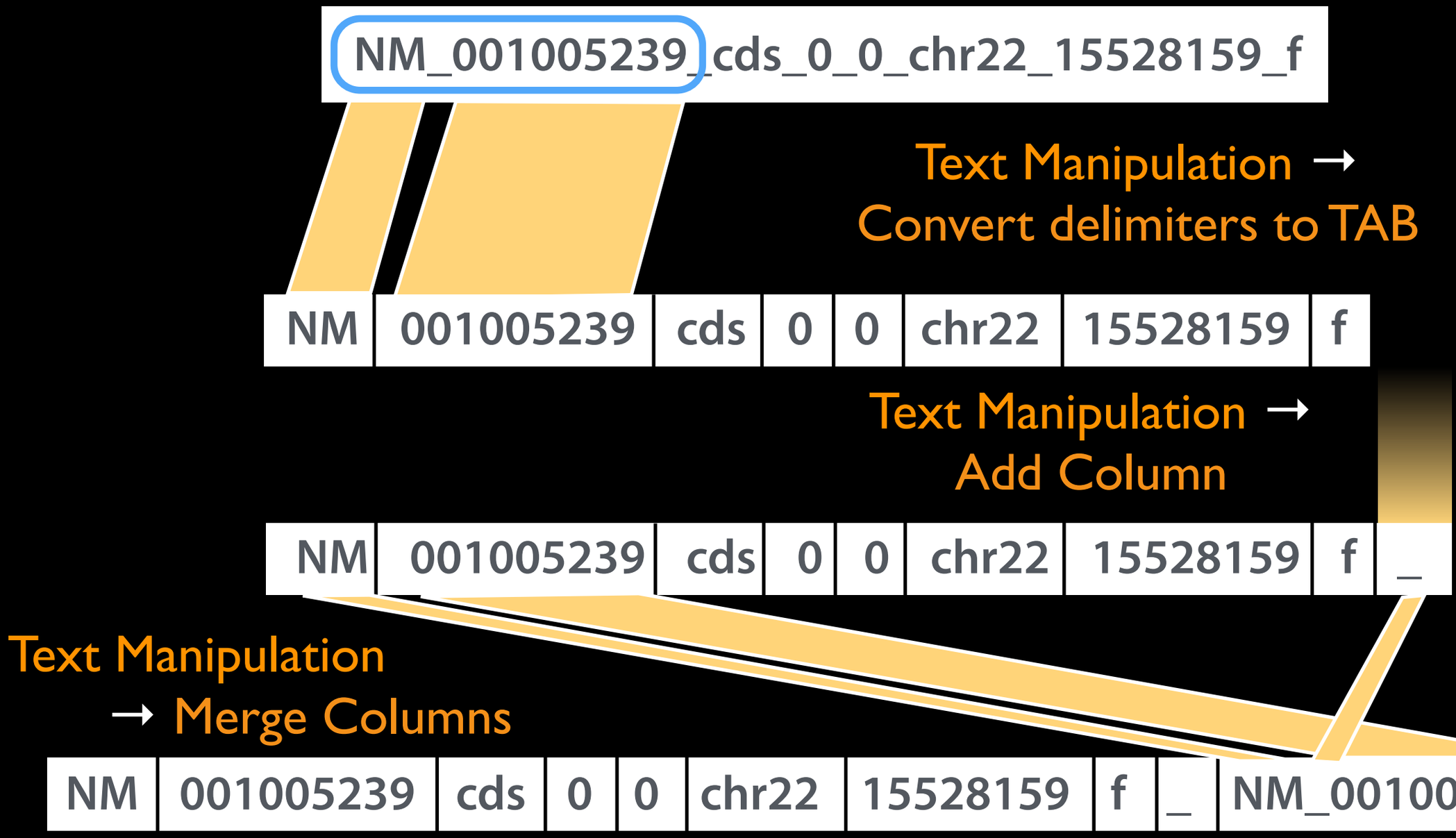
Text Manipulation →
Convert delimiters to TAB

NM	001005239	cds	0	0	chr22	15528159	f

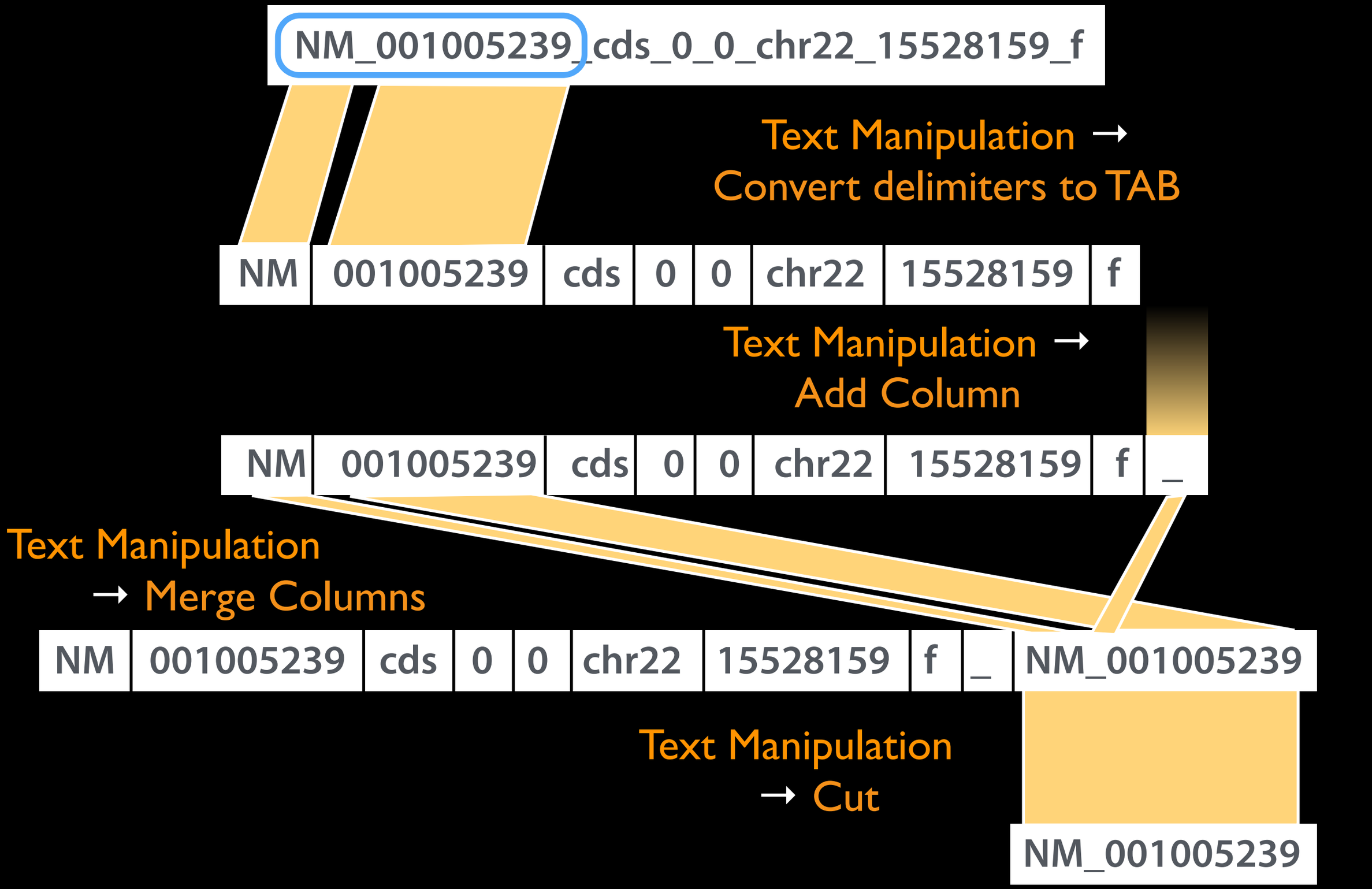
Text Manipulation →
Add Column

NM	001005239	cds	0	0	chr22	15528159	f	_

Get the gene


NM_001005239_cds_0_0_chr22_15528159_f
Text Manipulation →
Convert delimiters to TAB
NM 001005239 cds 0 0 chr22 15528159 f
Text Manipulation →
Add Column
NM 001005239 cds 0 0 chr22 15528159 f _
Text Manipulation
→ Merge Columns
NM 001005239 cds 0 0 chr22 15528159 f _ NM_001005239

Get the *damn* gene



Get the Genes

Remove duplicate gene names

Text Manipulation → Unique

Got the Gene*s*: Look for GO Enrichment

Let's see what those genes do.

http://geneontology.org/

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

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

Workflow:

A series of analysis steps
Can be repeated with different data

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

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

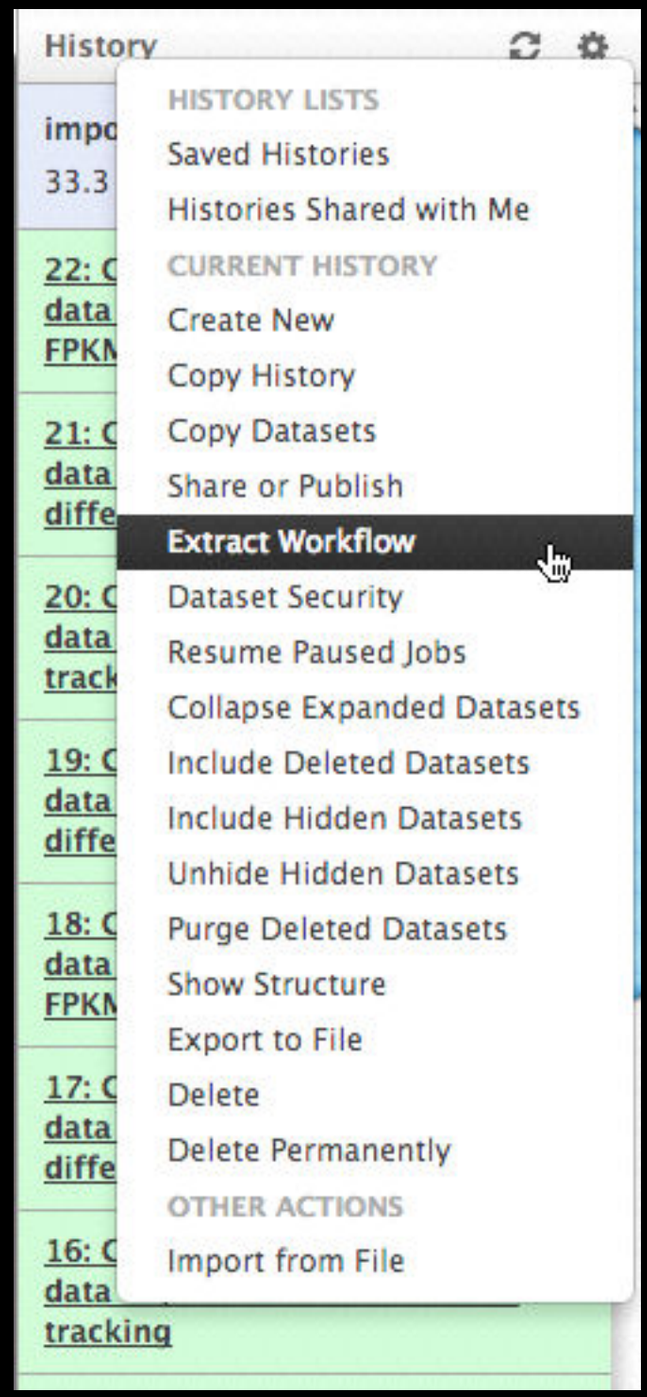
Create a Workflow from a History

Extract Workflow from history

Create a workflow from this history.

