DNA Sequence Bioinformatics Analysis with the Galaxy Platform

University of São Paulo, Brazil
28 July - 1 August 2014

Dave Clements
Johns Hopkins University

Robson Francisco de Souza
University of São Paulo

José Belizario
University of São Paulo

The Week's Agenda

Mon	Introductions: Cloud Computing, Nuvem Cloud, Basic Analysis in Galaxy
Tues	Nuvem or AWS?, Workflows, Sharing, Quality Control, ChIP-Seq
Wed	Genome Assembly, RNA-Seq
Thur	RNA-Seq continued, SNP and Variant Calling
Fri	Intro to Command Line, Genome Annotation using MAKER, CloudMan and AWS

bit.ly/gxyusp2014

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing

20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis

20 minute Break at around 3:20

5:00 Done

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing

20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis

20 minute Break at around 3:20

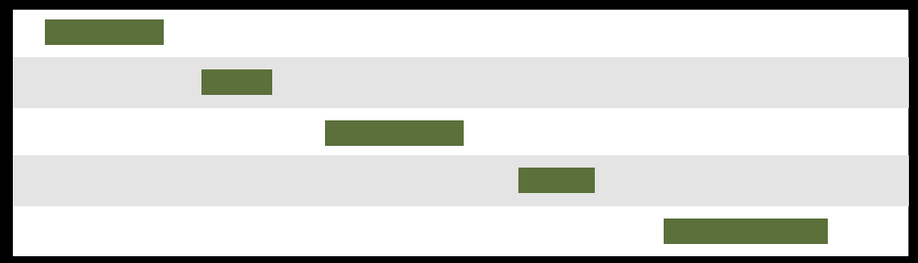
5:00 Done

Exons & Repeats: Exercise

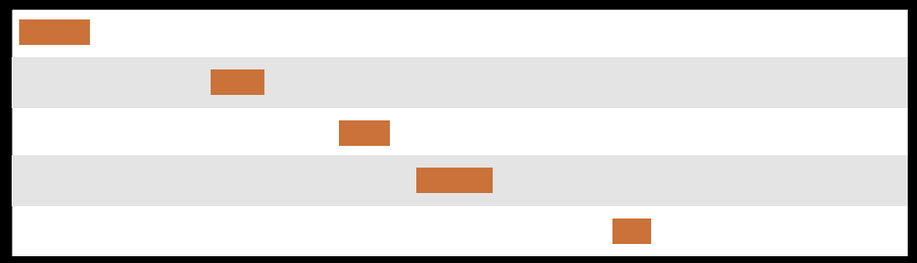
Include exons with no overlaps in final output.
Set the score for these to 0.

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

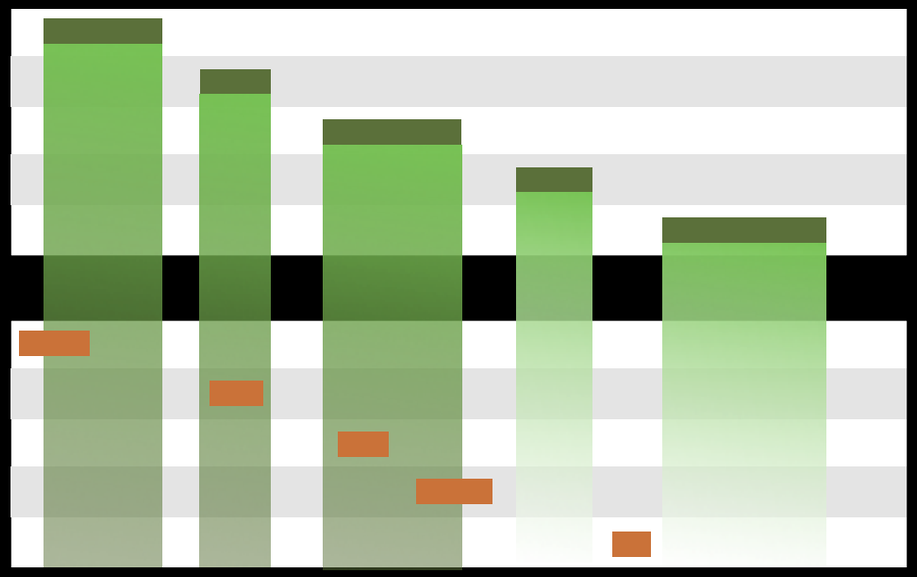
First, *what exactly did we do yesterday?*



