Introduction to Galaxy

Northwestern University
February 27, 2015

Dave Clements
Galaxy Project
Johns Hopkins University

Matt Schipma
NGS Core Facility
Center for Genetic Medicine (CGM)
Northwestern University
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>Welcome</td>
</tr>
<tr>
<td>9:30</td>
<td>Basic Analysis with Galaxy</td>
</tr>
<tr>
<td>10:45</td>
<td>Break</td>
</tr>
<tr>
<td>11:00</td>
<td>Basic Analysis into Reusable Workflows</td>
</tr>
<tr>
<td>11:30</td>
<td>RNA-Seq Example Part I</td>
</tr>
<tr>
<td>12:20</td>
<td>Lunch (on your own)</td>
</tr>
<tr>
<td>1:35</td>
<td>RNA-Seq Example Part II</td>
</tr>
<tr>
<td>2:45</td>
<td>Break</td>
</tr>
<tr>
<td>3:00</td>
<td>Sharing, Publishing, and Reproducibility</td>
</tr>
<tr>
<td>3:25</td>
<td>Setting up your own Galaxy Cluster on Amazon</td>
</tr>
<tr>
<td>4:00</td>
<td>Done</td>
</tr>
</tbody>
</table>
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 \textit{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 \textbf{RNA-Seq}. \textit{However, our emphasis today is on learning general NGS and Galaxy principles.}

It will still be enough to get you started with RNA-Seq.
Want more RNA-Seq?

Transcriptome Analysis with RNA-seq -- Mar. 19, 2015
Speaker: Matthew Schipma, NUCATS Next-Generation Sequencing Core
Direct registration link

Next generation sequencing has made it possible to sequence the entire transcriptome of a cell or tissue type. This allows researchers to evaluate the expression of each gene in the genome, including specific isoforms. RNA-seq also has the potential to identify novel splice variants and even novel genes. The ability to gain useful information from an RNA-seq project depends on several experimental factors, including the number of replicates and the sequencing depth. Here, we will discuss the best practices in designing an RNA-seq experiment as well as the information that can be obtained through RNA-seq.

Part of the
Computational Skills for Informatics Seminar Series

http://www.galter.northwestern.edu/News/computational-skills-for-informatics-seminar-series
What is Galaxy?

Data integration and analysis platform that emphasizes accessibility, reproducibility, and transparency

A free (for everyone) web server

Open source software

These options result in several ways to use Galaxy

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

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)
Galaxy Project: Further reading & Resources

http://galaxyproject.org
http://usegalaxy.org
http://getgalaxy.org
http://wiki.galaxyproject.org/Cloud
Agenda

9:00  Welcome

9:30  **Basic Analysis with Galaxy**

10:45  Break

11:00  Basic Analysis into Reusable Workflows

11:30  RNA-Seq Example Part I

12:20  Lunch (on your own)

1:35  RNA-Seq Example Part II

2:45  Break

3:00  Sharing, Publishing, and Reproducibility

3:25  Setting up your own Galaxy Cluster on Amazon

4:00  Done
Basic Analysis

Which exons have most overlapping Repeats?

Use Human, HG19, Chromosome 22

cloud1.galaxyproject.org
cloud2.galaxyproject.org
cloud3.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

Join on exon name

Join, Subtract, and Group → Join
(Incorporate the overlap count with rest of Exon information)
<table>
<thead>
<tr>
<th>Exon overlap counts</th>
<th>Exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Join on exon name

Rearrange columns with cut

Text Manipulation $\rightarrow$ Cut

(Incorporate the overlap count with rest of Exon information)
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.
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.
### 101: Overlapping Exons and Repeats

3.5 MB

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Exons, chr22</td>
<td></td>
</tr>
<tr>
<td>2: Repeats, chr22</td>
<td></td>
</tr>
<tr>
<td>3: Join on data 2 and data 1</td>
<td></td>
</tr>
<tr>
<td>4: Group on data 3</td>
<td></td>
</tr>
<tr>
<td>5: Join two Datasets on data 1 and data 4</td>
<td></td>
</tr>
<tr>
<td>6: Exons with overlapping repeats</td>
<td></td>
</tr>
</tbody>
</table>