Edit it to make some things clearer.

(cog) → Extract Workflow



Run / test it

Guided: rerun with same inputs

Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions

Agenda: Day 1

9:30 Welcome

10:00 Basic Analysis with Galaxy
A worked example demonstrating Galaxy Basics

10:50 Break

11:10 Basic Analysis (continued)

12:30 Lunch (on your own)

13:30 RNA-Seq Analysis
Quality Control and Galaxy Workflows

15:30 Break

15:50 RNA-Seq Analysis
Mapping and Splice Junction Identification

17:00 Done

http://bit.ly/GxyCam2015

Agenda: Day 1

9:30 Welcome

10:00 Basic Analysis with Galaxy
A worked example demonstrating Galaxy Basics

10:50 Break

11:10 Basic Analysis (continued)

12:30 Lunch (on your own)

13:30 RNA-Seq Analysis
Quality Control and Galaxy Workflows

15:30 Break

15:50 RNA-Seq Analysis
Mapping and Splice Junction Identification

17:00 Done

http://bit.ly/GxyCam2015

RNA-Seq Analysis: Get the Data

Create new history

(cog) → Create New

Import:

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

→ Raw Reads → Select

MeOH_REP1_R1, MeOH_REP1_R2

MeOH_REP2_R1, MeOH_REP2_R2

→ Reference

→ Select all

UCDAVIS Genome Center Bioinformatics Core

* 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!

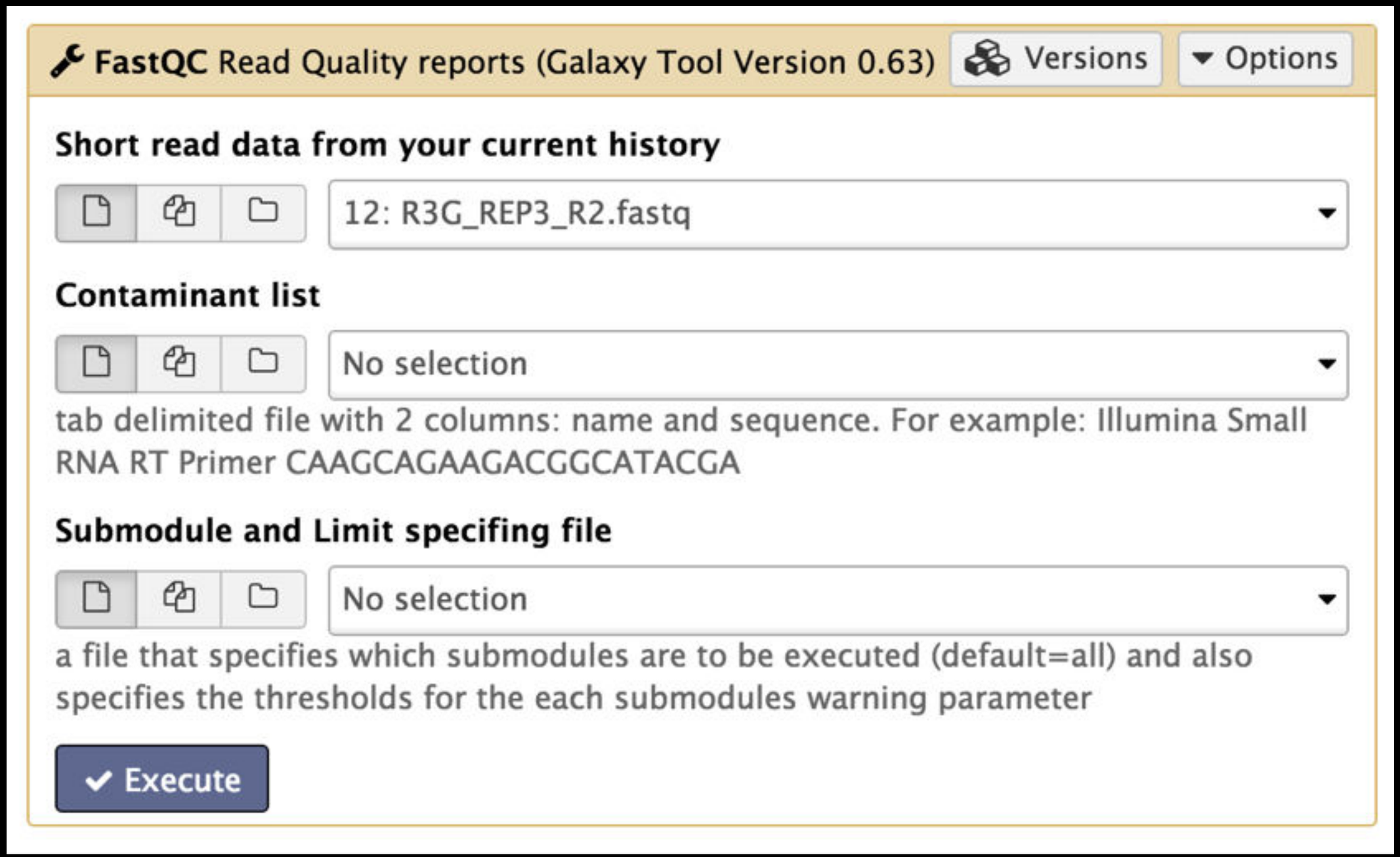
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  ............................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./0123456789:;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

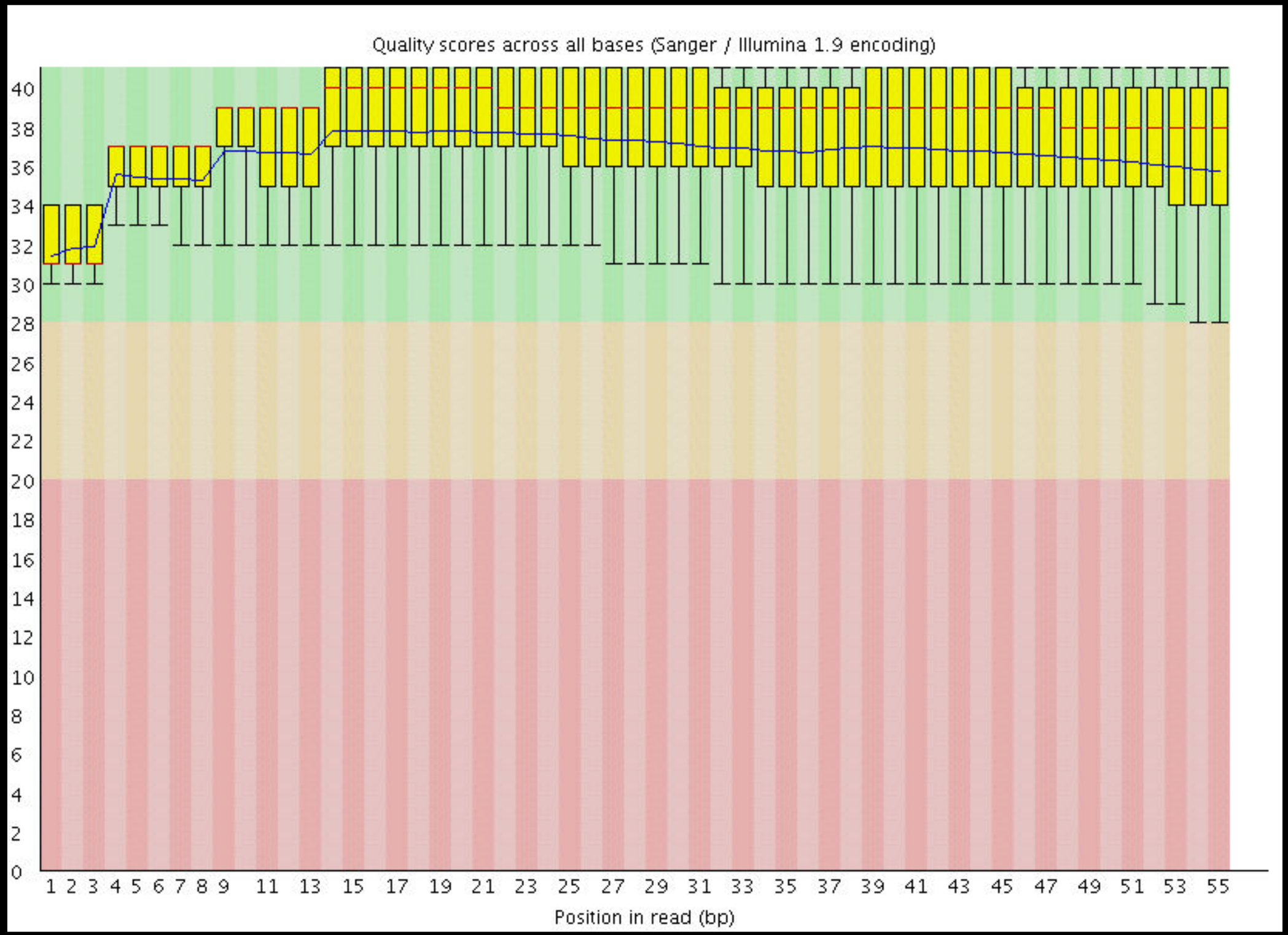
NGS Data Quality: Assessment tools

NGS QC and Manipulation → FastQC

Generates summary quality information.