Exons



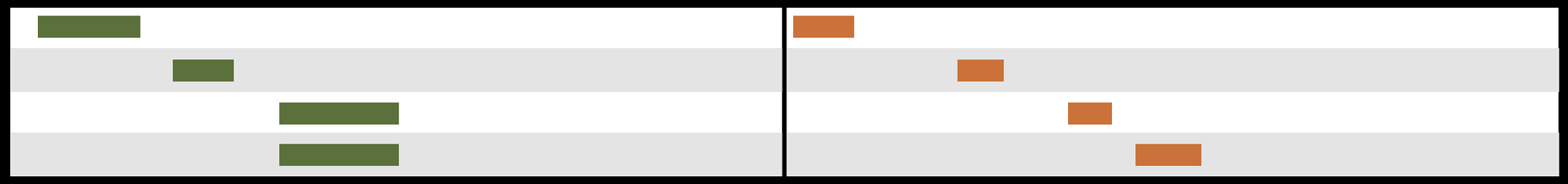
Repeats



Exons

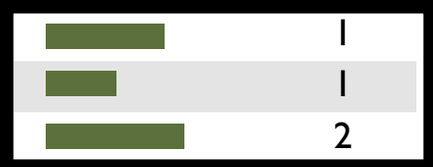
Repeats

Overlap pairings

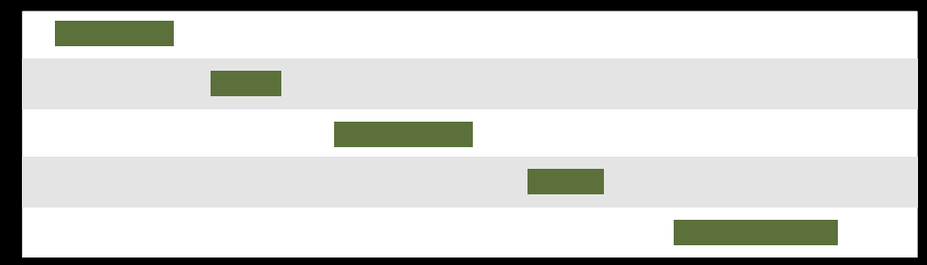


Exon overlap counts

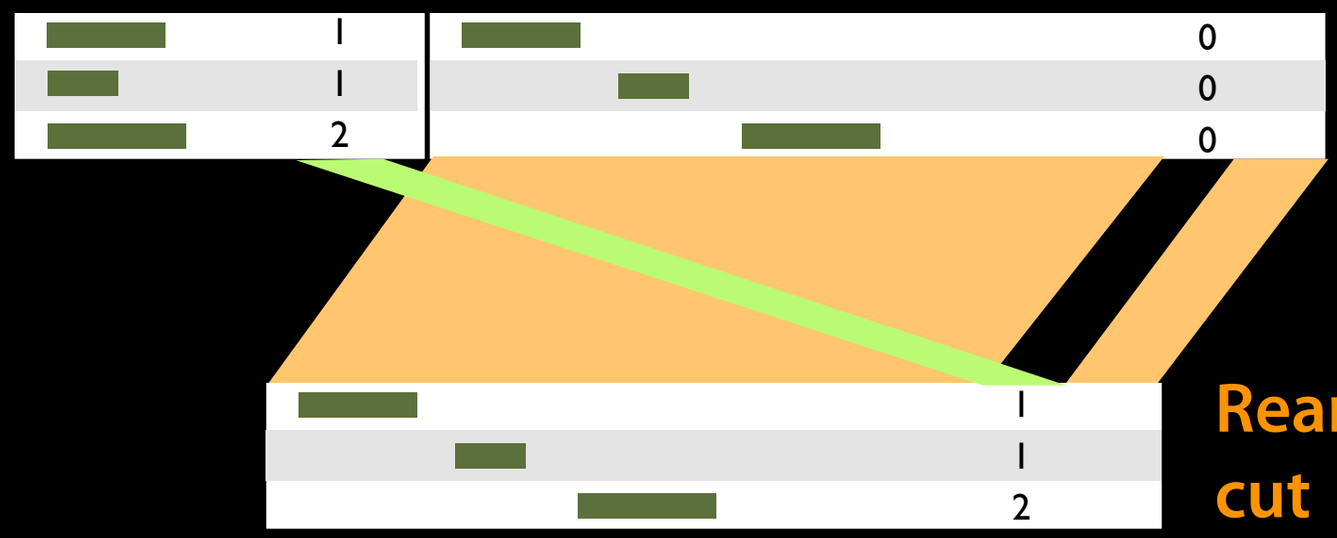
What we did yesterday: counting



Exon overlap counts



Exons



Join on exon name

Rearrange columns w/ cut

What we did yesterday: formatting

Galaxy / Nuvem USP Analyze Data Workflow Shared Data Visualization Help Us

Tools

search tools

Get Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Statistics
Graph/Display Data
Evolution
Motif Tools
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Simulation
Phenotype Association

Obrigado! Welcome Paulo
Galaxy on the Nuv

Data Libraries
Data Libraries Beta
Published Histories
Published Workflows
Published Visualizations
Published Pages

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy / Nuvem USP Analyze Data Workflow Shared Data Visualization Help User

Published Histories | outreach | 101: Overlapping Exons and Repeats

Import history

Make a copy of this history and switch to it

About t

101: Overlapping Exons and Repeats

3.5 MB

search datasets

Dataset	Annotation
1: Exons, chr22	
2: Repeats, chr22	
3: Join on data 2 and data 1	
4: Group on data 3	
5: Join two Datasets on data 1 and data 4	
6: Exons with overlapping repeats	

