http://bit.ly/glaxy2015slides

Introduction to Galaxy

University of Glasgow
8-9 June 2015

Dave Clements
Galaxy Project
Johns Hopkins University

Mani Mudaliar
Glasgow Polyomics
University of Glasgow

Graham Hamilton
Glasgow Polyomics
University of Glasgow

Agenda: Day 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

http://bit.ly/glaxy2015slides

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.

http://bit.ly/glaxy2015slides

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, RNA-Seq, variant analysis, ... However, you won't be an expert at any of these at the end of the workshop.

You will know enough to get you started.

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

openstack™
CLOUD SOFTWARE

amazon web services™

OpenNebula.org
The Open Source Toolkit for Cloud Computing

globus genomics

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

We are using the cloud today.

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

9:00 Welcome
9:30 Basic Analysis with Galaxy
10:45 Break
11:15 Basic Analysis (continued)
12:00 Basic Analysis into Reusable Workflows
12:30 Lunch (on your own)
13:30 ChIP-Seq Analysis
15:30 Break
16:00 Genome Assembly Concepts
16:30 Q & A Session
17:00 Done

Basic Analysis

Which exons have most overlapping Repeats?

Use Human, HG38, Chromosome 22

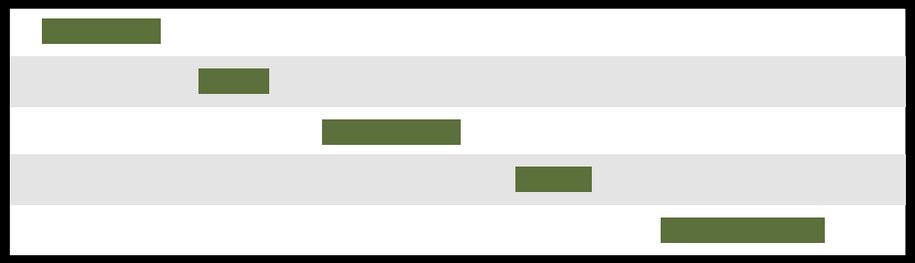
test.galaxyproject.org

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

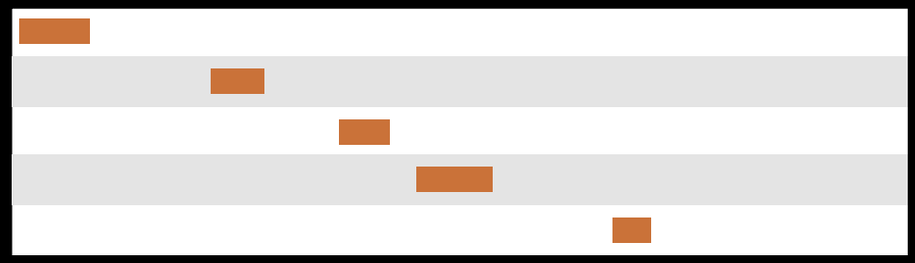
Exons & Repeats: A General Plan

- 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)

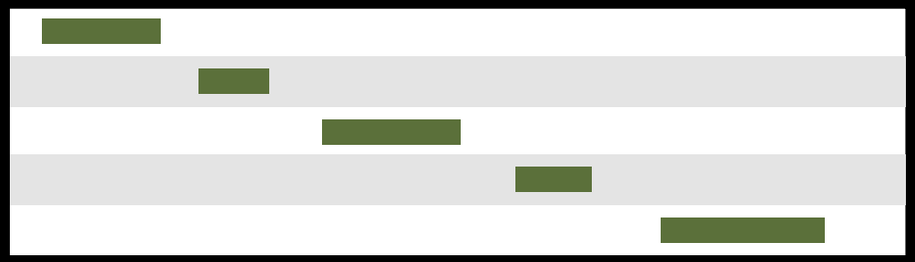


Exons

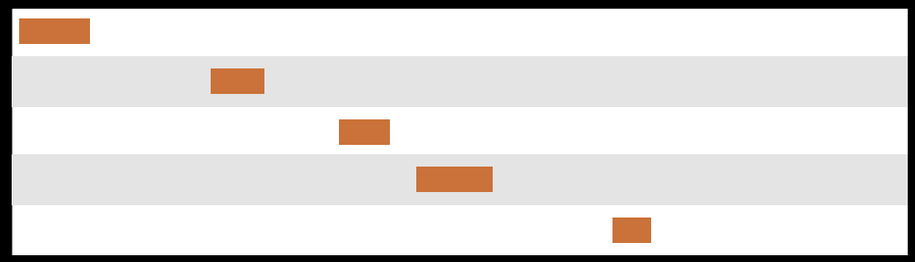


Repeats

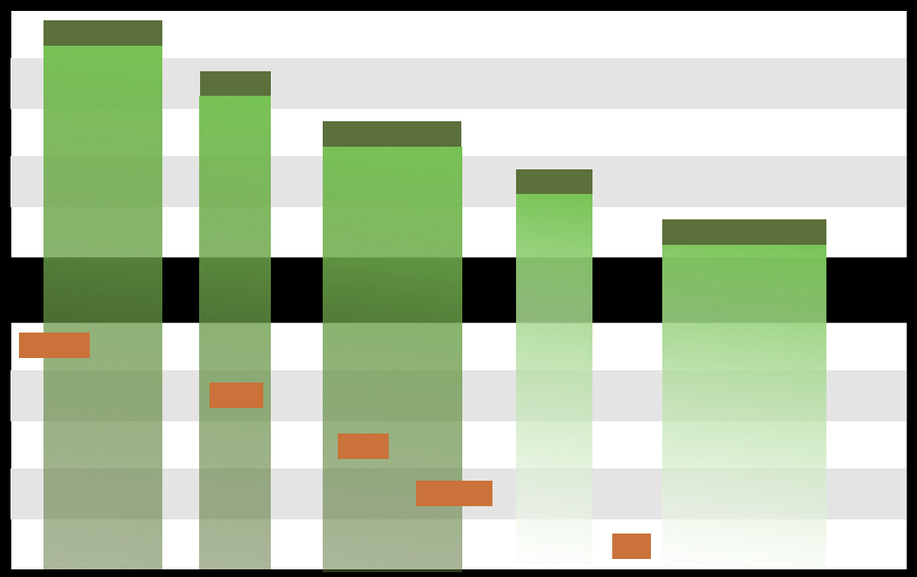
(Identify which exons have Repeats)