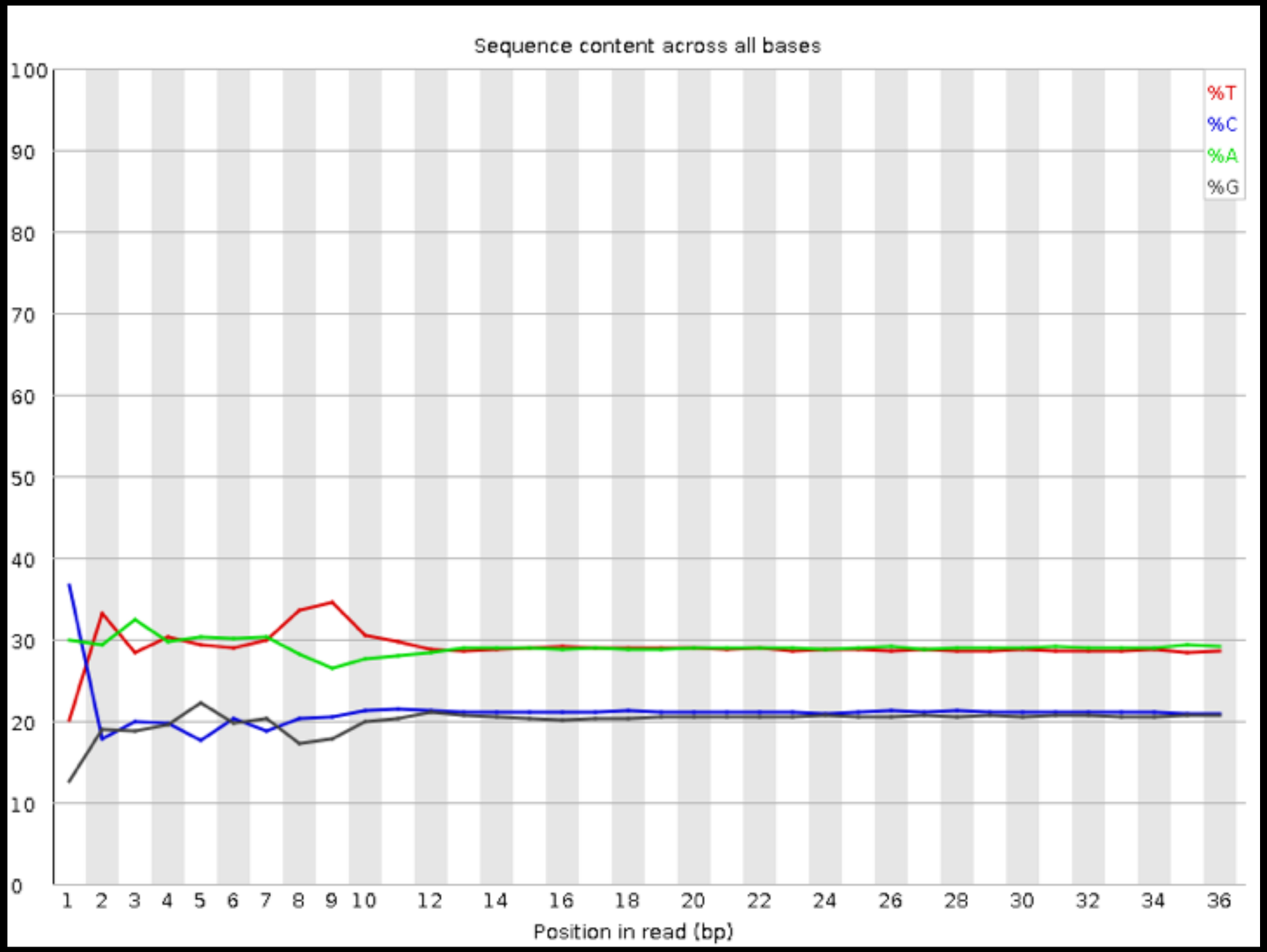


NGS Data Quality: Assessment tools



http://bit.ly/FastQCBoxPlot

NGS Data Quality: Sequence bias at front of reads?



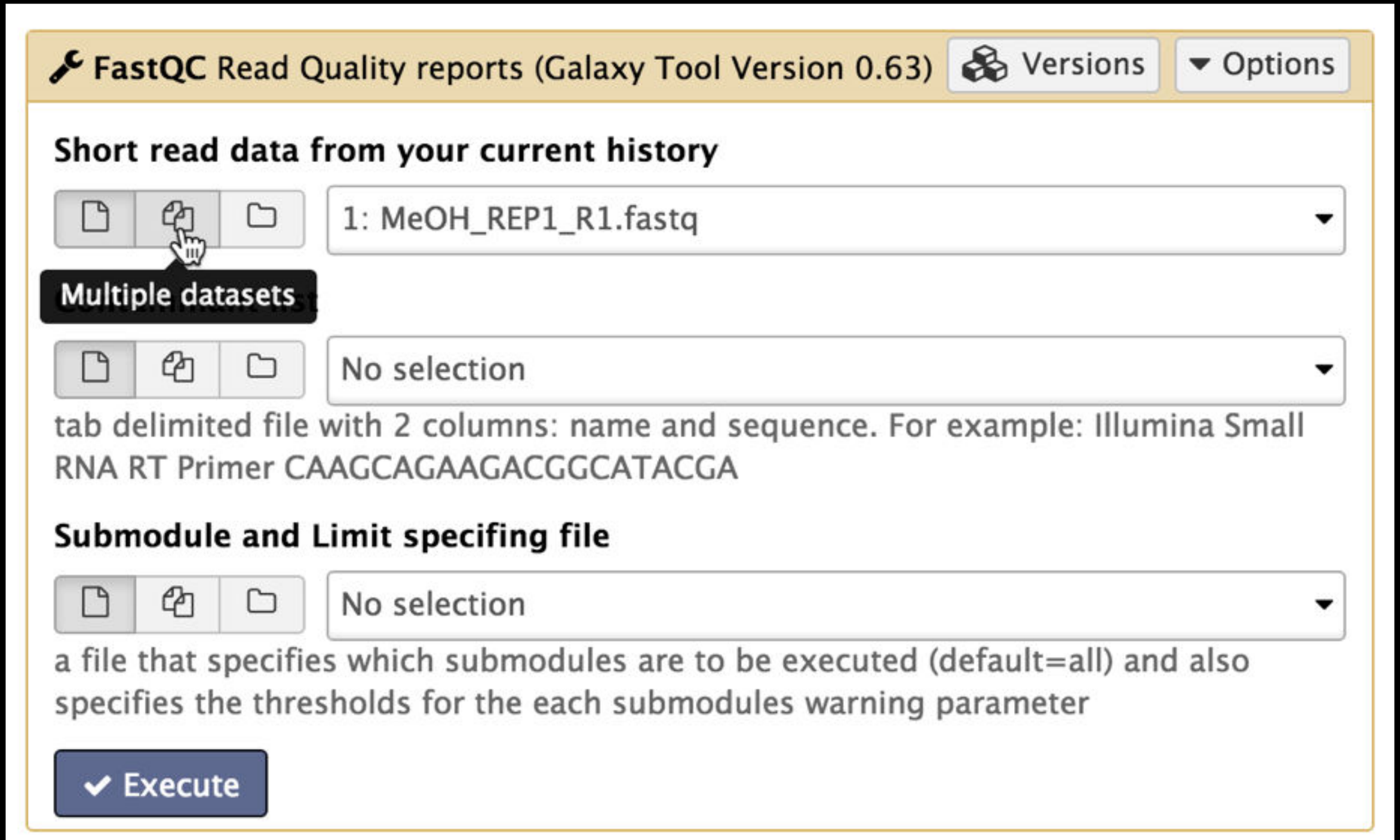
From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

NGS Data Quality Assessment: Done!

Now, just 11 more to go!

