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 2

9:00   Welcome
9:15   RNA-Seq Example Part 1
10:30  Break
10:45  RNA-Seq Example Part 2
12:00  Lunch (on your own)
13:00  SNP & Variant Calling Part 1
15:30  Break
15:45  SNP & Variant Calling Part 2
16:15  Galaxy on the Cloud
17:00  Done

Agenda: Day 2

9:00  Welcome

9:15  RNA-Seq Example Part 1

10:30  Break

10:45  RNA-Seq Example Part 2

12:00  Lunch (on your own)

13:00  SNP & Variant Calling Part 1

15:30  Break

15:45  SNP & Variant Calling Part 2

16:15  Galaxy on the Cloud

17:00  Done

RNA-Seq Analysis: Get the Data

Create new history

Gear (cog) → Create New

Import:

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

→ Raw Reads

→ Select all

→ Reference

→ Select all

NGS Quality Control Revisted

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

• 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
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

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**

Repeat this process with MeOH Rep1 R2 (the reverse reads) ...

... and now we notice a problem in Overrepresented sequences:

![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.
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

- Use Own Junctions → Yes
- Use Gene Annotation → Yes
- Gene Model Annotation → genes_chr12.gtf
- Use Raw Junctions → Yes (tab delimited file)
- Only look for supplied junctions → Yes
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.
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 $\rightarrow$ Filter

Mapping Quality $\rightarrow$ $\geq 20$

Insert Filter $\rightarrow$ isProperPair: Yes

Reference: chr12
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
<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>Welcome</td>
</tr>
<tr>
<td>9:15</td>
<td>RNA-Seq Example Part 1</td>
</tr>
<tr>
<td>10:30</td>
<td>Break</td>
</tr>
<tr>
<td>10:45</td>
<td>RNA-Seq Example Part 2</td>
</tr>
<tr>
<td>12:00</td>
<td>Lunch (on your own)</td>
</tr>
<tr>
<td>13:00</td>
<td>SNP &amp; Variant Calling Part 1</td>
</tr>
<tr>
<td>15:30</td>
<td>Break</td>
</tr>
<tr>
<td>15:45</td>
<td>SNP &amp; Variant Calling Part 2</td>
</tr>
<tr>
<td>16:15</td>
<td>Galaxy on the Cloud</td>
</tr>
<tr>
<td>17:00</td>
<td>Done</td>
</tr>
</tbody>
</table>

Agenda: Day 2

9:00  Welcome
9:15  RNA-Seq Example Part 1
10:30 Break
10:45 RNA-Seq Example Part 2
12:00 Lunch (on your own)
13:00 SNP & Variant Calling Part 1
15:30 Break
15:45 SNP & Variant Calling Part 2
16:15 Galaxy on the Cloud
17:00 Done

RNA-Seq Analysis: Transcript Prediction

Import:

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

RNA-Seq Analysis: Transcript Prediction

Run StringTie

- NGS RNA Analysis → StringTie
- Use GFF to guide? → No
- Use defaults
RNA-Seq Analysis: Transcript Prediction

Run Cuffmerge

NGS RNA Analysis → Cuffmerge
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

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.
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 Cuffmerge on MeOH & R3G Cufflinks transcripts
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
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?
RNA-Seq Differential Expression with Cuffdiff: Resources

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

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 2

9:00   Welcome
9:15   RNA-Seq Example Part 1
10:30  Break
10:45  RNA-Seq Example Part 2
12:00  Lunch (on your own)
13:00  SNP & Variant Calling Part 1
15:30  Break
15:45  SNP & Variant Calling Part 2
16:15  Galaxy on the Cloud
17:00  Done

Agenda: Day 2

9:00 Welcome
9:15 RNA-Seq Example Part 1
10:30 Break
10:45 RNA-Seq Example Part 2
12:00 Lunch (on your own)
13:00 SNP & Variant Calling Part 1
15:30 Break
15:45 SNP & Variant Calling Part 2
16:15 Galaxy on the Cloud
17:00 Done

Agenda: Day 2

9:00 Welcome
9:15 RNA-Seq Example Part 1
10:30 Break
10:45 RNA-Seq Example Part 2
12:00 Lunch (on your own)
13:00 SNP & Variant Calling Part 1
15:30 Break
15:45 SNP & Variant Calling Part 2
16:15 Galaxy on the Cloud
17:00 Done

Galaxy is available ...
AWS in Education Grants Program

http://aws.amazon.com/education
What is our path?