Exons



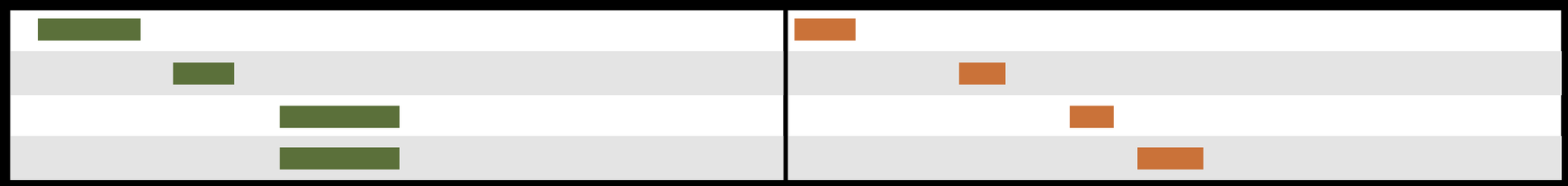
Repeats



Exons

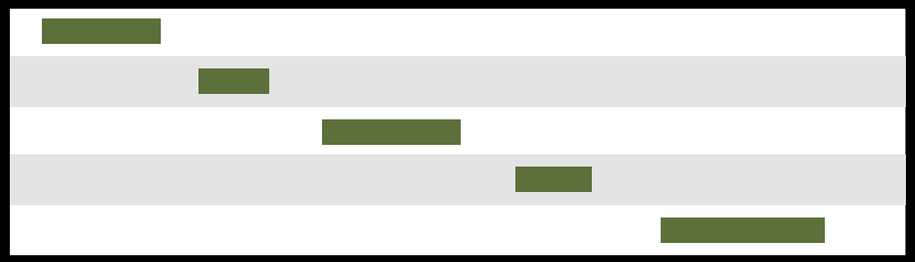
Repeats

Overlap pairings

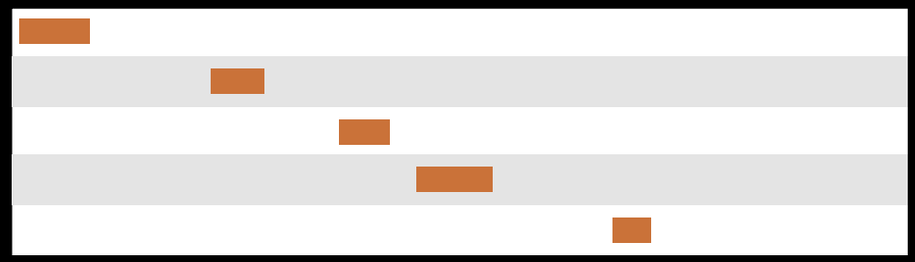


Operate on Genomic Intervals → **Join**

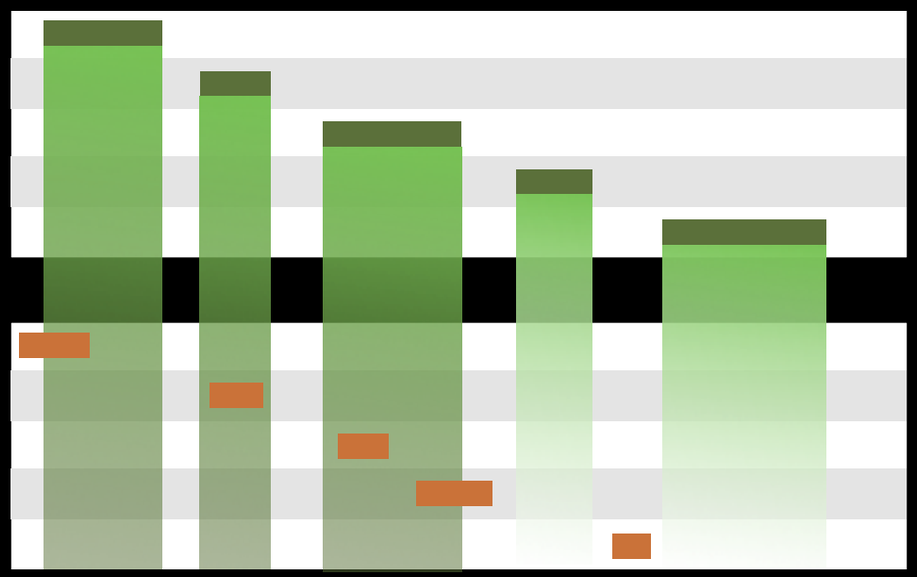
(Identify which exons have Repeats)



Exons



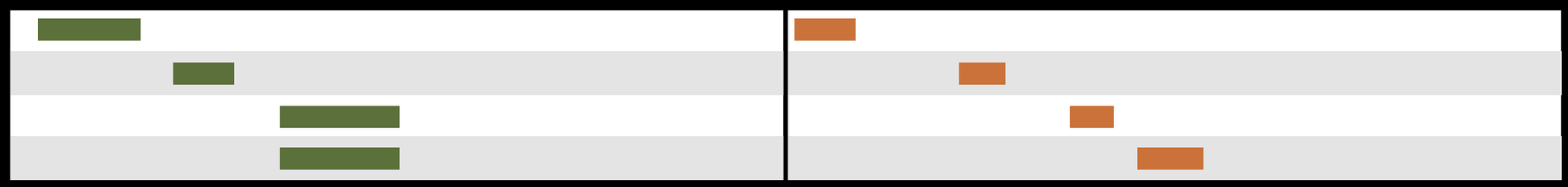
Repeats



Exons

Repeats

Overlap pairings

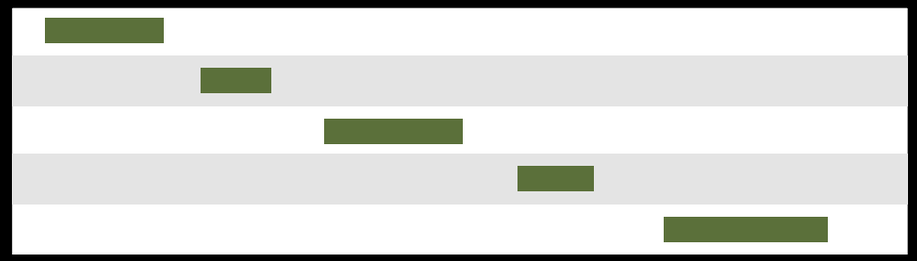


Exon overlap counts

Join, Subtract, and Group → Group

(Count Repeats per exon)