Autho

outreac

Relate

All publ

Publishe

Rating

Commu

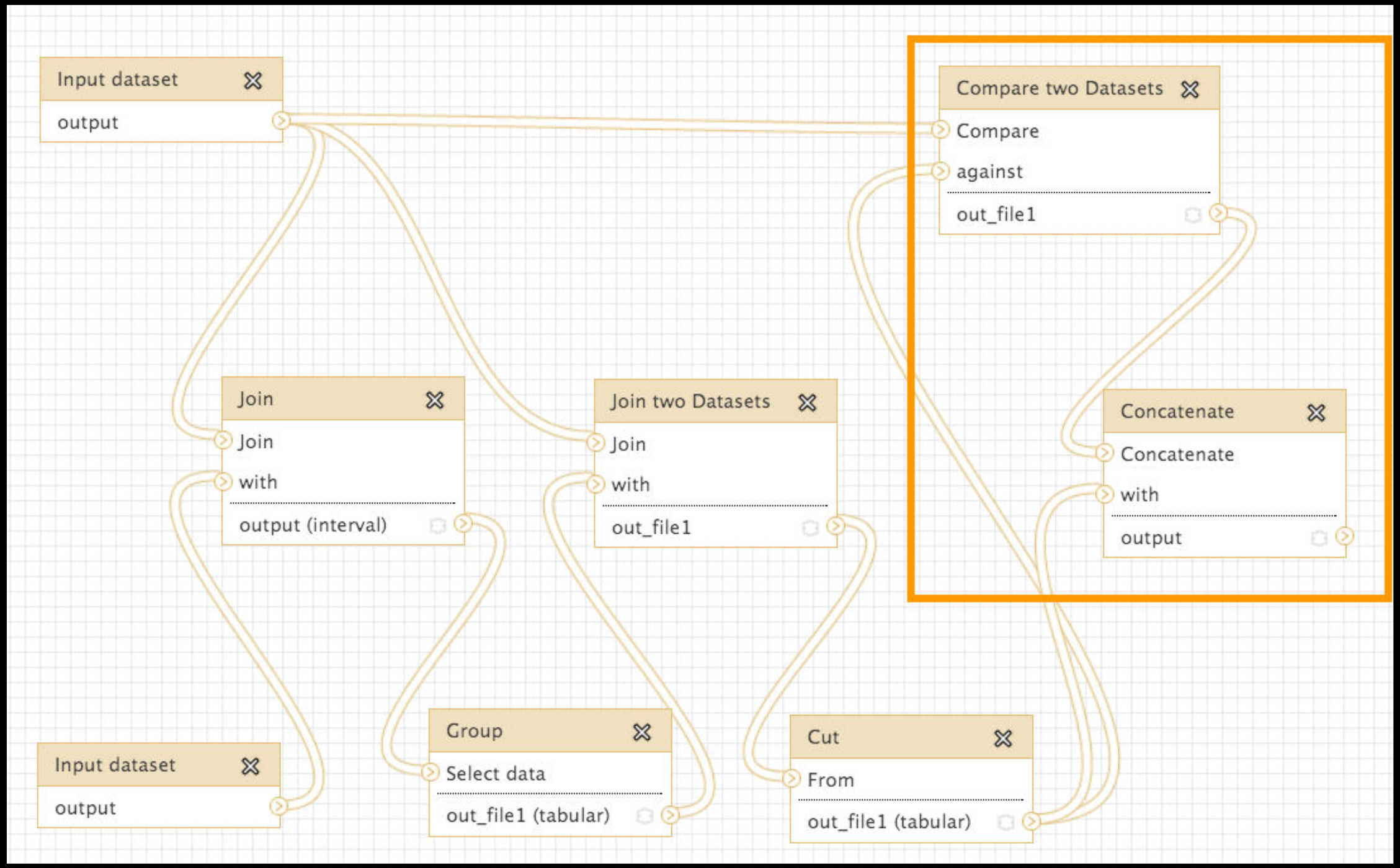
(0 ratings,

Tags

Commu

Note: In your solution, you can take advantage of the fact that Exons already have 0 scores.

One Possible Solution



Solution from Stanford Kwenda and Caron Griffiths in Pretoria.

