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

University of Cambridge
11-12 June 2015

Dave Clements
Galaxy Project
Johns Hopkins University

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

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)
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

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 ~)
(Identify which exons have Repeats)
Operate on Genomic Intervals → Join
(Identify which exons have Repeats)
Exons

Repeats

Overlap pairings

Exon overlap counts

Join, Subtract, and Group → Group
(Count Repeats per exon)
We’ve answered our question, but we can do better. Incorporate the overlap count with rest of Exon information.
### Exon overlap counts

<table>
<thead>
<tr>
<th>Exon</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Exons

<table>
<thead>
<tr>
<th>Exon</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Join on exon name

<table>
<thead>
<tr>
<th>Exon 1</th>
<th>Count 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

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)
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30</td>
<td>Welcome</td>
</tr>
<tr>
<td>10:00</td>
<td>Basic Analysis with Galaxy</td>
</tr>
<tr>
<td></td>
<td>A worked example demonstrating Galaxy Basics</td>
</tr>
<tr>
<td>10:50</td>
<td>Break</td>
</tr>
<tr>
<td>11:10</td>
<td>Basic Analysis (continued)</td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch (on your own)</td>
</tr>
<tr>
<td>13:30</td>
<td>RNA-Seq Analysis</td>
</tr>
<tr>
<td></td>
<td>Quality Control and Galaxy Workflows</td>
</tr>
<tr>
<td>15:30</td>
<td>Break</td>
</tr>
<tr>
<td>15:50</td>
<td>RNA-Seq Analysis</td>
</tr>
<tr>
<td></td>
<td>Mapping and Splice Junction Identification</td>
</tr>
<tr>
<td>17:00</td>
<td>Done</td>
</tr>
</tbody>
</table>

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

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

Get the gene

Text Manipulation → Convert delimiters to TAB

NM_001005239 cds_0_0_chr22_15528159_f

NM    001005239    cds    0    0    chr22    15528159    f
Get the gene

<table>
<thead>
<tr>
<th>NM</th>
<th>001005239</th>
<th>cds</th>
<th>0</th>
<th>0</th>
<th>chr22</th>
<th>15528159</th>
<th>f</th>
<th>_</th>
</tr>
</thead>
</table>

Text Manipulation → Convert delimiters to TAB

Text Manipulation → Add Column
Get the gene

Get the gene

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

<table>
<thead>
<tr>
<th>NM</th>
<th>001005239</th>
<th>cds</th>
<th>0</th>
<th>0</th>
<th>chr22</th>
<th>15528159</th>
<th>f</th>
</tr>
</thead>
</table>

**Text Manipulation** → Convert delimiters to TAB

<table>
<thead>
<tr>
<th>NM</th>
<th>001005239</th>
<th>cds</th>
<th>0</th>
<th>0</th>
<th>chr22</th>
<th>15528159</th>
<th>f</th>
<th>_</th>
</tr>
</thead>
</table>

**Text Manipulation** → Add Column

<table>
<thead>
<tr>
<th>NM</th>
<th>001005239</th>
<th>cds</th>
<th>0</th>
<th>0</th>
<th>chr22</th>
<th>15528159</th>
<th>f</th>
<th>_</th>
<th>NM_001005239</th>
</tr>
</thead>
</table>

**Text Manipulation** → Merge Columns

<table>
<thead>
<tr>
<th>NM</th>
<th>001005239</th>
<th>cds</th>
<th>0</th>
<th>0</th>
<th>chr22</th>
<th>15528159</th>
<th>f</th>
<th>_</th>
<th>NM_001005239</th>
</tr>
</thead>
</table>

**Text Manipulation** → Cut

<table>
<thead>
<tr>
<th>NM_001005239</th>
<th></th>
</tr>
</thead>
</table>
Get the Genes

Remove duplicate gene names

Text Manipulation → Unique
Got the Genes: 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
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 icon](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