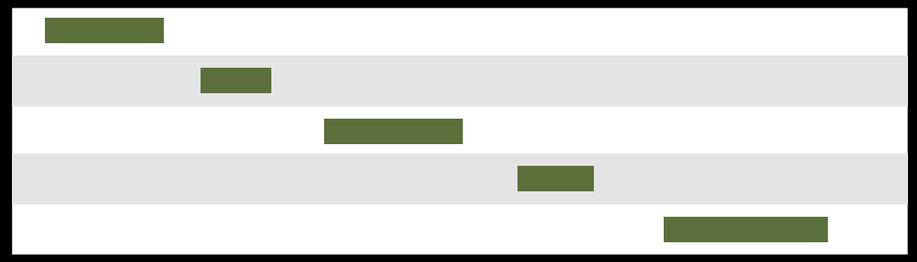
Exon overlap counts



Exons

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

Exon overlap counts



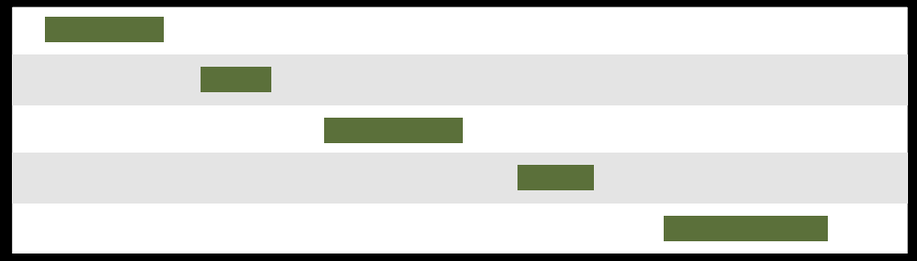
Exons

Join on exon name

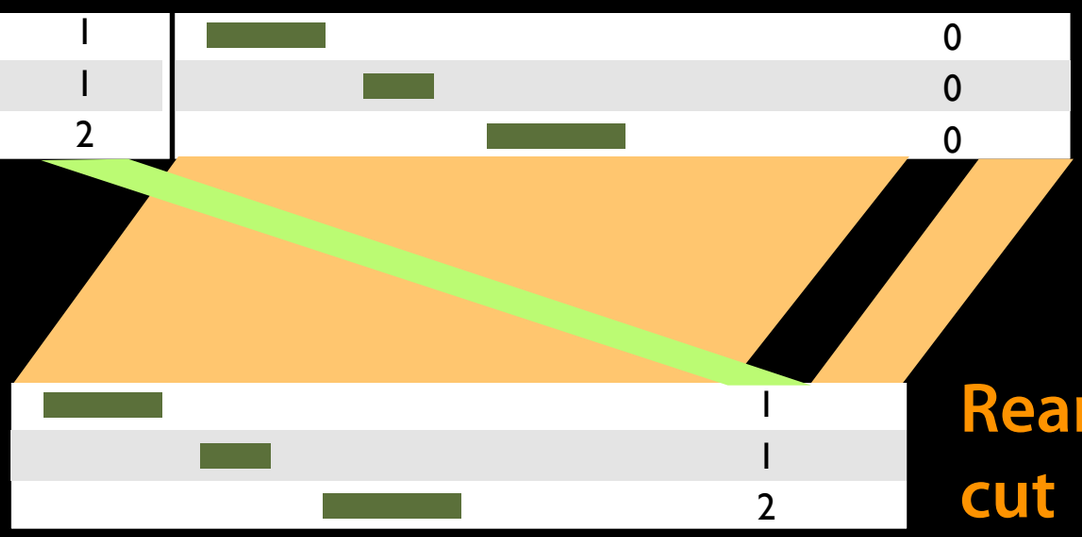
Join, Subtract, and Group → Join

(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

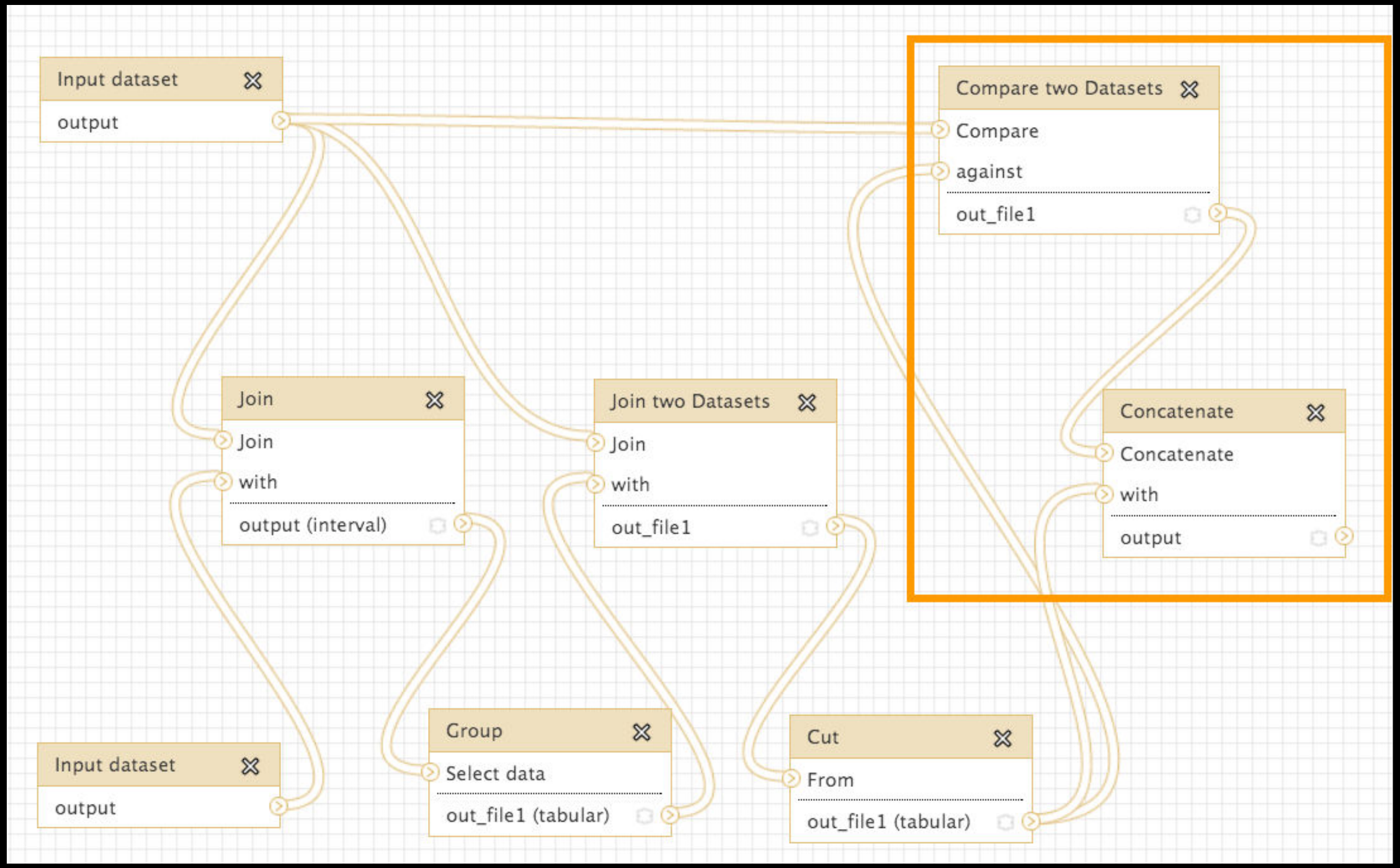
9:00 Welcome
9:30 Basic Analysis with Galaxy
10:45 Break
11:15 Basic Analysis (continued)
12:00 Basic Analysis into Reusable Workflows
12:30 Lunch (on your own)
13:30 ChIP-Seq Analysis
15:30 Break
16:00 Genome Assembly Concepts
16:30 Q & A Session
17:00 Done

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

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

Exons & Repeats: Done?

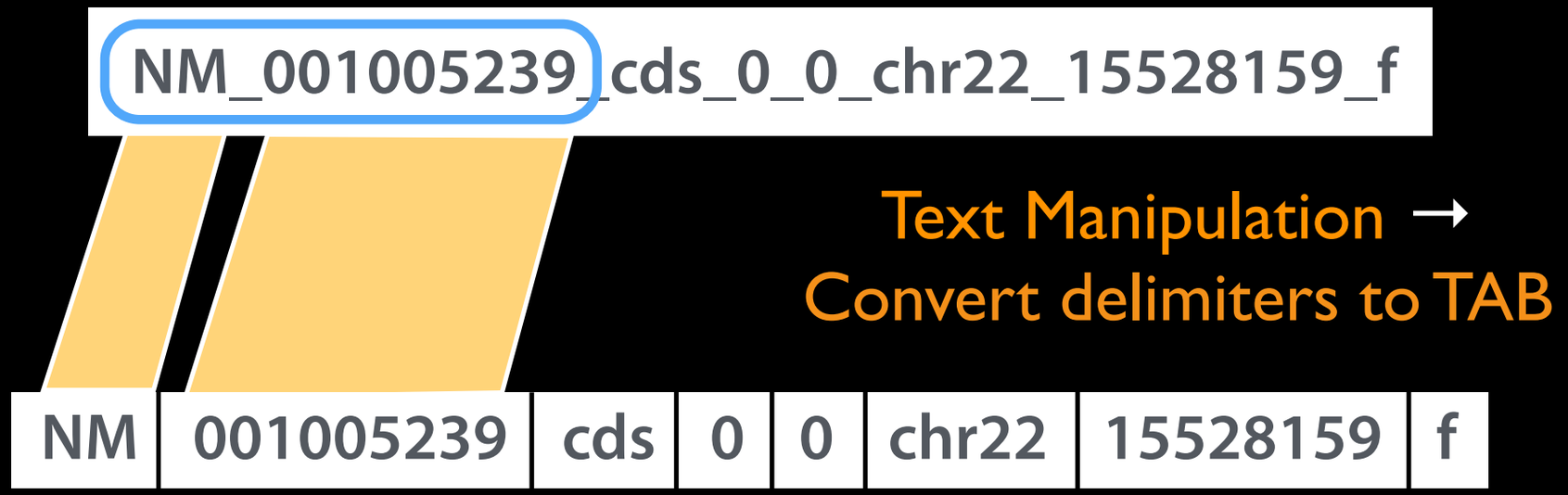
We now know which exons have repeats, and we have that information in a format that can be understood by many tools.