Today we will:

• **Launch our own Galaxy server on AWS**
• Make the server *dynamically scalable in response to demand.*
• Run some **basic analysis** on it.
• Make it **go away.**
Full Disclosure

To use AWS you must create an AWS account with a credit card associated with it.

You must also have created a key pair.

We will use the IAM account for this workshop.
IAM Accounts

Imagine, a link to a list of accounts, and credentials, here.
Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.
Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See Amazon's pricing for more information.

**Key ID**

This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Console.

**Secret Key**

This is your AWS Secret Key, also found in the Security Credentials section of the AWS Console.
Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See Amazon's pricing for more information.

Key ID
This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Console.

Secret Key
This is your AWS Secret Key, also found in the Security Credentials section of the AWS Console.

Instances in your account
- New Cluster

Cluster Name
This is the name for your cluster. You'll use this when you want to restart.

Cluster Password

Cluster Password - Confirmation

Key Pair
- cloudman_keypair

Instance Type
- Compute optimized Large (2 vCPU/4GB RAM)

Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page.
Launch a Galaxy Cloud Instance

Access Information
Your instance 'i-61503e9b' has been successfully launched using the 'ami-a7dbf6ce' AMI. While it may take a few moments to boot, you will be able to access the cloud control panel at ec2-54-196-164-110.compute-1.amazonaws.com/cloud. SSH access is also available using your private key. From the terminal, you would execute something like:
```
  `ssh -i cloudman_key_pair.pem ubuntu@ec2-54-196-164-110.compute-1.amazonaws.com`
```
CloudLaunch

Authentication Required

The server http://ec2-54-211-123-238.compute-1.amazonaws.com:80 requires a username and password. The server says: CM Administration.

User Name: 
Password: *********

Cancel  Log In
CloudLaunch

Welcome to CloudMan. This application will allow you to manage this cluster platform and the services provided within. To get started, choose the type of platform you'd like to work with and provide the associated value, if any.

- **Galaxy Cluster**: Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage type:
  - Volume - Default (10 GB)
  - Volume - Custom: [ ] GB
  - Transient Storage

Show more startup options

Choose platform type
Welcome to CloudMan. This application allows you to manage this cloud cluster and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

**Messages**

Initializing 'Galaxy' cluster type. Please wait... (2014-01-15 06:48:34)

**Status**

- **Cluster name:** PAG_CLOUD_2
- **Disk status:** 0 / 0 (0%)
- **Worker status:** Idle: 0 Available: 0 Requested: 0
- **Service status:** Applications • Data

**Cluster status log**
Cloud Launched

Messages

Initializing 'Galaxy' cluster type. Please wait... (2014-01-15 06:48:34)
All cluster services started; the cluster is ready for use. (2014-01-15 06:53:24)

is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

Cluster name: PAG_CLOUD_2
Disk status: 3.2G / 10G (32%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications ∙ Data

Cluster status log

Autoscaling is off. Turn on?
Cool things to do

- Create a **login**
- Become an **admin**
- Set up **autoscaling**
- Run ~ **Galaxy 101**
- [http://usegalaxy.org/galaxy101](http://usegalaxy.org/galaxy101)
- **Shut it down**
Agenda: Day 2

9:00   Welcome
9:15   RNA-Seq Example Part 1
10:30  Break
10:45  RNA-Seq Example Part 2
12:00  Lunch (on your own)
13:00  SNP & Variant Calling Part 1
15:30  Break
15:45  SNP & Variant Calling Part 2
16:15  Galaxy on the Cloud
17:00  Done (Almost)

Feedback: We *Need* It!

Galaxy is hiring post-docs and software engineers

Please help.
http://wiki.galaxyproject.org/GalaxyIsHiring
Also thanks to

Mani Mudalilar

Graham Hamilton

Amy Cattanach

Agenda: Day 2

9:00   Welcome
9:15   RNA-Seq Example Part 1
10:30  Break
10:45  RNA-Seq Example Part 2
12:00  Lunch (on your own)
13:00  SNP & Variant Calling Part 1
15:30  Break
15:45  SNP & Variant Calling Part 2
16:15  Galaxy on the Cloud
17:00  Done


Thanks

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

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 $\rightarrow$ Concatenate

Concatenate **Pou5f1 Rep 1 and 2 Peak files**

Operate on Genomic Intervals $\rightarrow$ 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