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

RNA-Seq Analysis: Get the Data

Create new history

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

UC Davis Bioinformatics Core

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
GATTTGGGTTTCAAGCAGTATCGATCAATAGTAAATCCATTTGTTCAACTCAGTTT
+
!*((**++)%&%++)(&$(&$).1**+-**'))**55CCF>><><><>CCCCCCC65
```

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

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS

http://en.wikipedia.org/wiki/FASTQ_format
NGS Data Quality: Assessment tools

NGS QC and Manipulation → FastQC

Generates summary quality information.

[Image of FastQC interface]
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.

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)

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 CAAGCAGAAGACGGCATAACGA

Submodule and Limit specifying file

No selection

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

Execute
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 CAAGCAGAAGACGGCATAACGA

Submodule and Limit specifying 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://biostars.org/)
  - [http://seqanswers.com/](http://seqanswers.com/)
  - [http://galaxyproject.org/search](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](image)

**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.*
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
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.
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

Thanks

Dave Clements
Galaxy Project
Johns Hopkins University
clements@galaxyproject.org
Introduction to Galaxy: RNA-seq & ChIP-seq Data Analysis

University of Cambridge
11-12 June 2015

Dave Clements
Galaxy Project
Johns Hopkins University

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30</td>
<td>RNA-Seq Analysis</td>
</tr>
<tr>
<td>10:50</td>
<td>Break</td>
</tr>
<tr>
<td>11:10</td>
<td>RNA-Seq Analysis</td>
</tr>
<tr>
<td></td>
<td>continued</td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch (on your own)</td>
</tr>
<tr>
<td>13:30</td>
<td>ChIP-Seq Analysis</td>
</tr>
<tr>
<td></td>
<td>Quality Control and Mapping</td>
</tr>
<tr>
<td>15:00</td>
<td>Break</td>
</tr>
<tr>
<td>15:20</td>
<td>ChIP-Seq Analysis</td>
</tr>
<tr>
<td></td>
<td>Differential Binding and Comparing Results</td>
</tr>
<tr>
<td>16:30</td>
<td>Done</td>
</tr>
</tbody>
</table>
RNA-Seq Analysis: Transcript Prediction

Create new history

Gear (cog) → Create New

Import:

Shared Data → Data Libraries → Training → RNA-Seq*
  → Mapped Filtered Reads
    → Select all
  → Reference
    → Select all

### RNA-Seq Analysis: Collections

<table>
<thead>
<tr>
<th>History</th>
<th>History</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cam RNA-Seq Day 2 Test</strong>&lt;br&gt;8 shown&lt;br&gt;189.3 MB</td>
<td><strong>Cam RNA-Seq Day 2 Test</strong>&lt;br&gt;8 shown&lt;br&gt;189.3 MB</td>
<td><strong>Cam RNA-Seq Day 2 Test</strong>&lt;br&gt;8 shown&lt;br&gt;189.3 MB</td>
</tr>
<tr>
<td><strong>8: genes_chr12.gtf</strong>&lt;br&gt;Operations on multiple datasets</td>
<td><strong>8: genes_chr12.gtf</strong>&lt;br&gt;Operations on multiple datasets</td>
<td><strong>Build Dataset List</strong>&lt;br&gt;Build Dataset Pair&lt;br&gt;Build List of Dataset Pairs</td>
</tr>
<tr>
<td><strong>7: chr12.fa</strong></td>
<td><strong>7: chr12.fa</strong></td>
<td><strong>7: chr12.fa</strong></td>
</tr>
<tr>
<td><strong>6: R3G_REP3 Mapped &amp; Filtered</strong></td>
<td><strong>6: R3G_REP3 Mapped &amp; Filtered</strong></td>
<td><strong>6: R3G_REP3 Mapped &amp; Filtered</strong></td>
</tr>
<tr>
<td><strong>5: R3G_REP2 Mapped &amp; Filtered</strong></td>
<td><strong>5: R3G_REP2 Mapped &amp; Filtered</strong></td>
<td><strong>5: R3G_REP2 Mapped &amp; Filtered</strong></td>
</tr>
<tr>
<td><strong>4: R3G_REP1 Mapped &amp; Filtered</strong></td>
<td><strong>4: R3G_REP1 Mapped &amp; Filtered</strong></td>
<td><strong>4: R3G_REP1 Mapped &amp; Filtered</strong></td>
</tr>
<tr>
<td><strong>3: MeOH_REP3 Mapped &amp; Filtered</strong></td>
<td><strong>3: MeOH_REP3 Mapped &amp; Filtered</strong></td>
<td><strong>3: MeOH_REP3 Mapped &amp; Filtered</strong></td>
</tr>
<tr>
<td><strong>2: MeOH_REP2 Mapped &amp; Filtered</strong></td>
<td><strong>2: MeOH_REP2 Mapped &amp; Filtered</strong></td>
<td><strong>2: MeOH_REP2 Mapped &amp; Filtered</strong></td>
</tr>
<tr>
<td><strong>1: MeOH_REP1 Mapped &amp; Filtered</strong></td>
<td><strong>1: MeOH_REP1 Mapped &amp; Filtered</strong></td>
<td><strong>1: MeOH_REP1 Mapped &amp; Filtered</strong></td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>All</strong></td>