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

Basic Analysis: Further reading & Resources

http://usegalaxy.org/galaxy101

https://vimeo.com/76343659

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?*

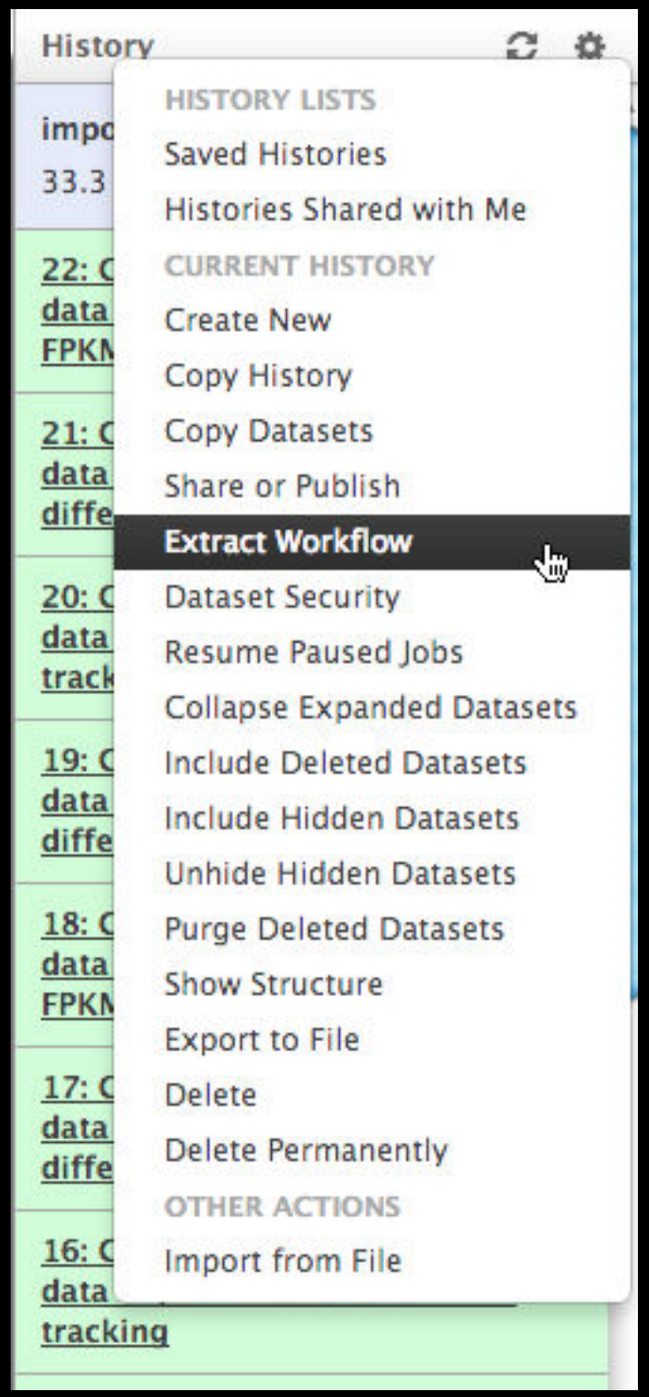
- The analysis we just finished was about
  - Human chr22
  - Overlap between exons and Repeats
- But, ...
  - there is nothing inherent in the analysis about humans, exons or repeats
  - It is a series of steps that sets the score of one set of features to the number of overlaps from another set of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

More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Let's all share...

Sharing & Publishing enables Reproducibility

Reproducibility: Everybody talks about it, but ...

Galaxy aims to push the goal of reproducibility from the bench to the bioinformatics realm

All analysis in Galaxy is recorded without any extra effort from the user.

Histories, workflows, visualizations and ***pages*** can be shared with others or published to the world.

Sharing & Publishing enables Reproducibility

CSH PRESS GENOME RESEARCH

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

- » Abstract *Free*

Current Issue

October 2010, 20 (10)

Sharing & Publishing enables Reproducibility

CSH PRESS GENOME RESEARCH

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

- » Abstract *Free*

Current Issue

October 2010, 20 (10)

Footnotes

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]

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

Published Pages | aun1 | Windshield Splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis
Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From

About this Page

Author

aun1

Related Pages

All published pages
Published pages by aun1

Rating

Community
(6 ratings, 5.0 average)

Tags

Community:
paper | galaxy | megan

http://usegalaxy.org/u/aun1/p/windshield-splatter

Sharing for Galaxy Administrators Too

Data Libraries

Make data easy to find

Genome Builds

Care about a particular subset of life?

Galaxy Tool Shed

Wrapping tools and datatypes

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing
20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis
20 minute Break at around 3:20

5:00 Done

ChIP-Seq: FASTQ data and quality control

http://scriptogr.am/ohofmann

By Shannan Ho Sui

H3ABioNet

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

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

ChIP-Seq Analysis: Get the Data

Import

Shared Data → Data Libraries →

ChIP-Seq Datasets → Unfiltered Reads

H1hesc_Input_Rep1_chr12_unfiltered.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.

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!

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./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