Your Friend: The Multiple datasets button



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

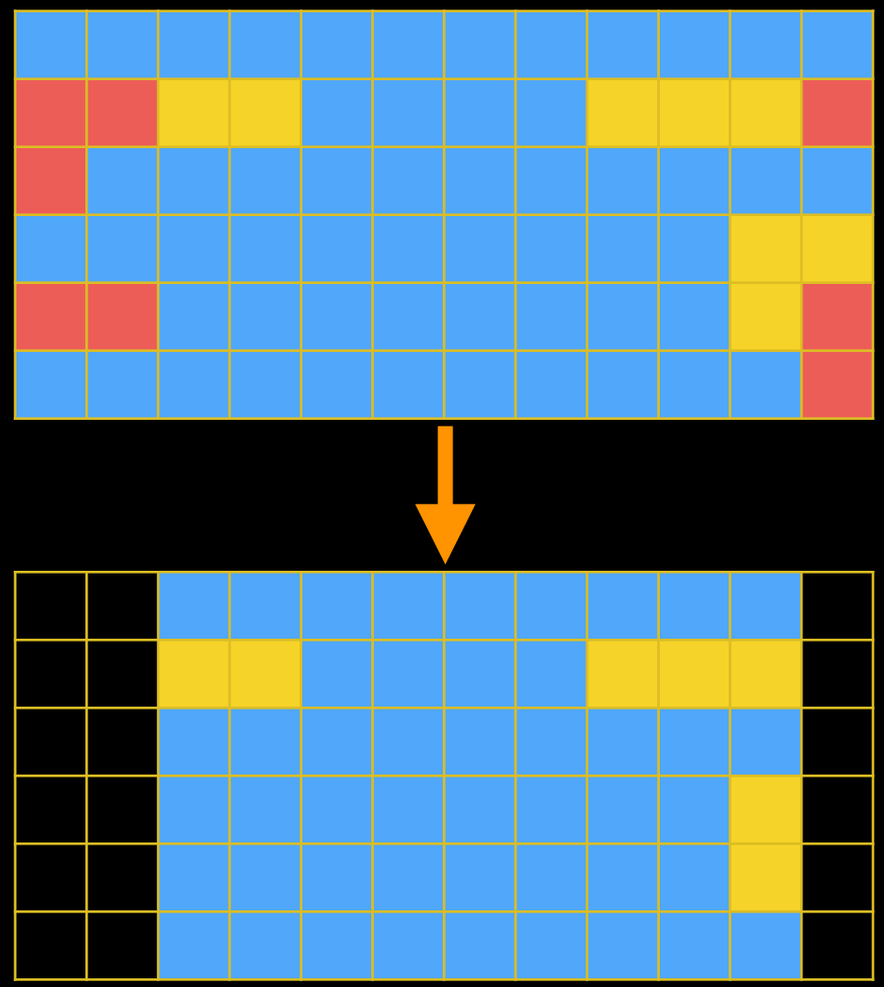
No selection

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

Execute

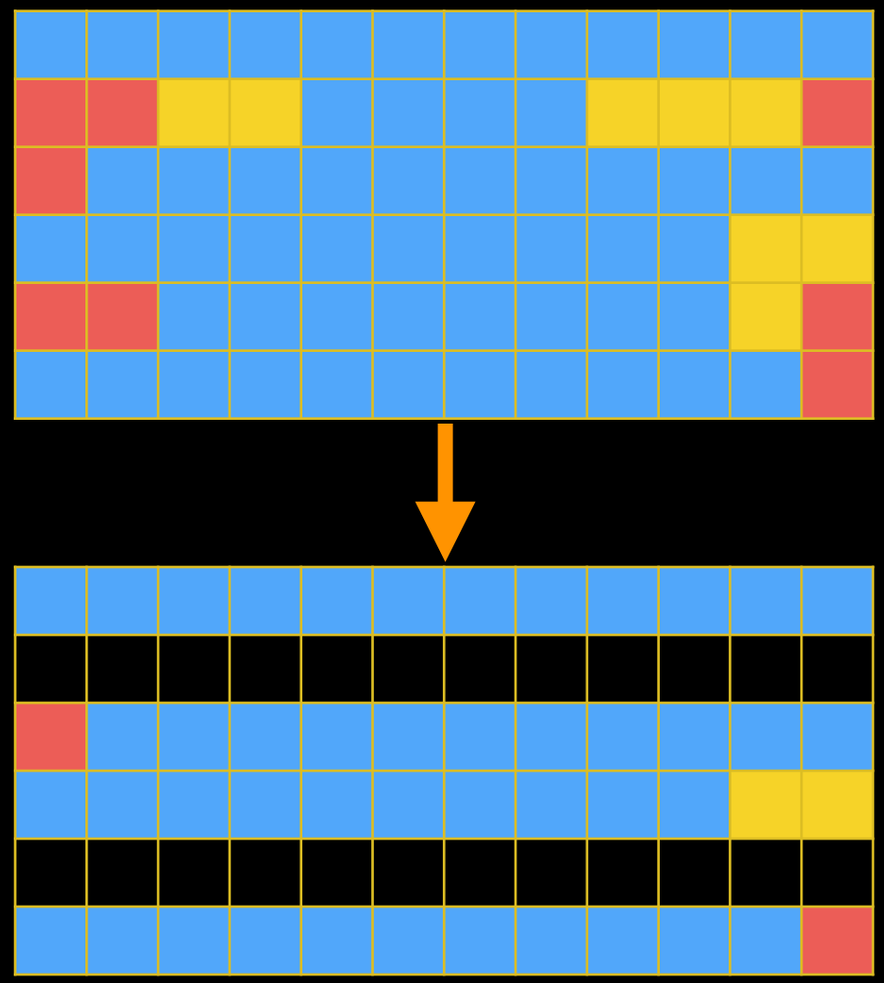
NGS Data Quality: Trim as we see fit

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



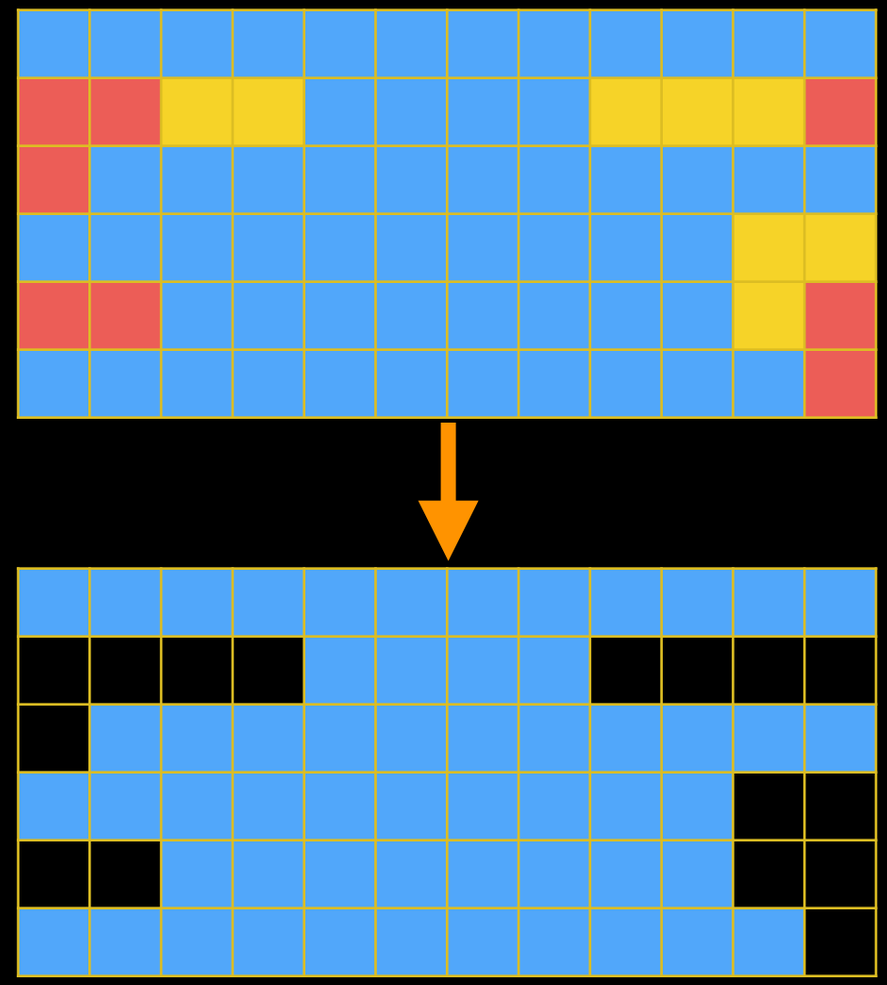
NGS Data Quality: Base Quality Trimming

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

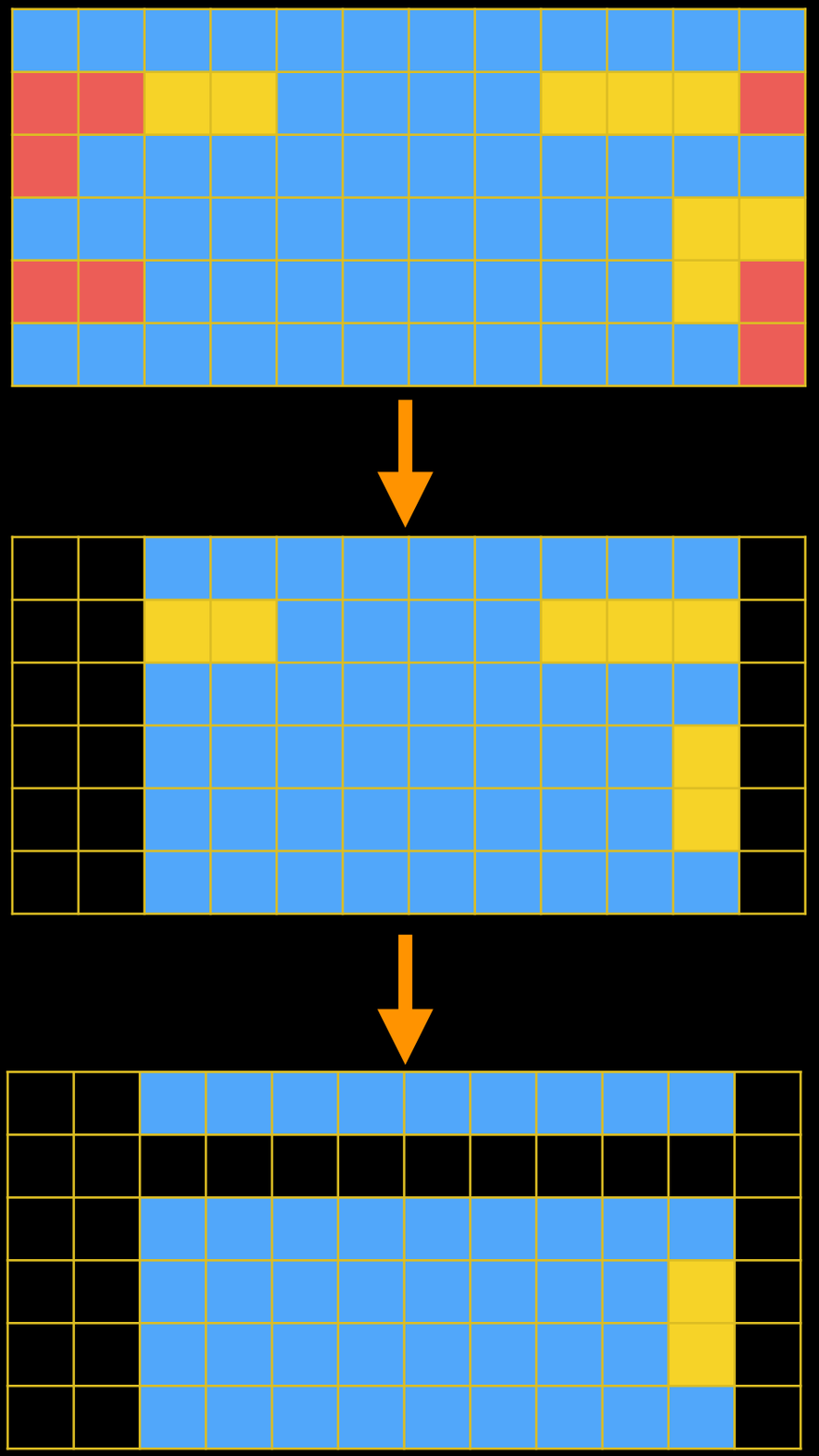


NGS Data Quality: Base Quality Trimming

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



Options are not mutually exclusive



Option 1
(by column)

+

Option 2
(by entire row)