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

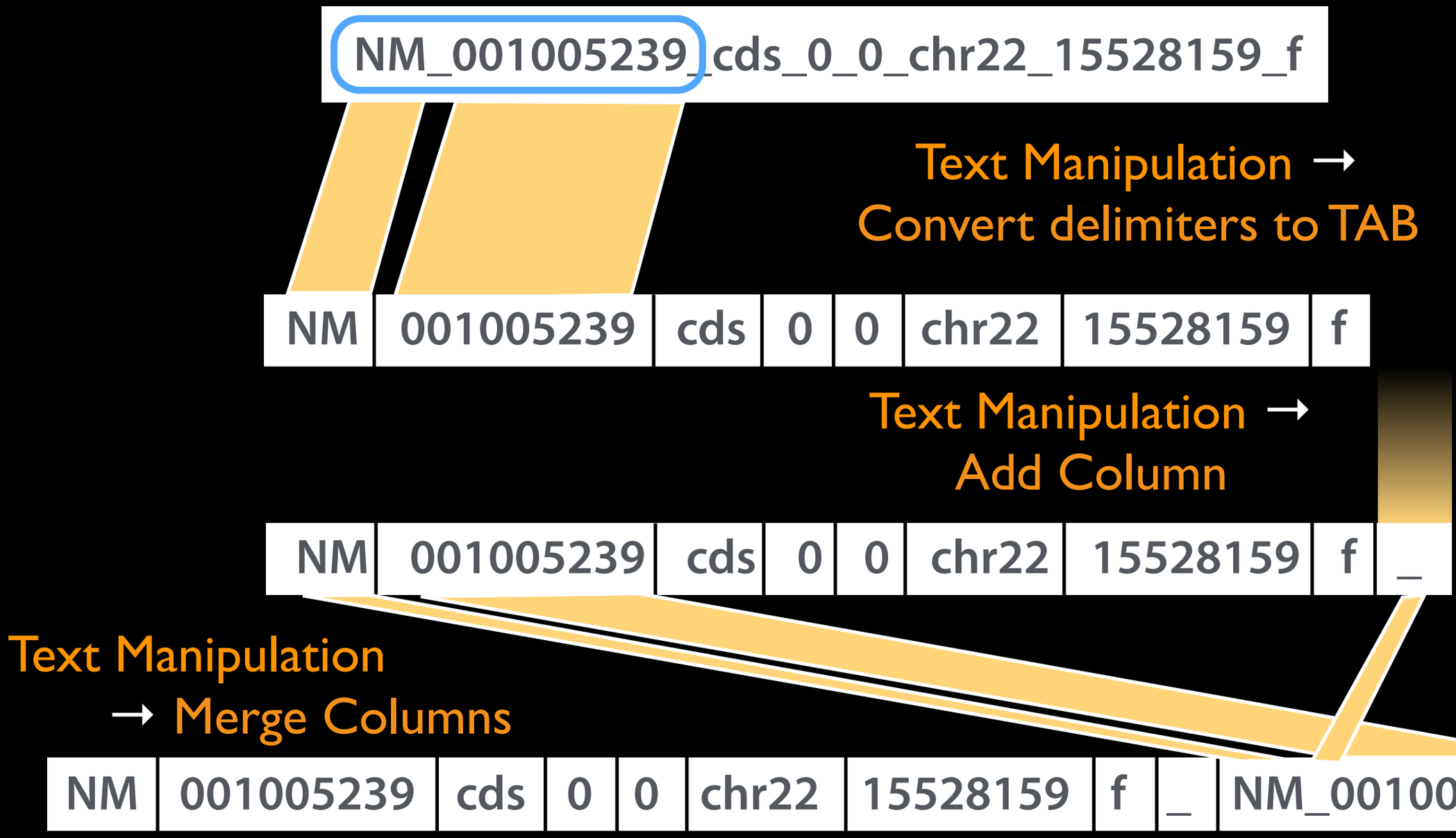
Text Manipulation →
Convert delimiters to TAB

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 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

Text Manipulation
→ Cut

NM_001005239

Get the Genes

Still not done

Text Manipulation → Unique

Got the Gene*s*: Look for GO Enrichment

Let's see what those genes do.

http://geneontology.org/

Agenda: Day 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

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

Workflow:

A series of analysis steps
Can be repeated with different data

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

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

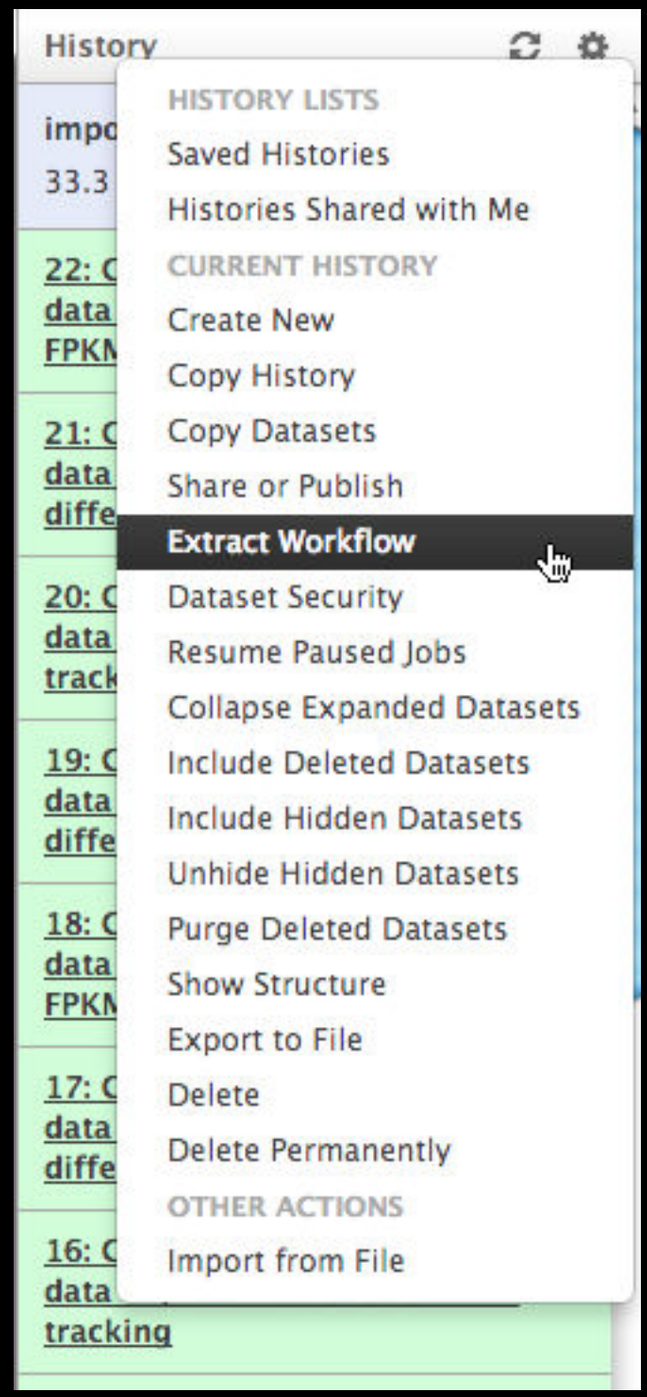
Create a Workflow from a History

Extract Workflow from history

Create a workflow from this history.