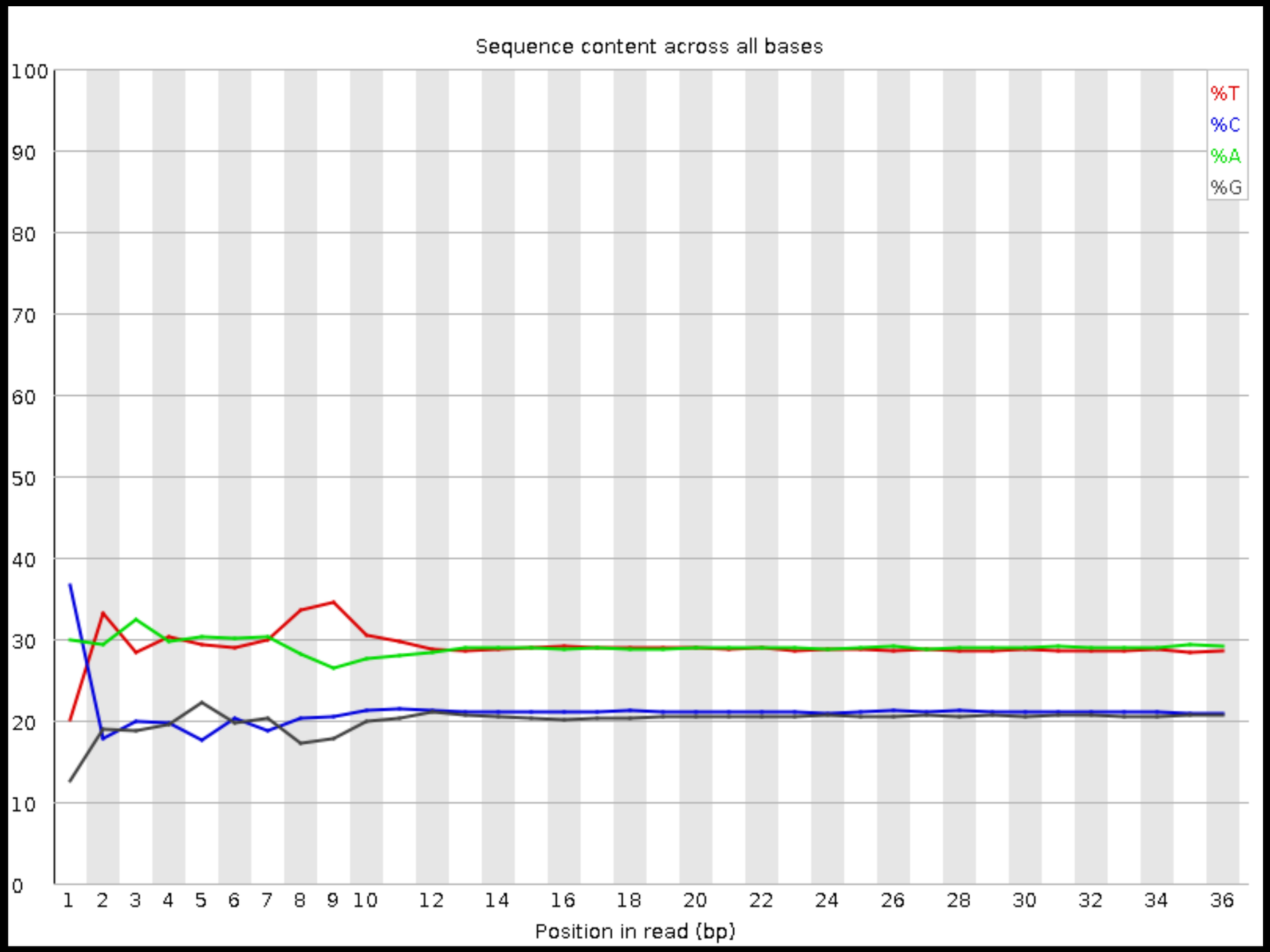
NGS Data Quality: Assessment tools

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?

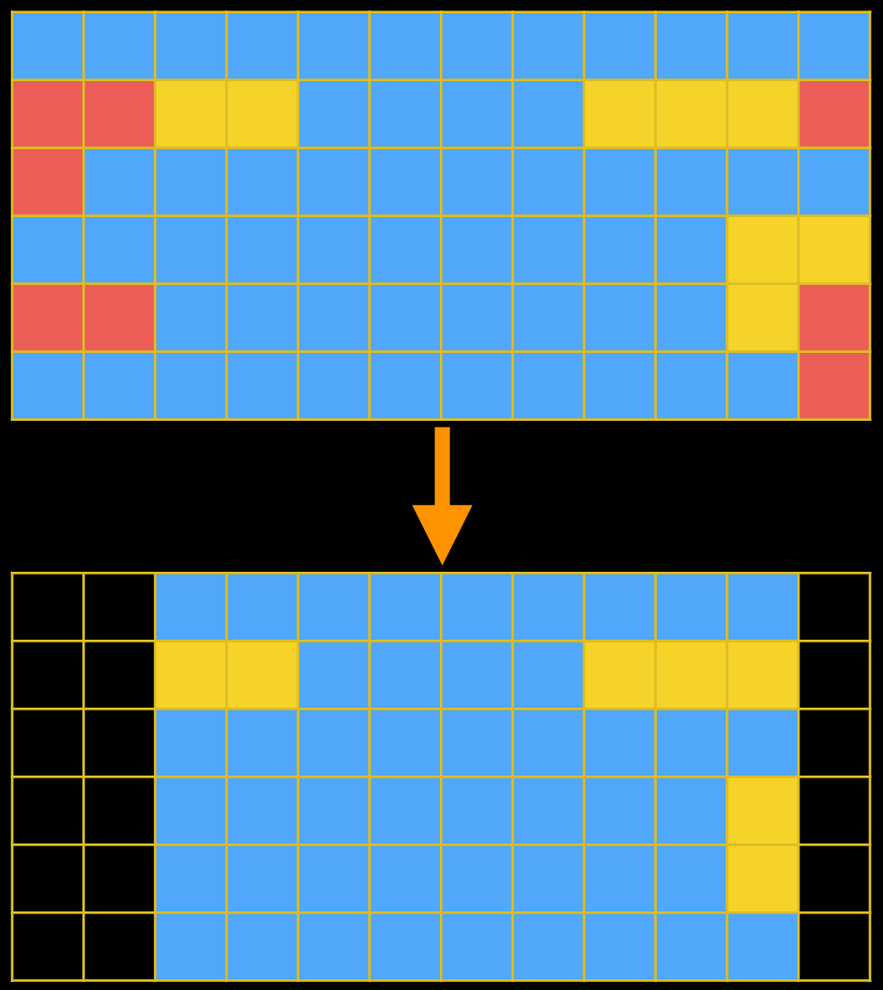


From a sequence specific bias that is caused by use of random hexamers in library preparation.