Trim? *As we see fit?*

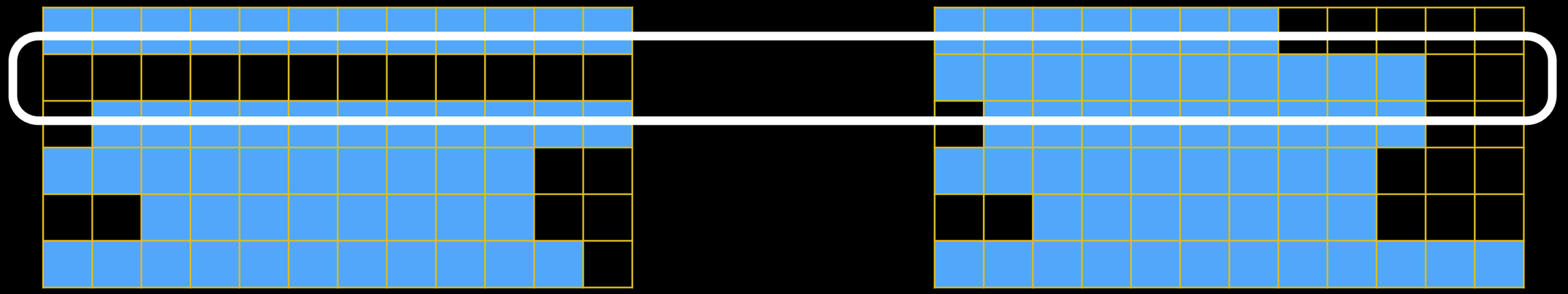
- 3 options
  - One preserves original read length, two don't
  - One preserves number of reads, two don't
  - Two keep/make every read the same length, one does not
  - One preserves pairings, two don't

Trim? *As we see fit?*

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

Trim? *As we see fit?*

- 3 options
  - ...
  - **One preserves pairings, two don't**



Keeping paired ends paired: Things to Try

- Don't bother.
- Run a workflow (try the "Re-Pair Paired ends after QC may have broken them" workflow on usegalaxy.org) that removes any unpaired reads before mapping
- Run the Picard Paired Read Mate Fixer after mapping reads.

RNA-Seq Analysis: Restore Pairings

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

Shared Data → Published Workflows

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

→ Import

And then

Workflows

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

→ Run

Re-Pair Paired ends after QC may have broken them

Workflow takes 4 inputs

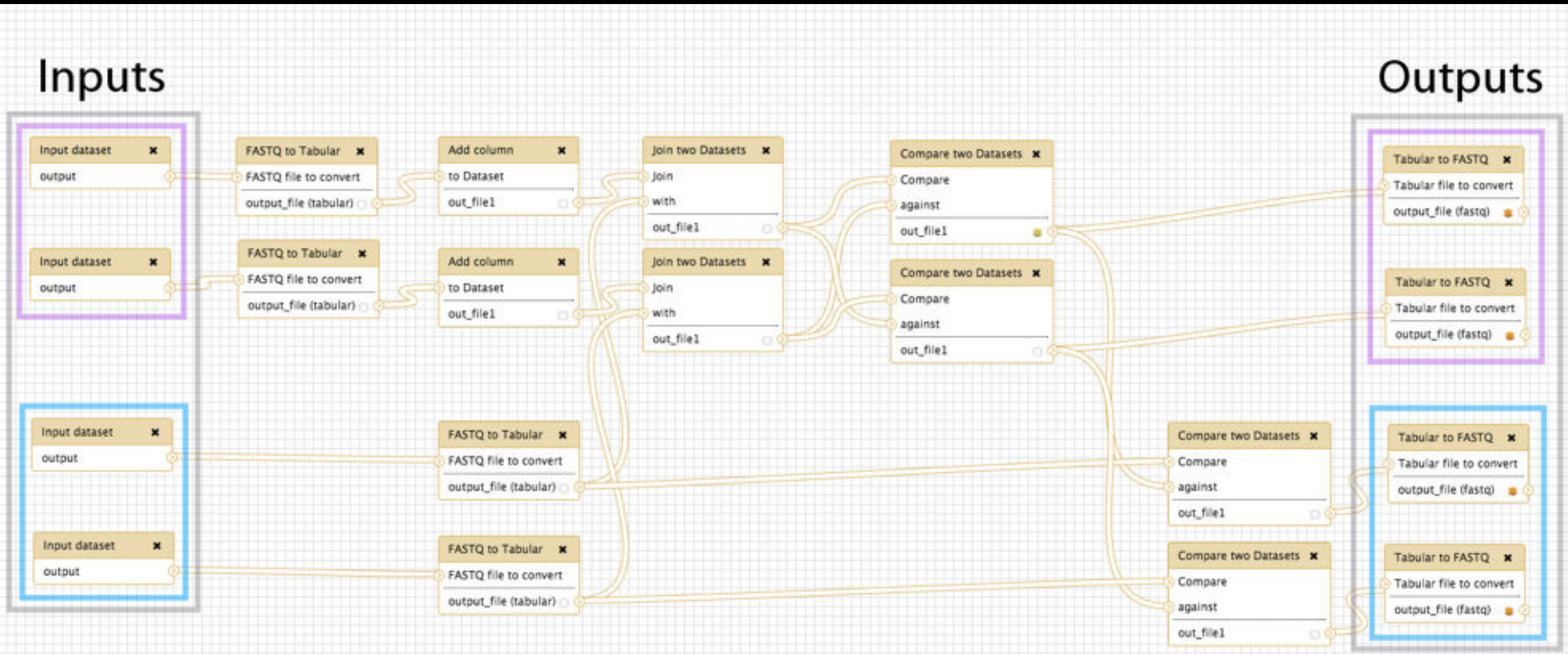
- Forward Reads, before QC
- Reverse Reads, before QC
- Forward Reads, after QC
- Reverse Reads, after QC

And produces 4 outputs

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

Workflow assumes pre-QC reads are correctly paired

Re-Pair Paired ends after QC may have broken them



NGS Data Quality: 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
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	342	0.2052981325073385	No Hit
CGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	325	0.1950932545756871	No Hit
CTGCATTATAAAAAGGACAGCCAGATATCAACTGTTACAGAAATGAAATAAGACG	230	0.13806599554587093	No Hit
CGGCCGCAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGA	199	0.11945710049403614	No Hit
GTCAGCTCAACTTGTAGGCCCCAAAAGAAAACAGCGTCTTACTGGGGAGGGATAT	197	0.11825652661972422	No Hit

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

But this will break pairings (if we still have them).

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

NGS Quality Control *Revisted*

"Quality Control is not sexy.

But it is vital."

Really?

Do QC, or rely on bad data not to map?

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 with Tophat: Max # of Alignments Allowed

Some reads align to more than one place equally well.

For such reads, how many should Tophat include?

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

Tophat breaks ties randomly.

Tophat assigns equal fractional credit to all *n* mappings

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

TopHat Manual

Mapping With Tophat: Cleanup

Use only the good stuff!

NGS BAM Tools → Filter

Mapping Quality → **>=20**

Insert Filter → **isProperPair: Yes**

Insert Filter → **reference: chr12**

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

RNA-Seq Mapping With Tophat: Resources

RNA-Seq Concepts, Terminology, and Work Flows
by Monica Britton

Aligning PE RNA-Seq Reads to a Genome
by Monica Britton

both from the UC Davis 2013 Bioinformatics Short Course

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

from the GCC2013 Training Day

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

from the GCC2012 Training Day

Agenda: Day 1

9:30 Welcome

10:00 Basic Analysis with Galaxy
A worked example demonstrating Galaxy Basics

10:50 Break

11:10 Basic Analysis (continued)

12:30 Lunch (on your own)

13:30 RNA-Seq Analysis
Quality Control and Galaxy Workflows

15:30 Break

15:50 RNA-Seq Analysis
Mapping and Splice Junction Identification

17:00 Done

http://bit.ly/GxyCam2015

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

clements@galaxyproject.org

http://bit.ly/GxyCam2015

Introduction to Galaxy: RNA-seq & ChIP-seq Data Analysis

University of Cambridge
11-12 June 2015

Dave Clements
Galaxy Project
Johns Hopkins University

http://bit.ly/GxyCam2015

Agenda: Day 2

9:30 RNA-Seq Analysis
Differential expression

10:50 Break

11:10 RNA-Seq Analysis
continued

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis
Quality Control and Mapping

15:00 Break

15:20 ChIP-Seq Analysis
Differential Binding and Comparing Results

16:30 Done

http://bit.ly/GxyCam2015

RNA-Seq Analysis: Transcript Prediction

Create new history

(cog) → Create New

Import:

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

→ Mapped Filtered Reads

→ Select all

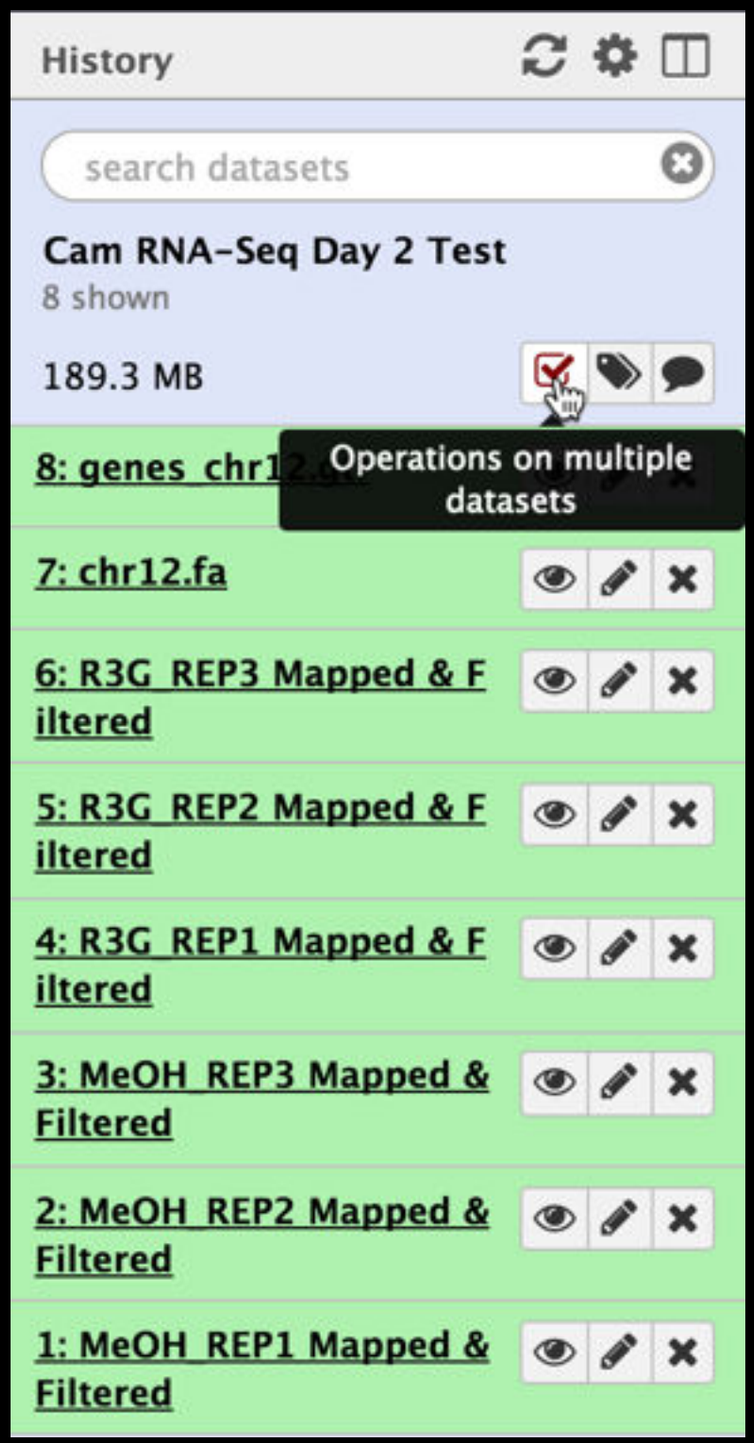
→ Reference

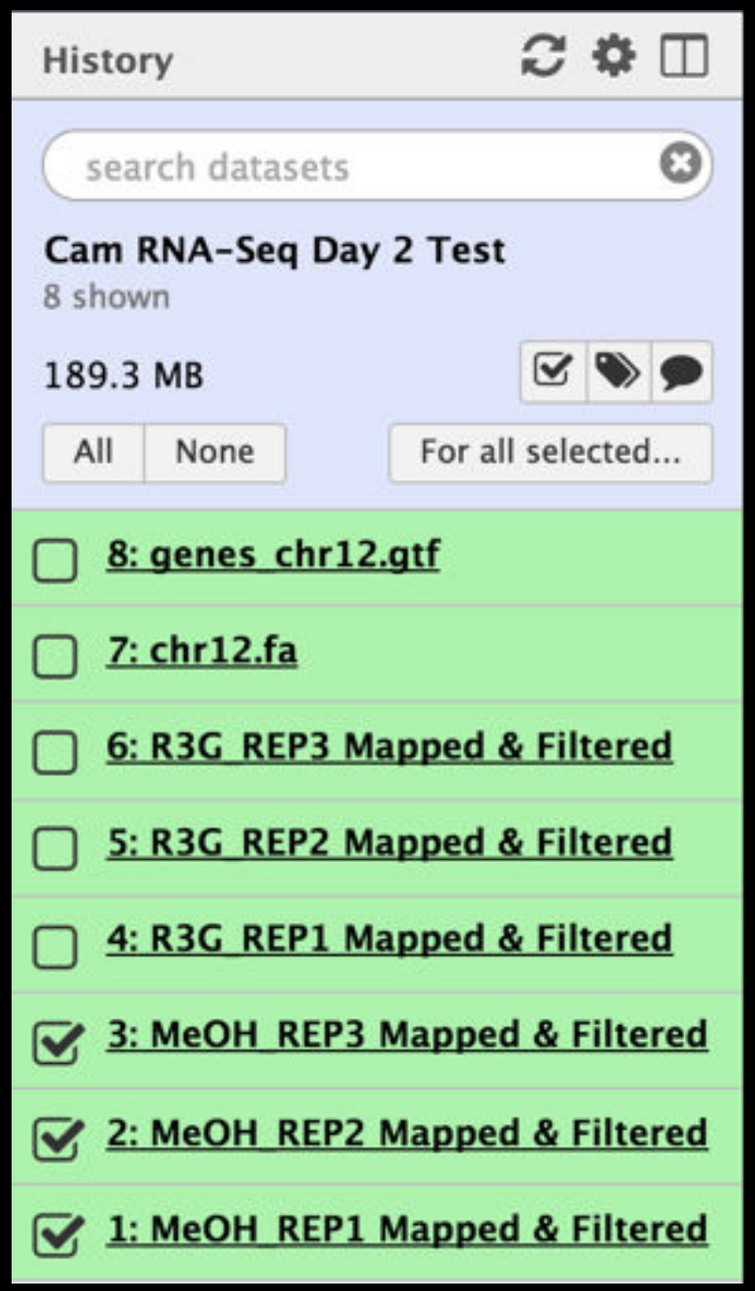
→ Select all

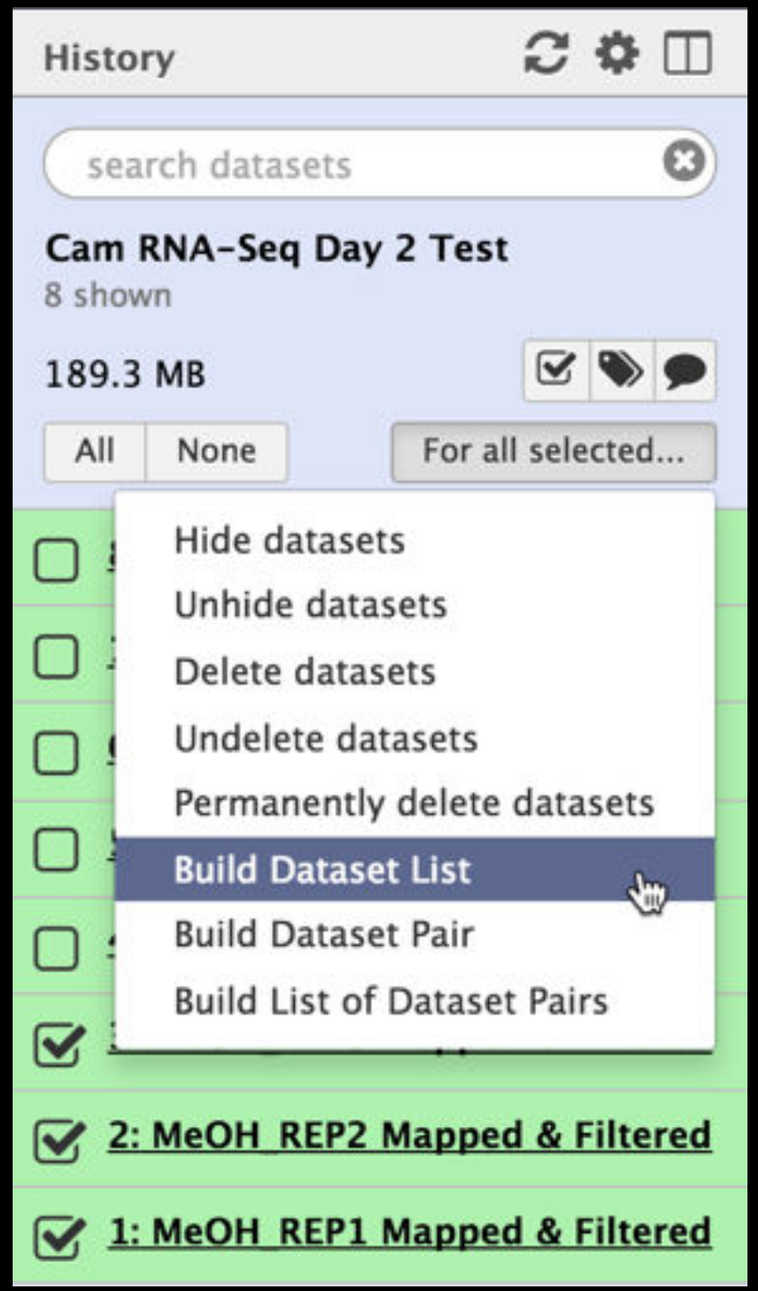
UCDAVIS Genome Center Bioinformatics Core

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

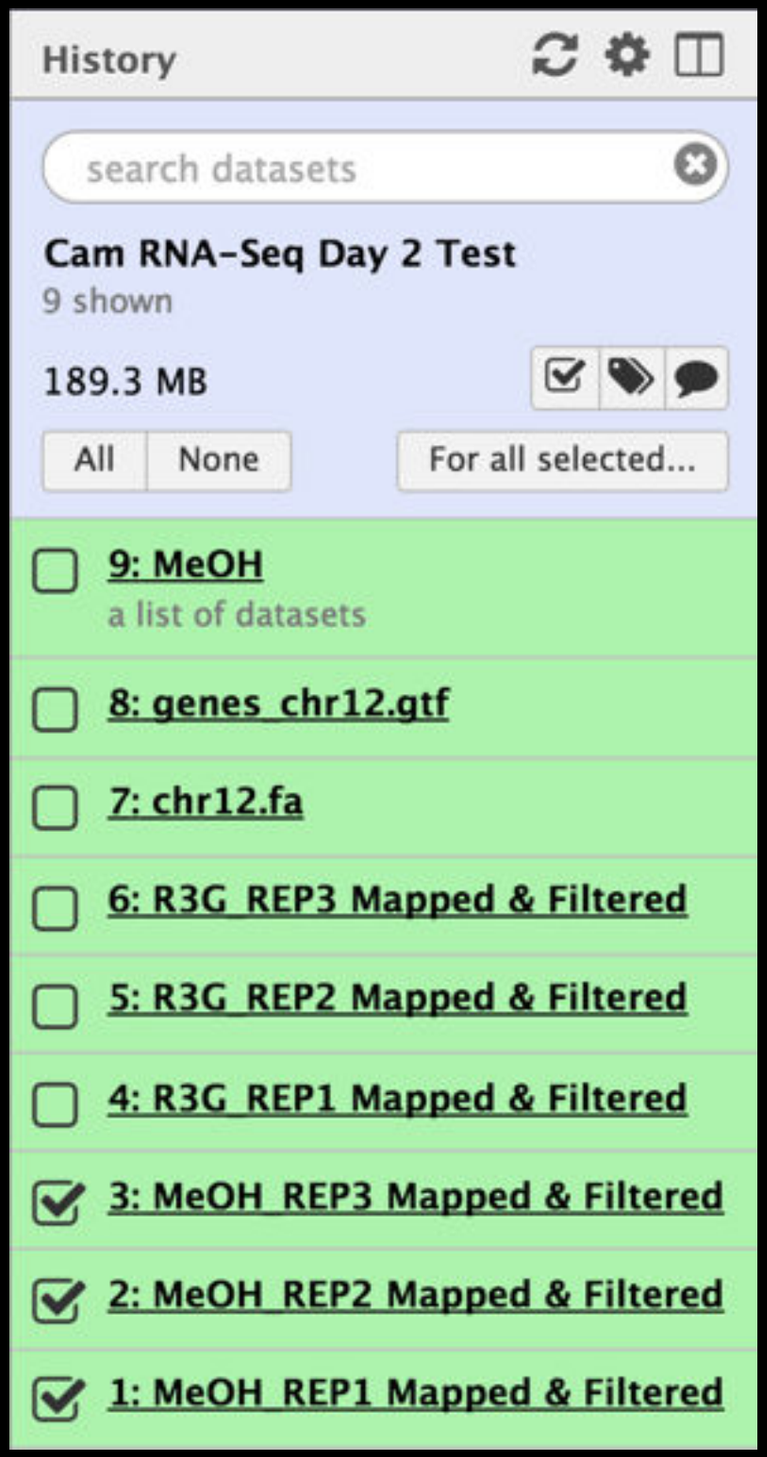
RNA-Seq Analysis: Collections







RNA-Seq Analysis: Collections



RNA-Seq Analysis: Transcript Prediction

Cufflinks and StringTie are both tools that predict transcripts based on mapped reads.