Edit it to make some things clearer.

(cog) → Extract Workflow



Run / test it

Guided: rerun with same inputs

Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions

Agenda: Day 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

Agenda: Day 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

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/

ChIP-Seq Analysis: Get the Data

Import

Shared Data → Data Libraries → Training →

ChIP-Seq → Raw Reads

H1hesc_Input_Rep1_chr12.fastq

NGS Data Quality Control

- FASTQ format
- Examine quality in an Chip-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

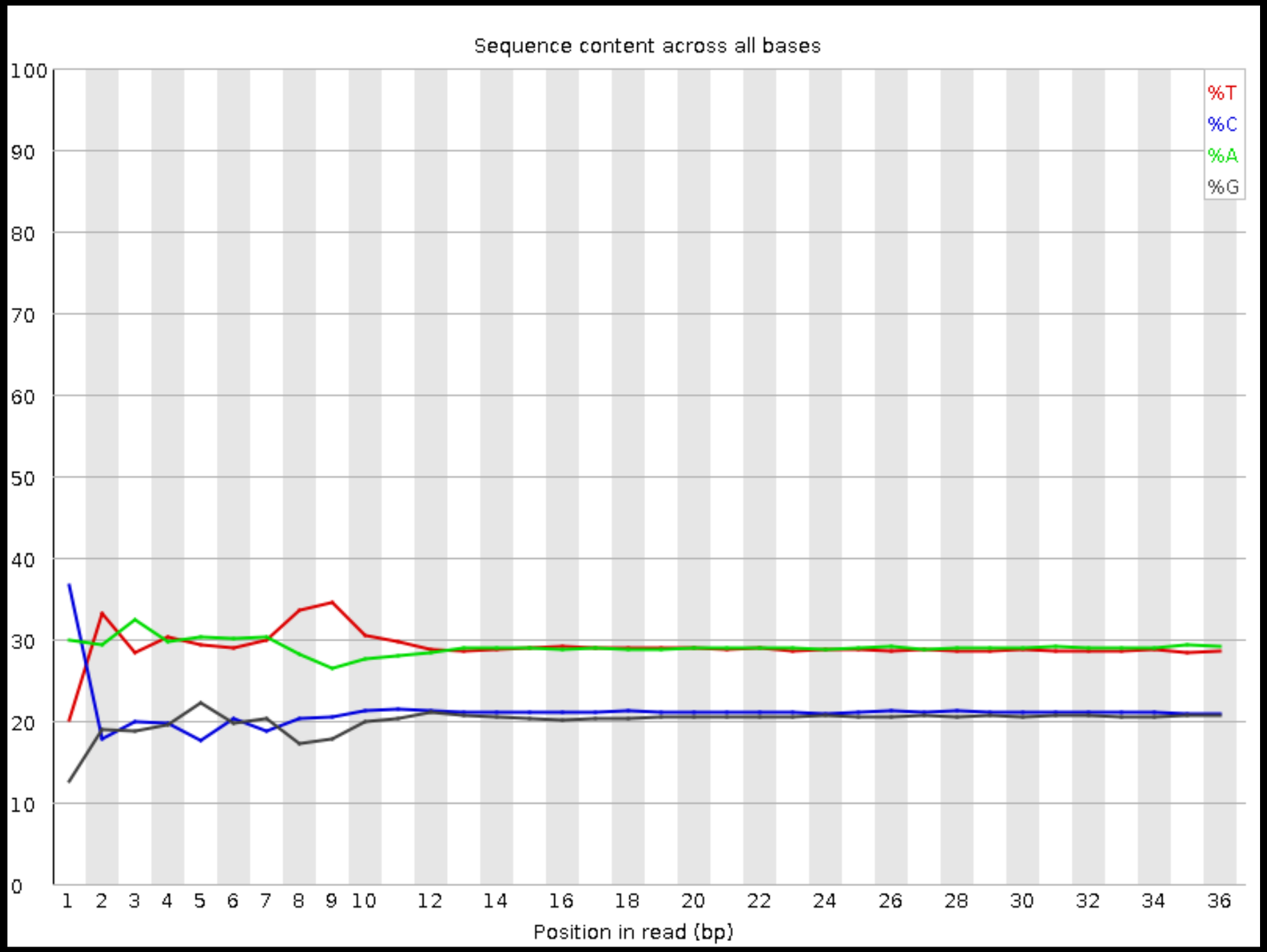
NGS Data Quality: Assessment tools

NGS QC and Manipulation → FastQC

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

http://bit.ly/FastQCBoxPlot

NGS Data Quality: Sequence bias at front of reads?