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

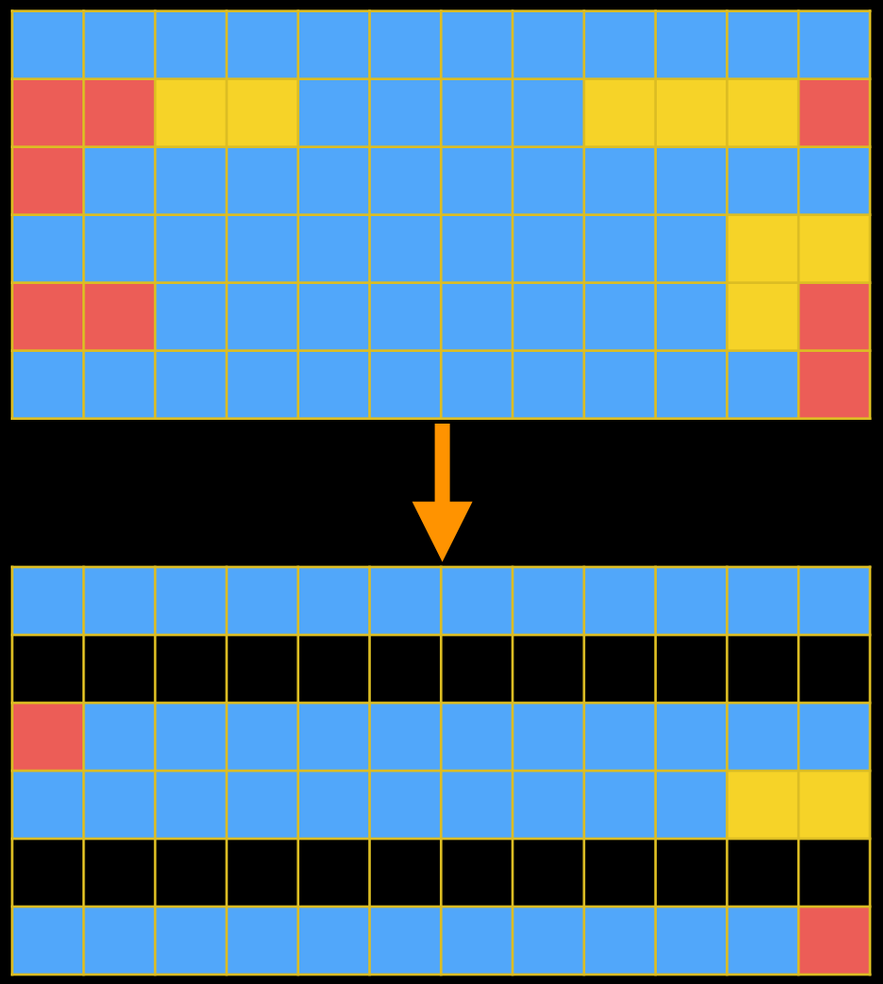
NGS Data Quality: Trim as we see fit

- 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



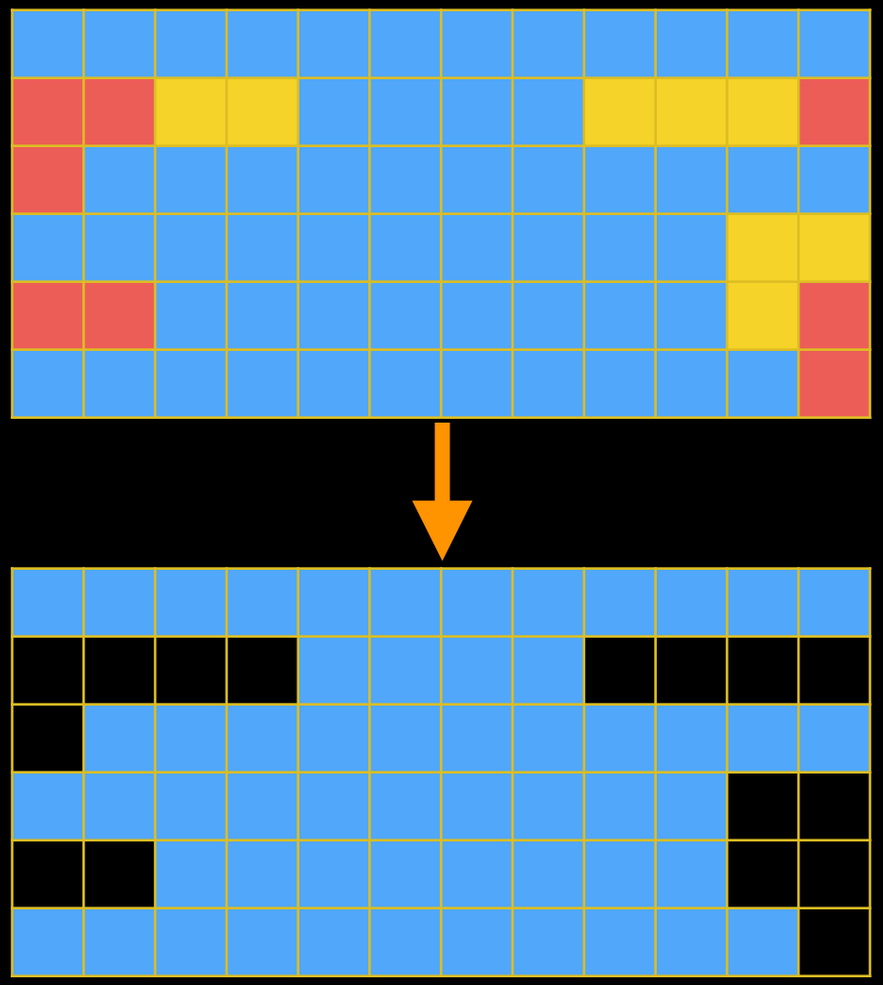
NGS Data Quality: Base Quality Trimming

- ~~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.