We'll use StringTie as it's rumoured to be faster and more accurate.

NGS RNA Analysis → StringTie

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT & Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads *Nature Biotechnology* 2015, doi:10.1038/nbt.3122

RNA-Seq Analysis: StringTie

StringTie transcript assembly and quantification (Galaxy Tool Version 1.0.3) ▾ Options

Mapped reads to assemble transcripts from

6: R3G_REP3 Mapped & Filtered

Dataset collection

Use **ssembly**

Do not use GFF

Options

Use defaults

Job Resource Para

Use default job res

✔ Execute

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

Do not use GFF

Options

Use defaults

Job Resource Parameters

Use default job resource parameters

✔ Execute

RNA-Seq Analysis: Transcript Prediction

NGS RNA Analysis → StringTie

Use GFF to guide assembly → Yes

Perform abundance information only of input transcripts → No

Output additional files for use in Ballgown → No

Options → Use defaults

RNA-Seq Analysis: Transcript Prediction

StringTie outputs:

Assembled Transcripts:
Transcripts that StringTie successfully assembled

Coverage:
Reference transcripts that are fully covered by reads

RNA-Seq Analysis: Unify Predictions

Have transcript predictions from 6 replicates
Have reference transcripts as well

Cuffmerge unifies these 7 sets of predictions into a single rationalised set of transcripts.

RNA-Seq Analysis: Transcript Prediction

Run Cuffmerge

NGS RNA Analysis → Cuffmerge
Run it with the two assembled transcripts collections from StringTie
Use reference Annotation? → Yes
Use sequence data? → Yes
Source for reference List → History, chr12.fa

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.

Cuffdiff

- Which Transcript definitions to use?
  - Official (genes_chr12.gtf in our case)
  - MeOH or R3G Cufflinks transcripts
  - Results of Cuffmerge on MeOH & R3G Cufflinks transcripts
- Depends on what you care about
  - I'll use transcripts from 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 in FPKM
p_value	Uncorrected P value for test statistic
q_value	FDR-adjusted p-value for the test statistic
status	Was there enough data to run the test?
significant	and, was the gene differentially expressed?

Cuffdiff

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

Returns the list of genes with

1) enough data to make a call, and

2) that are called as differentially expressed.

Cuffdiff: Next Steps

Try running Cuffdiff with different normalization and dispersion estimation methods.

Compare the differentially expressed gene lists.

Which settings have what type of impacts on the results?

Are there any patterns to the identified genes?

Agenda: Day 2

9:30 RNA-Seq Analysis
Differential expression

10:50 Break

11:10 RNA-Seq Analysis
continued

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis
Quality Control and Mapping

15:00 Break

15:20 ChIP-Seq Analysis
Differential Binding and Comparing Results

16:30 Done

http://bit.ly/GxyCam2015

Agenda: Day 2

9:30 RNA-Seq Analysis
Differential expression

10:50 Break

11:10 RNA-Seq Analysis
continued

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis
Quality Control and Mapping

15:00 Break

15:20 ChIP-Seq Analysis
Differential Binding and Comparing Results

16:30 Done

http://bit.ly/GxyCam2015

ChIP-Seq: FASTQ data and quality control

By Shannan Ho Sui

Look at two transcription factor proteins, Pou5f1 and Nanog, in H1hesc cell lines.

Both are involved in self-renewal of undifferentiated embryonic stem cells

H3ABioNet

http://hbc.github.io/ngs-workshops/courses/introduction-to-chip-seq/

http://bit.ly/nanogname

ChIP-Seq Analysis: Get the Data

Import

Shared Data → Data Libraries → Training →

ChIP-Seq → Raw Reads

H1hesc_Input_Rep1_chr12.fastq

H1hesc_Input_Rep2_chr12.fastq

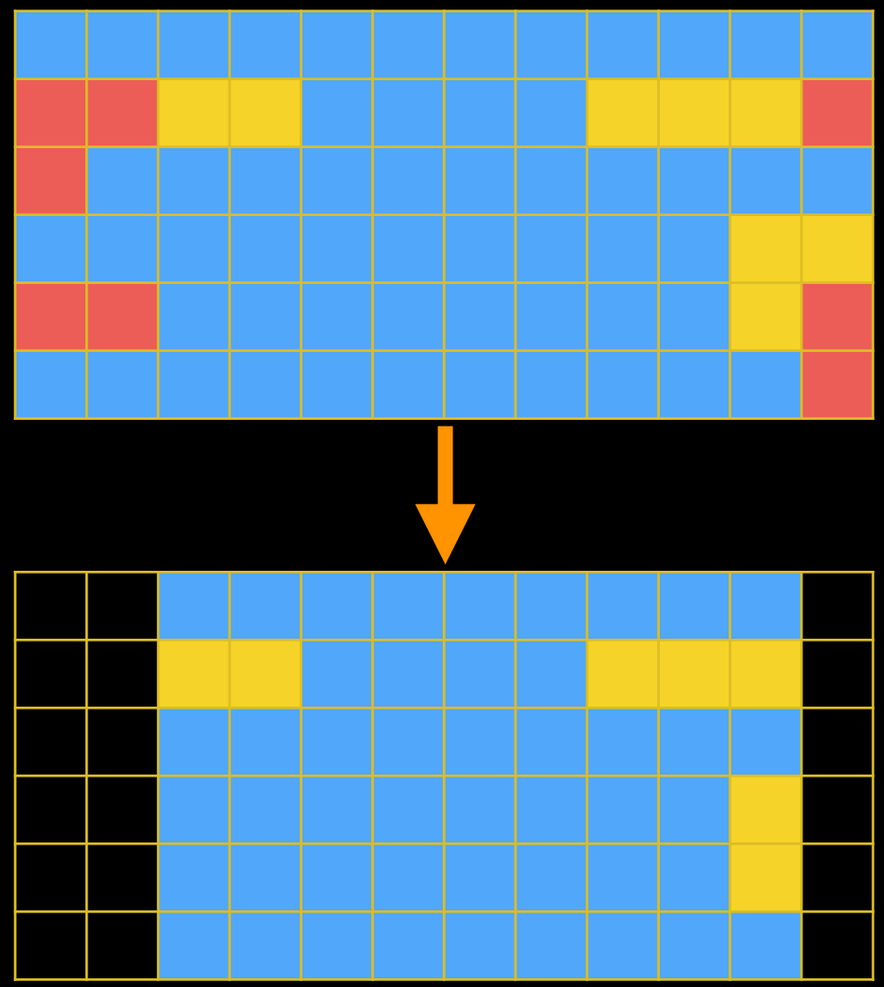
NGS Data Quality: Assessment tools

Same tools available as yesterday:

FastQC, Sliding window, Trimmer by Column, by quality score and length

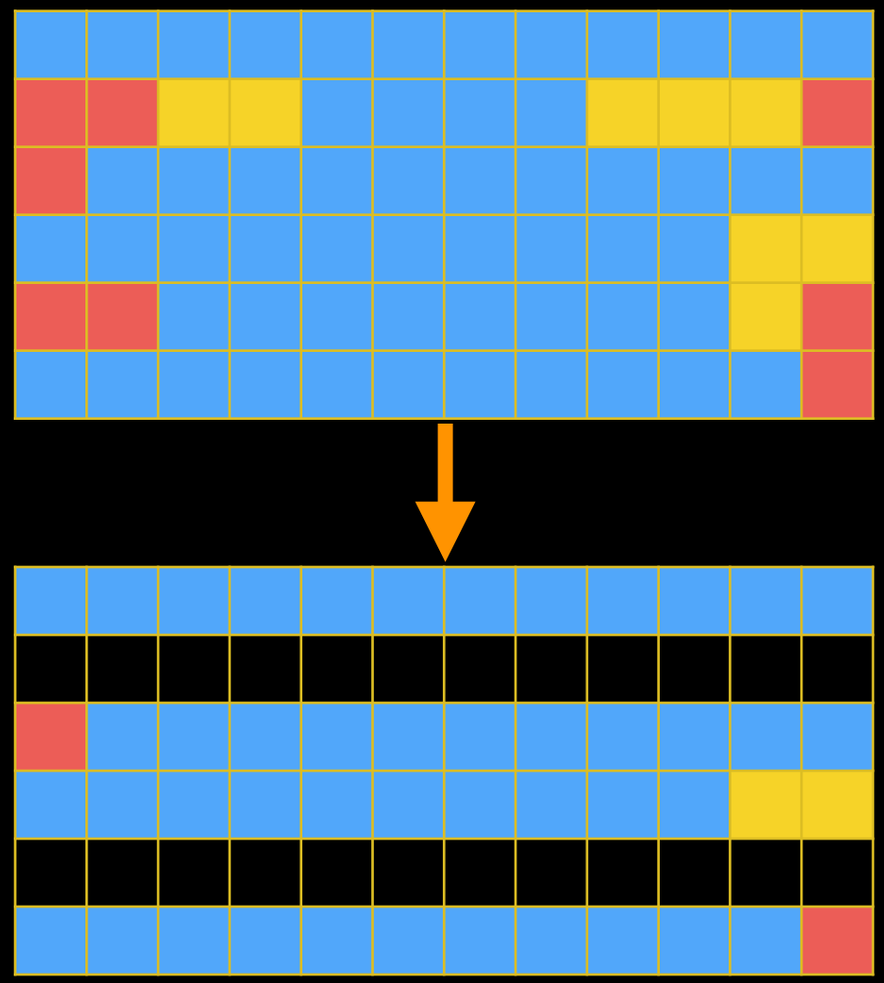
NGS Data Quality: Trim as we see fit

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



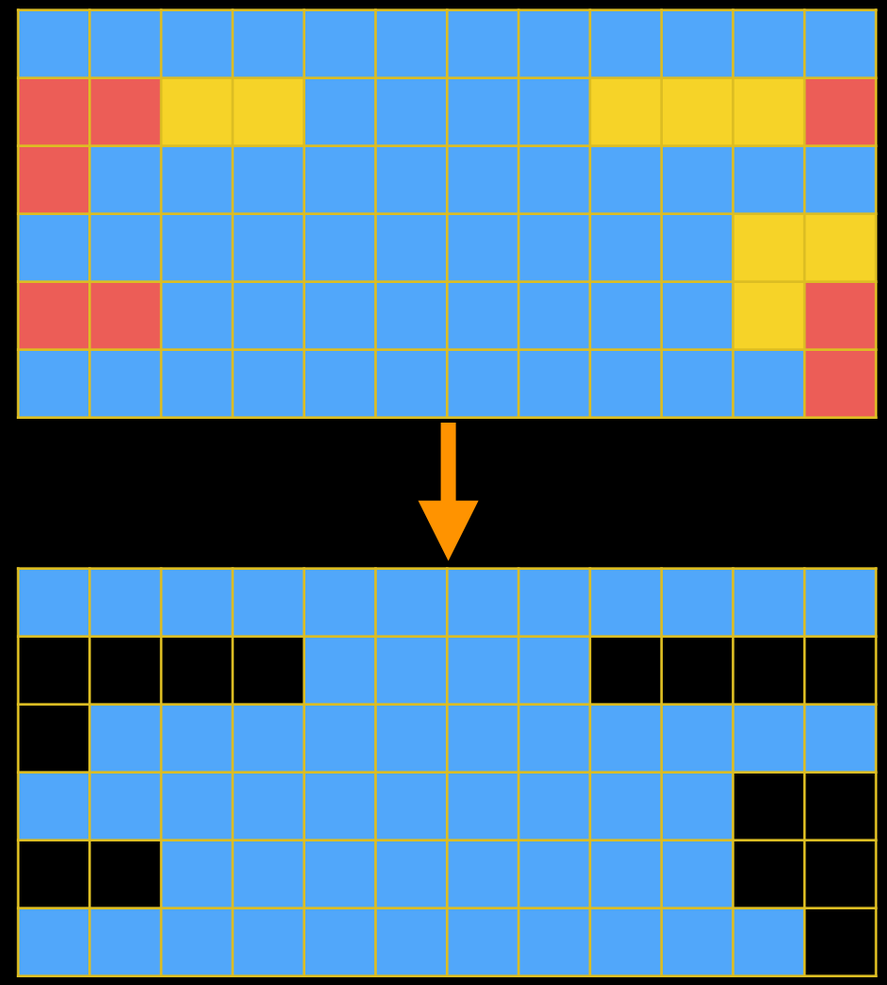
NGS Data Quality: Base Quality Trimming

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



NGS Data Quality: Base Quality Trimming

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



Trim? *As we see fit?*

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

Does MACS2 care? *No.*

From the MACS Announcement mailing list

Ian 10/22/14

Call me Dr. Impatient, but has anyone an answer for this?

Thanks again.
- show quoted text -

Tao Liu 10/24/14

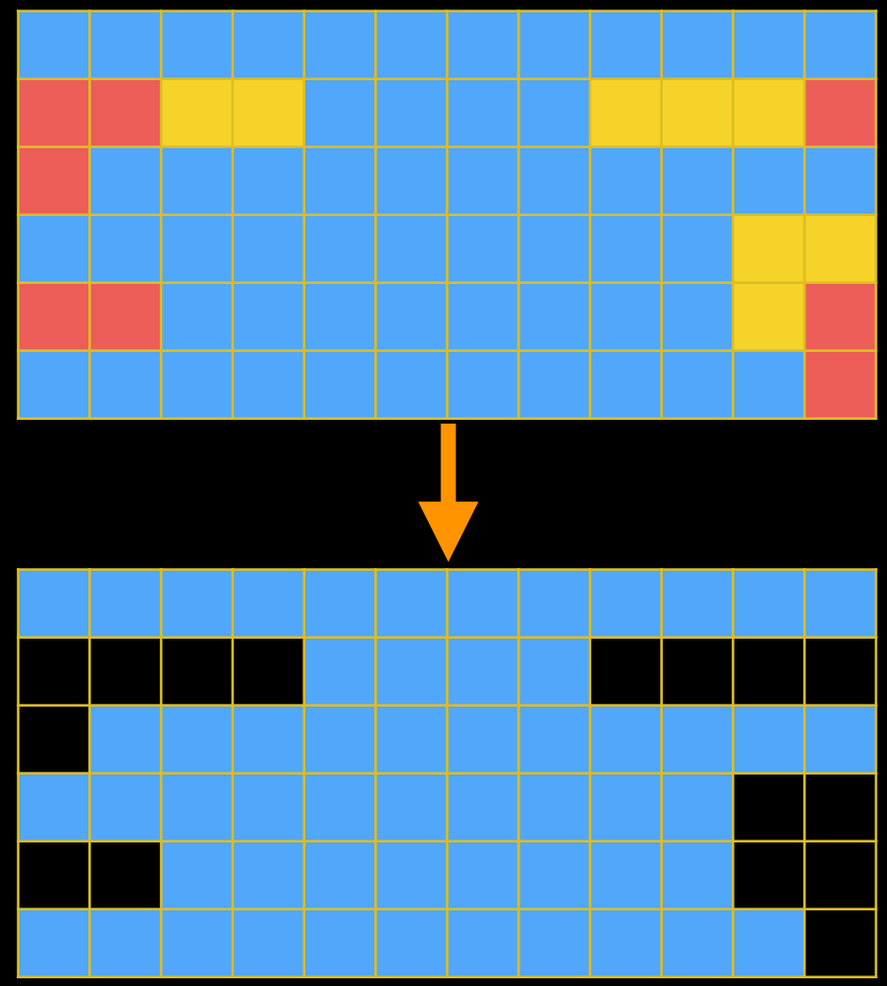
Dear Dr. Impatient,

Tag size only affects how MACS (version 1) builds strand model to compute fragment size. And in MACS2, it's not even effective while computing fragment size since only 'cutting' positions are informative. But in MACS2, the so-called maximum gap (an internal value) for merging nearby significant regions is set as read length since we regard this as the resolution of your data. In fact, it has very little impact on peak calling. So… briefly, you don't need to worry about this parameter. Longer reads help a lot for the reads alignment, but not much for peak calling.

Best,
Tao

Does MACS2 care? No

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



ChIP-Seq Analysis: Get the Data

Shared Data → Data Libraries → Training →

ChIP-Seq

Select everything in the Filtered Reads folder

Also grab genes_chr12.gtf from library

ChIP-Seq Exercise: Mapping with Bowtie

Use Bowtie2 (could also use BWA)

NGS Mapping: → Bowtie2

FASTQ file → H1hesc_Nanog_Rep1 post-QC

Single End

ChIP-Seq Analysis: remove unmapped reads

NGS Picard → FilterSamReads

Filtering Type → Include Aligned

ChIP-Seq Analysis: Get the Data

Shared Data → Data Libraries → Training →

ChIP-Seq

Select everything in the **Mapped Reads** folder

(These already have unmapped removed.)

ChIP-Seq Analysis: Find Peaks

NGS: ChIP-seq → MACS2 callpeak

Treatment File → Nanog Rep 1

Control File → H1hesc_Input_Rep2_chr12 Mapped BAM file

Outputs → Peaks, Scores

https://github.com/taoliu/MACS/

ChIP-Seq Analysis: Replicates

Shared Data → Data Libraries → Training → ChIP-Seq →

MACS Outputs → Peaks in BED format

Import files for

Nanog Rep 2

Pou5f1 Rep 1

Pou5f1 Rep 2

(or just get all 4)

ChIP-Seq Analysis: Unify Replicates

Operate on Genomic Intervals → Concatenate

Concatenate Nanog Rep 1 and 2 peak files

Operate on Genomic Intervals → Cluster

Use default parameters

Rename the output dataset

ChIP-Seq Analysis: Unify Replicates

Repeat for Pou5f1 replicates

Operate on Genomic Intervals → Concatenate

Concatenate Pou5f1 Rep 1 and 2 Peak files

Operate on Genomic Intervals → Cluster

Use default parameters

Rename the output dataset

ChIP-Seq Analysis: Differential binding

Operate on Genomic Intervals → Subtract

First dataset clustered → Pou5f1

Second dataset clustered → Nanog

Return → Intervals with no overlap

ChIP-Seq Mapping With MACS
Further reading & Resources

ChIP-Seq: FASTQ data and quality control

by Shannan Ho Sui

HAIB TFBS ENCODE collection

MACS Documentation

Model-based analysis of ChIP-Seq (MACS)

by Zhang *et al.*

Cistrome and Nebula Galaxy Servers

Nebula Tutorial

by Valentina Boeva

Agenda: Day 2

9:30 RNA-Seq Analysis
Differential expression

10:50 Break

11:10 RNA-Seq Analysis
continued

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis
Quality Control and Mapping

15:00 Break

15:20 ChIP-Seq Analysis
Differential Binding and Comparing Results

16:30 Done (almost)

http://bit.ly/GxyCam2015

Your Feedback: We *need* it

http://bit.ly/GxyCamFeedback

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

Galaxy is hiring post-docs and software engineers

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Also thanks to

Gabriella Rustici

Paul Judge

Cathy Hemmings

Anne Pajon

&

You

http://bit.ly/GxyCamFeedback

Agenda: Day 2

9:30 RNA-Seq Analysis
Differential expression

10:50 Break

11:10 RNA-Seq Analysis
continued

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis
Quality Control and Mapping

15:00 Break

15:20 ChIP-Seq Analysis
Differential Binding and Comparing Results

16:30 Done

http://bit.ly/GxyCam2015

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

clements@galaxyproject.org

http://bit.ly/GxyCamFeedback