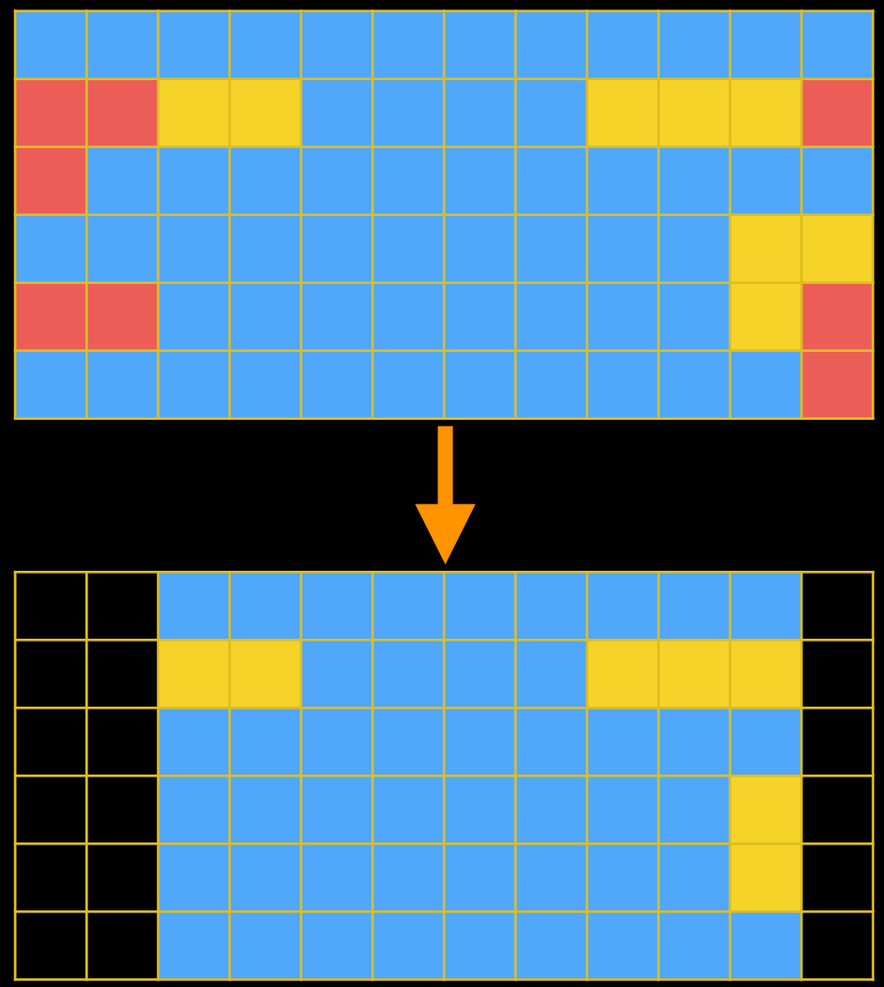


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

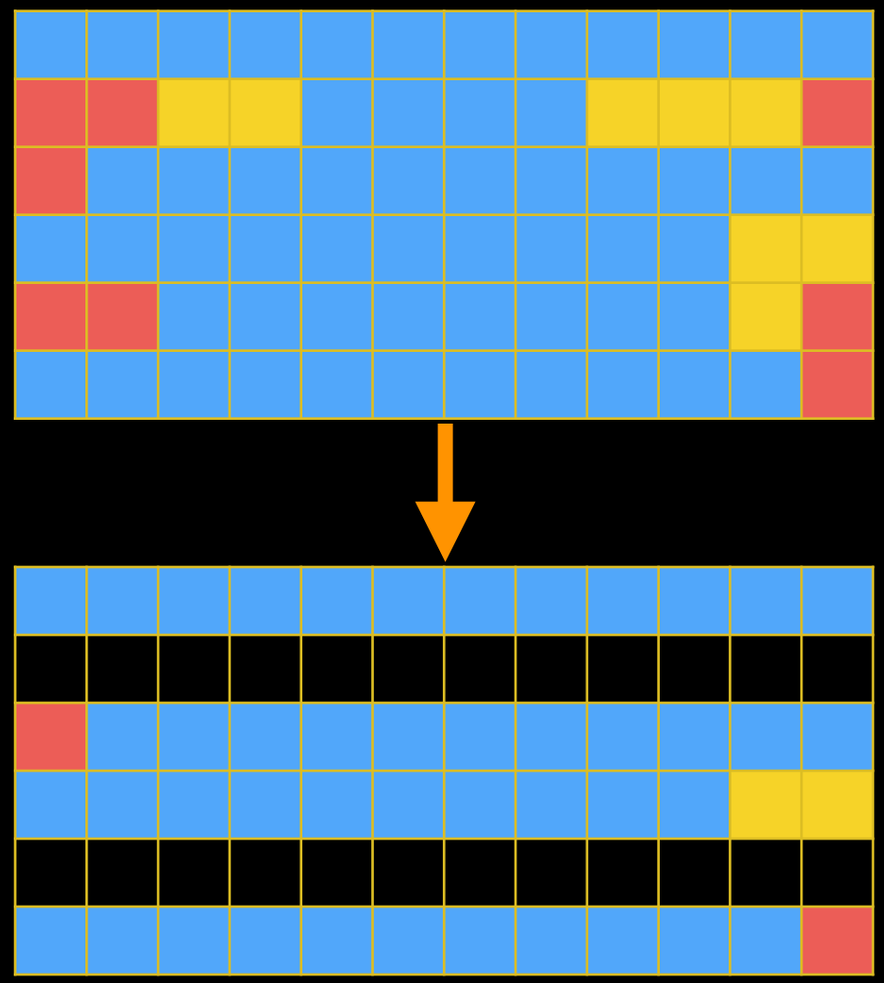
NGS Data Quality: Trim as we see fit

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



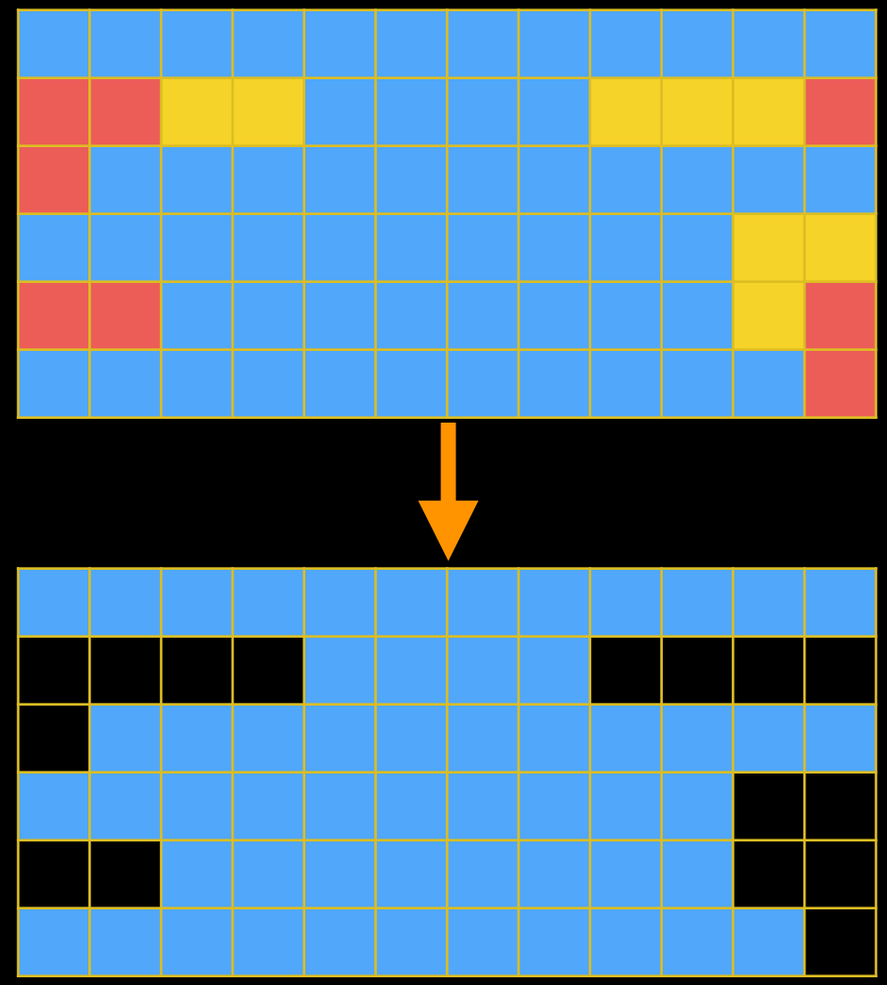
NGS Data Quality: Base Quality Trimming

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

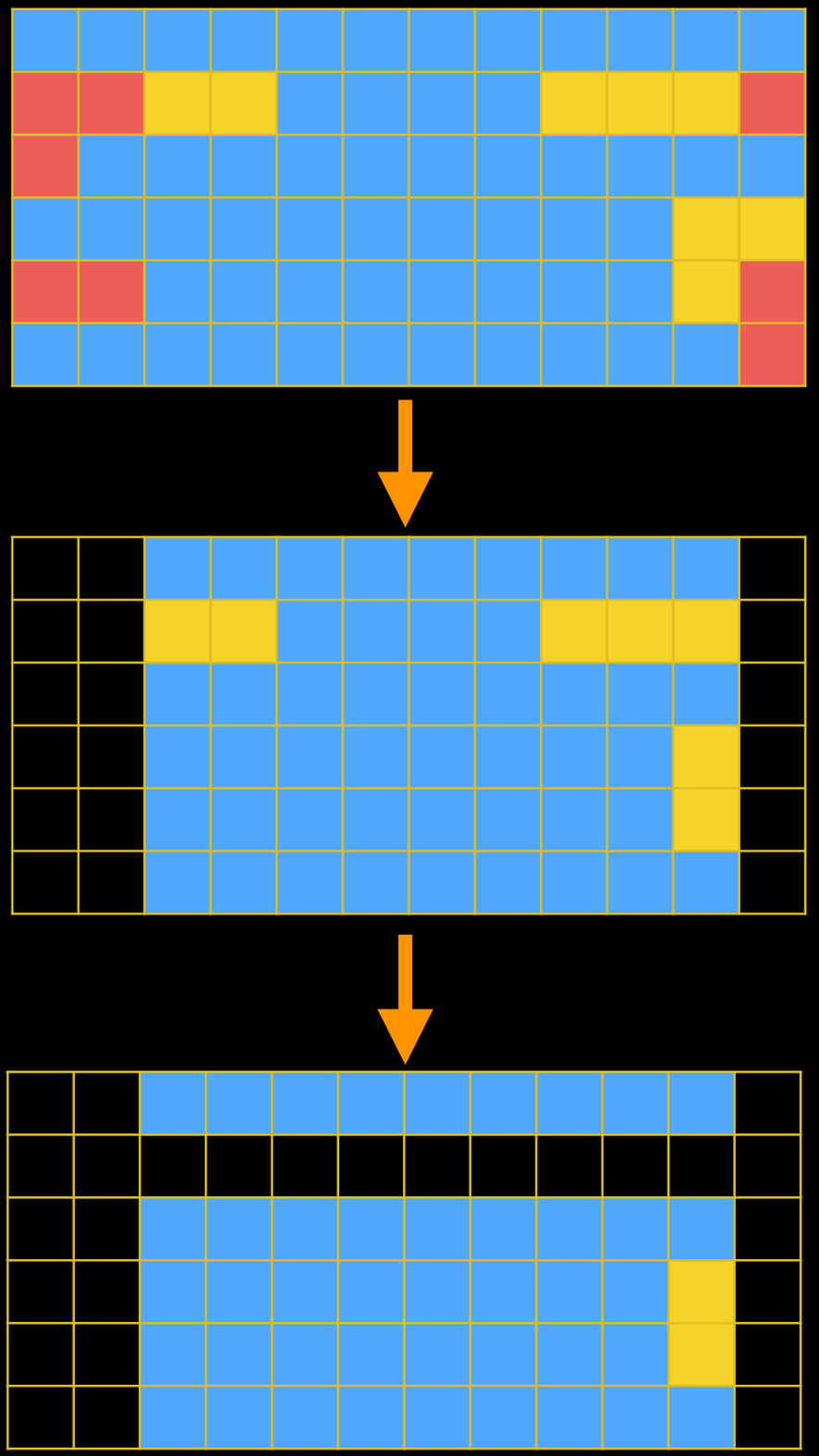


NGS Data Quality: Base Quality Trimming

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



Options are not mutually exclusive



Option 1
(by column)

+

Option 2
(by entire row)

Trim? *As we see fit?*

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

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**

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

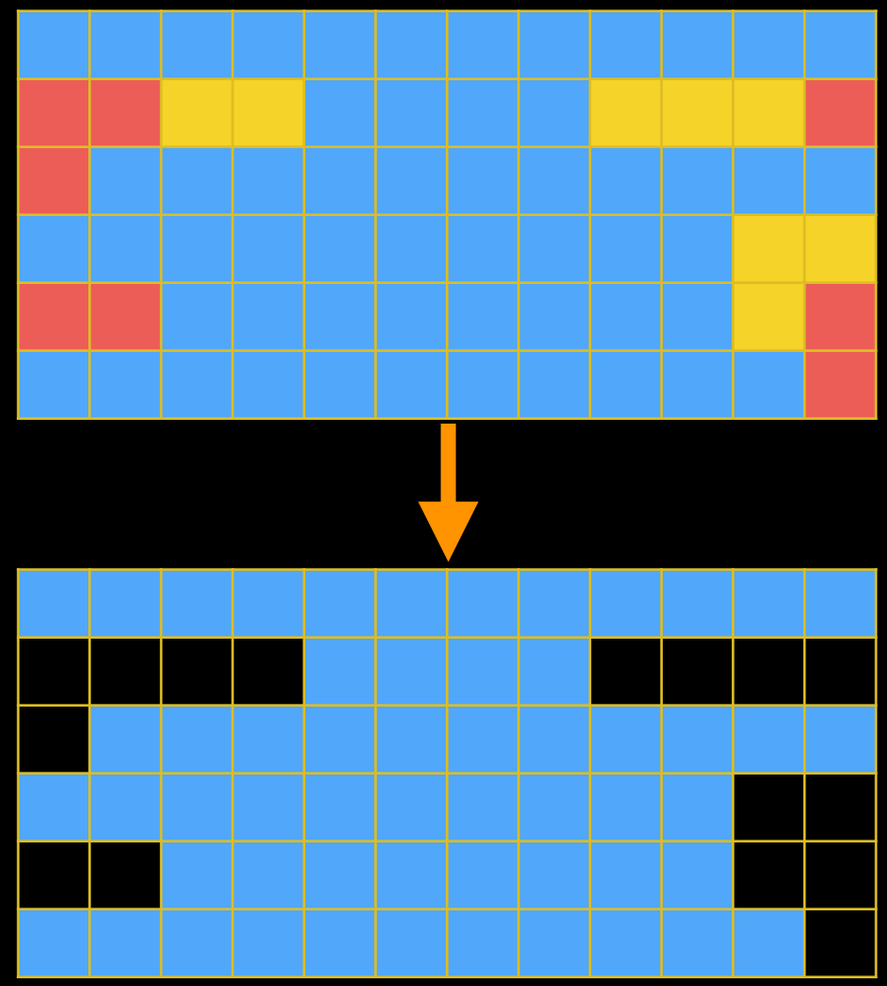
Dear Dr. Impatient,

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



NGS Data Quality: Further reading & Resources

FastQC Documenation

Read Quality Assessment & Improvement

by Joe Fass

From the UC Davis 2013 Bioinformatics Short Course

Manipulation of FASTQ data with Galaxy

by Blankenberg, *et al*.

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: Find Peaks

NGS: ChIP-seq → MACS2 callpeak

Treatment File → Nanog Rep 1

Control File → Nanog Rep2 BAM file

Control File → H1hesc_Input_Rep2_chr12 Mapped BAM file

Outputs → Everything except summits

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

ChIP-Seq Analysis: Visualize Results

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

Reference → genes_chr12.gtf

Launch a Trackster visualization and bring in

the Peaks in BED format

the Bedgraph Treatment

the Bedgraph Control

the gene definitions

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

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

Add the Nanog cluster output to your visualization

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

Add the Pou5f1 cluster output to your visualization

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 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

Agenda: Day 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

Evolutionary Applications

Open Access

Evolutionary Applications ISSN 1752-4571

REVIEWS AND SYNTHESIS

A field guide to whole-genome sequencing, assembly and annotation

Robert Ekblom and Jochen B. W. Wolf

Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

Box 2: Before you start

Some important points to consider

- Availability of appropriate computational resources
- Collaboration with sequencing facility and bioinformatics groups
- Plan for amount and type of sequencing data needed
- Does funding allow to produce sufficient sequence coverage? If not, alternative approaches should be considered rather than producing a poor, low coverage, assembly
- Familiarization with data handling pipelines and file formats (see below)
- High-quality DNA sample (with individual metadata)
- Plan for analyses and publication