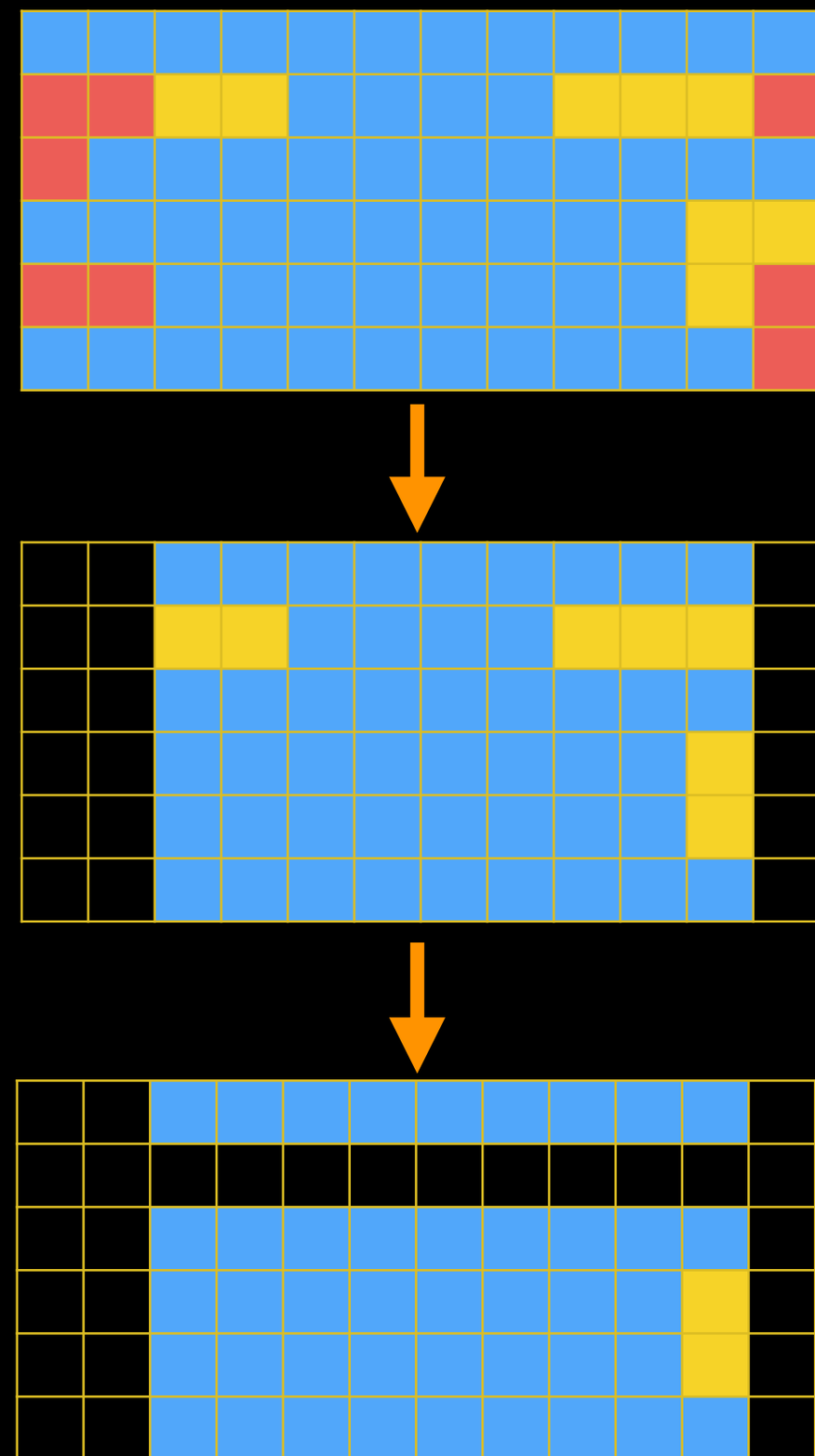


NGS Data Quality: Base Quality Trimming

- 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?*

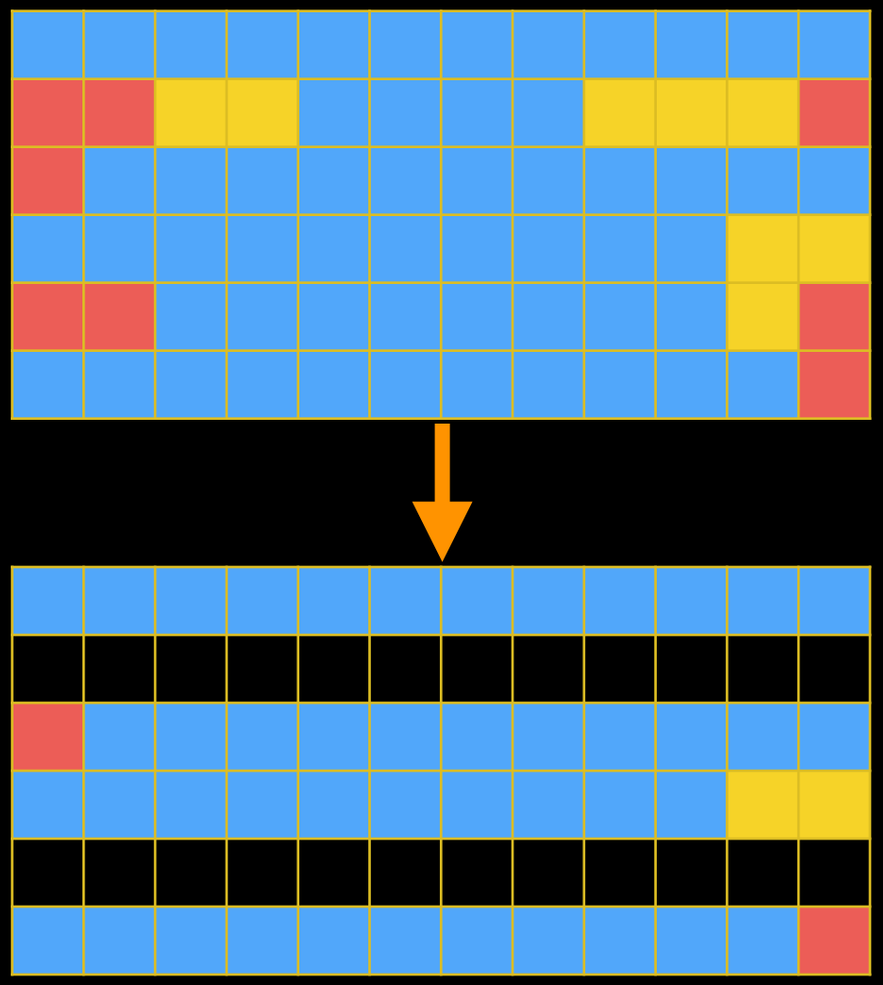
- Choice depends on downstream tools
- Find out assumptions & requirements for downstream tools and make appropriate choice(s) now.
- How to do that?
  - Read the tool documentation
  - http://biostars.org/
  - http://seqanswers.com/
  - http://galaxyproject.org/search

SEQanswers
the next generation sequencing community

Biostar
BIOINFORMATICS EXPLAINED

Does MACS care? Maybe

- ~~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: 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*.

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing
20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis
20 minute Break at around 3:20

5:00 Done

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing
20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis
20 minute Break at around 3:20

5:00 Done

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing
20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis
20 minute Break at around 3:20

5:00 Done

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing
20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis
20 minute Break at around 3:20

5:00 Done

Tuesday's Agenda

9:00 Nuvem or AWS?

9:40 Introduction to Using Galaxy, continued
Exercise, Workflows and sharing
20 minute Break at around 10:20