<td><strong>All</strong></td>
</tr>
<tr>
<td><strong>None</strong></td>
<td><strong>None</strong></td>
<td><strong>None</strong></td>
</tr>
<tr>
<td><strong>For all selected...</strong></td>
<td><strong>For all selected...</strong></td>
<td><strong>For all selected...</strong></td>
</tr>
</tbody>
</table>
RNA-Seq Analysis: Collections

Cam RNA-Seq Day 2 Test
9 shown
189.3 MB

- 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

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
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**
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 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

<table>
<thead>
<tr>
<th>test_id</th>
<th>gene_id</th>
<th>gene</th>
<th>locus</th>
<th>sample_1</th>
<th>sample_2</th>
<th>status</th>
<th>value_1</th>
<th>value_2</th>
<th>log2(fold_change)</th>
<th>test_stat</th>
<th>p_value</th>
<th>q_value</th>
<th>significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2M</td>
<td>A2M</td>
<td>A2M</td>
<td>chr12:9217772-9268558</td>
<td>MeOH</td>
<td>R3G</td>
<td>NOTEST</td>
<td>3.32147</td>
<td>3.13694</td>
<td>-0.0824644</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>A2M-AS1</td>
<td>A2M-AS1</td>
<td>A2M-AS1</td>
<td>chr12:9217772-9268558</td>
<td>MeOH</td>
<td>R3G</td>
<td>NOTEST</td>
<td>7.45797</td>
<td>13.9413</td>
<td>0.902515</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>A2ML1</td>
<td>A2ML1</td>
<td>A2ML1</td>
<td>chr12:8975149-9029381</td>
<td>MeOH</td>
<td>R3G</td>
<td>NOTEST</td>
<td>4.83055</td>
<td>7.79884</td>
<td>0.691072</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>A2MP1</td>
<td>A2MP1</td>
<td>A2M-P1</td>
<td>chr12:9381128-9386803</td>
<td>MeOH</td>
<td>R3G</td>
<td>NOTEST</td>
<td>2.49656</td>
<td>0</td>
<td>-inf</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>AAAS</td>
<td>AAAS</td>
<td>AAAS</td>
<td>chr12:53701239-53715412</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>269.035</td>
<td>159.23</td>
<td>-0.756683</td>
<td>0</td>
<td>0.0005</td>
<td>0.00194017</td>
<td>yes</td>
</tr>
<tr>
<td>AACS</td>
<td>AACS</td>
<td>AACS</td>
<td>chr12:125549924-125627871</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>29.2933</td>
<td>35.0339</td>
<td>0.258178</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ABCB9</td>
<td>ABCB9</td>
<td>ABCB9</td>
<td>chr12:123405497-123451056</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>4.68869</td>
<td>1.7732</td>
<td>-1.40283</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ABCC9</td>
<td>ABCC9</td>
<td>ABCC9</td>
<td>chr12:21950323-22086286</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>553.247</td>
<td>487.261</td>
<td>-0.18323</td>
<td>1</td>
<td>0.0004</td>
<td>0.00162143</td>
<td>yes</td>
</tr>
<tr>
<td>ABCD2</td>
<td>ABCD2</td>
<td>ABCD2</td>
<td>chr12:39945021-40013843</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>86.1377</td>
<td>172.795</td>
<td>1.00435</td>
<td>1</td>
<td>0.0004</td>
<td>0.00246739</td>
<td>yes</td>
</tr>
<tr>
<td>ACACB</td>
<td>ACACB</td>
<td>ACACB</td>
<td>chr12:109577201-109706030</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>8.45306</td>
<td>15.5772</td>
<td>0.881885</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ACAD10</td>
<td>ACAD10</td>
<td>ACAD10</td>
<td>chr12:112123856-112194911</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>21.8237</td>
<td>27.8326</td>
<td>0.350882</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ACADS</td>
<td>ACADS</td>
<td>ACADS</td>
<td>chr12:121163570-121177811</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>38.644</td>
<td>16.1739</td>
<td>-1.25658</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ACARB</td>
<td>ACARB</td>
<td>ACARB</td>
<td>chr12:6747241-6756580</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>2.96987</td>
<td>3.26939</td>
<td>0.138621</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ACSM4</td>
<td>ACSM4</td>
<td>ACSM4</td>
<td>chr12:7456927-7480969</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ACTR6</td>
<td>ACTR6</td>
<td>ACTR6</td>
<td>chr12:100593864-100618202</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>475.594</td>
<td>421.324</td>
<td>-0.174799</td>
<td>1</td>
<td>0.1588</td>
<td>0.258406</td>
<td>no</td>
</tr>
<tr>
<td>ACVRI1B</td>
<td>ACVRI1B</td>
<td>ACVRI1B</td>
<td>chr12:52345450-52390860</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>32.5737</td>
<td>38.3075</td>
<td>0.233922</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ACVR1L1</td>
<td>ACVR1L1</td>
<td>ACVR1L1</td>
<td>chr12:52301201-52317145</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>1.7773</td>