Basic considerations

Genome assembly is a challenging problem that requires time, resources and expertise. Before engaging in a genome sequencing project, it should thus be carefully considered whether a genome reference sequence is strictly necessary for the purpose in question.

it needs to be considered whether sufficient financial and computational resources are available to produce a genome of satisfactory quality. If funding is not available to obtain the appropriate read depth, it is advisable to utilize alternative approaches where possible (such as genotyping-by-sequencing or transcriptome sequencing), rather than settle for low-coverage whole-genome sequencing data. The latter would be a waste of funding, effort and time.

even more encouragement from Ekblom & Wolf

- it is essentially impossible to sequence and assemble all nucleotides in the genome (Ellengren 2014)
- there will also be some degree of error in the characterized genome sequence
- every genome assembly is the result of a series of assembly heuristics and should accordingly be treated as a working hypothesis
- it is often not realistic to aim for a chromosome level assembly

Best Practices

- Use several libraries covering different and longer insert sizes
- If using only short reads, ~100x coverage is needed. Suggested breakdown for mammals:
  - 45x coverage with short insert
  - 45x coverage with medium insert (3-10kb)
  - 1-5x coverage with long insert (10-40kb)
  - From Nagarajan and Pop, 2013

Best Practices

- Estimate genome size, sequencing error rates, repeat content and amount of genome duplication
- Can perform a pilot study to get these estimates.
- More repeats or duplication mean higher coverage
- Use inbred, parthenogenic or gynogenetic individuals. Heterozygosity is not your friend.

NGS Assembly: What next?

Scaffolding

Want to tie together those contigs into larger units called scaffolds.

Some software solutions for this.

Can also use related genomes.

Get more reads, possibly on a different platform,

or different insert length.

These can be provided at initial assembly time.

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam[1*†], Joseph N Fass[1†], Anton Alexandrov[36], Paul Baranay[2], Michael Bechner[39], Inanç Birol[33], Sébastien Boisvert[10,11], Jarrod A Cha
Wen-Chi Chou[14,16], Jacques Corbeil[1
Scott Emrich[3], Pavel Fedotov[36], Nun
Sante Gnerre[22], Élénie Godzaridis[11],
Joseph B Hiatt[41], Isaac Y Ho[20], Jason
Huaiyang Jiang[32], Sergey Kazakov[36],
Tak-Wah Lam[29], Dominique Lavenie
Yue Liu[32], Ruibang Luo[28,29], Iain Mac
Delphine Naquin[8,9], Zemin Ning[34], T
Francisco Pina-Martins[31], Michael Pla
Stephen Richards[32], Daniel S Rokhsa
David C Schwartz[39], Alexey Sergushi
Jared T Simpson[34], Henry Song[32], Fe
Jun Wang[28], Kim C Worley[32], Shuan
Shiguo Zhou[39] and Ian F Korf[1*]

NGS Assembly: What's *better*?

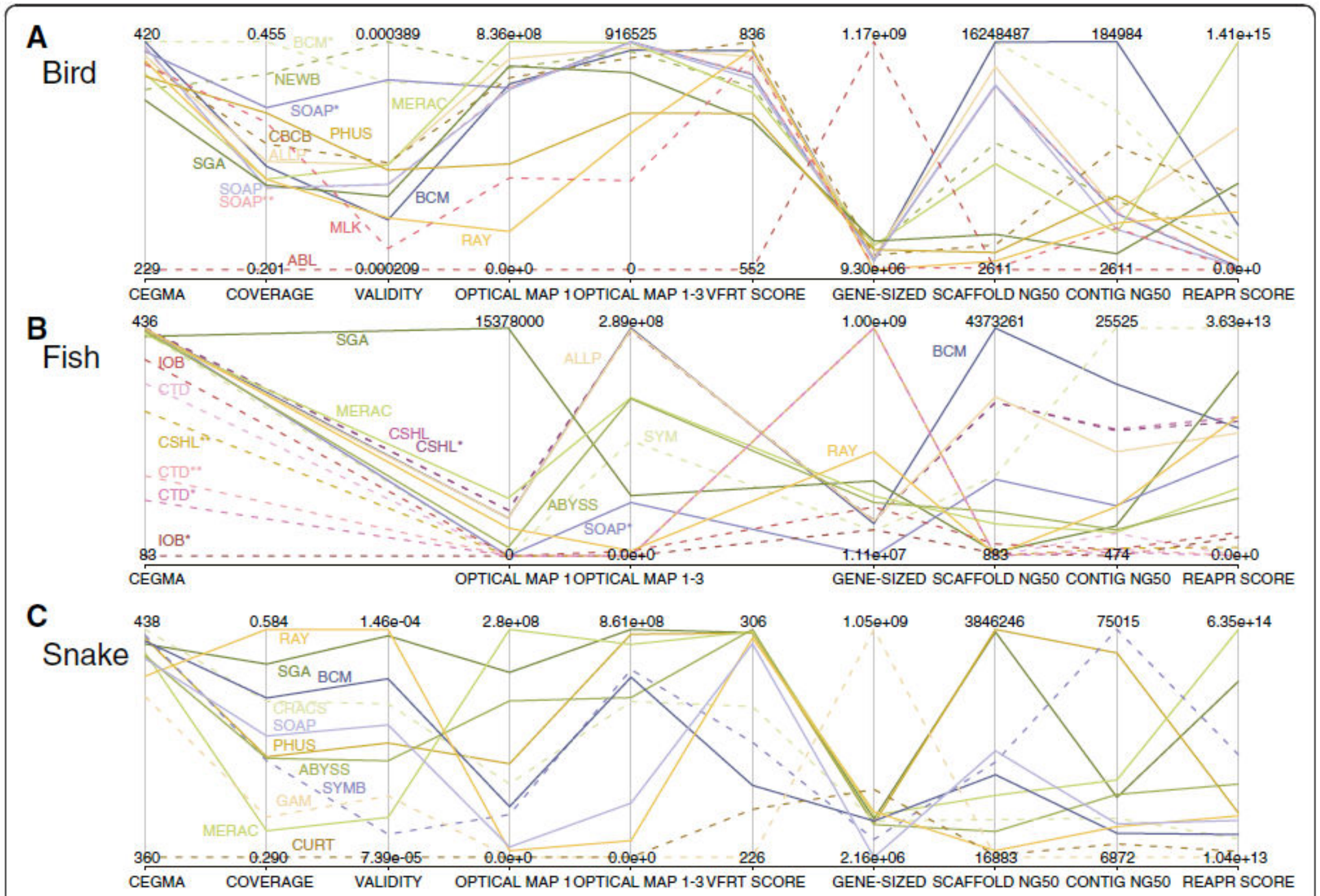


Figure 21 Parallel coordinate mosaic plot showing performance of all assemblies in each key metric. Performance of bird, fish, and snake

NGS Assembly: Resources and Reading

Beginner's guide to comparative bacterial genome analysis using next-generation sequence data

Bacterial Comparative Genomics Tutorial

By David J Edwards and Kathryn E Holt

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Bradnam, *et al.*

Whole Genome Assembly and Alignment

Michael Schatz

Agenda: Day 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

Agenda: Day 1

9:00 Welcome

9:30 Basic Analysis with Galaxy

10:45 Break

11:15 Basic Analysis (continued)

12:00 Basic Analysis into Reusable Workflows

12:30 Lunch (on your own)

13:30 ChIP-Seq Analysis

15:30 Break

16:00 Genome Assembly Concepts

16:30 Q & A Session

17:00 Done

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

clements@galaxyproject.org

http://bit.ly/glaxy2015slides