10:50 Next Generation Sequencing (NGS) Data Quality Control

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 NGS QC, continuted

3:00 ChIP-Seq Analysis
20 minute Break at around 3:20

5:00 Done

ChIP-Seq Analysis: Get the Data

Shared Data → Data Libraries →

ChIP-Seq Datasets

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 Exercise: Mapping with Bowtie

Convert BAM to SAM

SAM Tools → BAM-to-SAM

ChIP-Seq Analysis: remove unmapped reads

SAM Tools → Filter SAM

- Click Add a new Flag
- Set Type to The read is unmapped
- Set flag state to No.

ChIP-Seq Analysis: Put mapped reads in BAM

SAM Tools → SAM-to-BAM

Get the the control (already mapped for us)

Shared Data → Data Libraries → Aligned → Import H1hesc_Input_Rep1 Mapped into current history

ChIP-Seq Analysis: Find Peaks

NGS: Peak Calling → MACS

Experiment name → MACS NanogRep1

Tag File → Nanog Rep1 BAM file

Control File → H1hesc_Input_Rep1 Mapped BAM file

Tag Size → 36

Leave MFOLD → 32

Save shifted raw tag count ... → Save (leave resolution at 10)

Check Perform the new peak detection method (futuredir)

ChIP-Seq Analysis: Visualize Results

Look at the HTML report dataset

Launch a Trackster visualization and bring in

- the called peaks
- the Treatment WIG
- the Control WIG
- the gene definitions

ChIP-Seq Analysis: Replicates

Shared Data → Data Libraries → ChIP-Seq Datasets → MACS Outputs

Import Peaks 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

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org