<td>2.1616</td>
<td>0.759201</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ADAMA1A</td>
<td>ADAMA1A</td>
<td>ADAMA1A</td>
<td>chr12:112236666-112339706</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>30.0162</td>
<td>55.2154</td>
<td>0.879331</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ADAMT20</td>
<td>ADAMT20</td>
<td>ADAMT20</td>
<td>chr12:43748011-43945724</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>0.453322</td>
<td>0.502067</td>
<td>0.147346</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ADCY6</td>
<td>ADCY6</td>
<td>ADCY6</td>
<td>chr12:4915997-49198280</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>9.32272</td>
<td>17.6743</td>
<td>0.922135</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ADIP2R2</td>
<td>ADIP2R2</td>
<td>ADIP2R2</td>
<td>chr12:1800246-1897845</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>207.468</td>
<td>179.333</td>
<td>-0.210248</td>
<td>1</td>
<td>0.09</td>
<td>0.158988</td>
<td>no</td>
</tr>
<tr>
<td>AEBP2</td>
<td>AEBP2</td>
<td>AEBP2</td>
<td>chr12:19592607-19675173</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>143.039</td>
<td>128.293</td>
<td>-0.156957</td>
<td>1</td>
<td>0.2254</td>
<td>0.344537</td>
<td>no</td>
</tr>
<tr>
<td>AGAP2</td>
<td>AGAP2</td>
<td>AGAP2</td>
<td>chr12:8118075-8135944</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>98.2385</td>
<td>116.302</td>
<td>0.243511</td>
<td>1</td>
<td>0.935119</td>
<td>0.11475</td>
<td>no</td>
</tr>
<tr>
<td>AICDA</td>
<td>AICDA</td>
<td>AICDA</td>
<td>chr12:8754761-8765442</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>78.1514</td>
<td>63.4313</td>
<td>-0.301077</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>AKAP3</td>
<td>AKAP3</td>
<td>AKAP3</td>
<td>chr12:4724675-4754343</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>6.12385</td>
<td>7.89266</td>
<td>0.366731</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ALDH1L2</td>
<td>ALDH1L2</td>
<td>ALDH1L2</td>
<td>chr12:105413561-105478341</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>7.11374</td>
<td>8.11722</td>
<td>0.193077</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ALDH2</td>
<td>ALDH2</td>
<td>ALDH2</td>
<td>chr12:112204690-112247789</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>12.8033</td>
<td>8.05635</td>
<td>-0.668321</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ALG10</td>
<td>ALG10</td>
<td>ALG10</td>
<td>chr12:34175215-34181236</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>54.8575</td>
<td>59.3459</td>
<td>0.113461</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ALG10B</td>
<td>ALG10B</td>
<td>ALG10B</td>
<td>chr12:38710556-38723528</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>43.8157</td>
<td>63.0457</td>
<td>0.524952</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>ALK1H2</td>
<td>ALK1H2</td>
<td>ALK1H2</td>
<td>chr12:109525992-109531293</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>679.517</td>
<td>297.183</td>
<td>-1.19316</td>
<td>1</td>
<td>0.000246739</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>ALK1X</td>
<td>ALK1X</td>
<td>ALK1X</td>
<td>chr12:86574035-86595561</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
</tbody>
</table>
Cuffdiff: differentially expressed genes

<table>
<thead>
<tr>
<th>Column</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>test_stat</td>
<td>value of the test statistic used to compute significance of the observed change in FPKM</td>
</tr>
<tr>
<td>p_value</td>
<td>Uncorrected P value for test statistic</td>
</tr>
<tr>
<td>q_value</td>
<td>FDR-adjusted p-value for the test statistic</td>
</tr>
<tr>
<td>status</td>
<td>Was there enough data to run the test?</td>
</tr>
<tr>
<td>significant</td>
<td>and, was the gene differentially expressed?</td>
</tr>
</tbody>
</table>
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

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

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

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

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
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

Call me Dr. Impatient, but has anyone an answer for this?
Thanks again.
- show quoted text -

Tao Liu

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 the 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)

Your Feedback: We need it

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

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

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

Thanks

Dave Clements
Galaxy Project
Johns Hopkins University
clements@galaxyproject.org