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

9:00  Welcome
9:30  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35  RNA-Seq Example Part II
2:45  Break
3:00  Sharing, Publishing, and Reproducibility
3:25  Setting up your own Galaxy Cluster on Amazon
4:00  Done
Solution from Stanford Kwenda and Caron Griffiths, Pretoria. Takes advantage of the fact that Exons already have 0 scores.
Agenda

9:00  Welcome

9:30  Basic Analysis with Galaxy

10:45 Break

11:00 Basic Analysis into Reusable Workflows

11:30  RNA-Seq Example Part I

12:20 Lunch (on your own)

1:35  RNA-Seq Example Part II

2:45  Break

3:00  Sharing, Publishing, and Reproducibility

3:25  Setting up your own Galaxy Cluster on Amazon

4:00  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

9:00  Welcome
9:30  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35  RNA-Seq Example Part II
2:45  Break
3:00  Sharing, Publishing, and Reproducibility
3:25  Setting up your own Galaxy Cluster on Amazon
4:00  Done
RNA-Seq Analysis: Get the Data

Create new history

Gear (cog) → Create New

Import:

Shared Data → Data Libraries

→ RNA-Seq UCDavis 2013 Example Data*

→ Unfiltered Reads

→ MeOH_REP1_R1.fastq and MeOH_REP1_R2.fastq

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

```plaintext
@SEQ_ID
GATTTGGGGTTCAGTGATCGATCAATAGTAAATCCATTGTCAACTCACTAGTTT
+
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

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)
```

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

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.

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: 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
  - One preserves pairings, two don't
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://galaxyproject.org/search](http://galaxyproject.org/search)

**TopHat Overview**

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie(2), and then analyzes the mapping results to identify splice junctions between exons. Please cite: Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36, 2013.
“Mixing paired- and single-end reads together is not supported.”

Tophat Manual

“If you are performing RNA-seq analysis, there is no need to filter the data to ensure exact pairs before running Tophat.”

Jen Jackson
Galaxy User Support Person Extraordinaire

“Dang.”

Most of us

Running Tophat on no-longer-cleanly-paired data does map the reads, but, it no longer keeps track of read pairs in the SAM/BAM file.
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) that removes any unpaired reads before mapping:
- Run the Picard **Paired Read Mate Fixer** after mapping reads.
- Use sliding windows for QC, **but keep empty reads. (This does not work with Tophat.)**
I'll use Option3, sliding windows, and run a workflow afterward to patch up pairings.

- NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**

Run again:

- NGS QC and Manipulation → **FastQC** on trimmed dataset
NGS Data Quality: Base Quality Trimming

New Problem?
Now some reads are so short they are just noise and can’t be meaningfully mapped. Have potential to bog down mapping.

Option 2 can fix this, but breaks pairings (if you still have them). Or, your mapper may have an option to ignore shorter reads.
RNA-Seq Analysis

I'll use option 2, since my pairings are already broken.

NGS QC and Manipulation
→ Filter FASTQ reads by quality score and length

Pick a minimum length. I used 32.
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 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: Done with 1st Replicate!

Now, only 5 more to go!

Workflows?

Create a QC workflow that does the trimming

Or, cheat and import trimmed+paired datasets from the RNA-Seq UCDavis 2013 Example Data → Reads, Post-QC, Re-Paired

shared data library
NGS Data Quality: Further reading & Resources

**FastQC Documentation**

**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.*
Mapping with Tophat
RNA-Seq: Mapping with Tophat

Create new history

Create New

Get filtered reads

Shared Data → Data Libraries

→ RNA-Seq UCDavis 2013 Example Data*

→ Reads, Post-QC

→ Select MeOH_REP1_R1, MeOH_REP1_R2

Also select genes_chr12.gtf

And then Import to current history

RNA-seq Exercise: Mapping with Tophat

- 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
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
Mapping with Tophat: Make it quicker?

Warning: Here be dragons!

- **Allow indel search** → No
- **Use Coverage Search** → No (wee dragons)

TopHat generates its database of possible splice junctions from two sources of evidence. The first and strongest source of evidence for a splice junction is when two segments from the same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found *ab initio*. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. *We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (= 10 million).* This latter option will only report alignments across "GT-AG" introns.

TopHat Manual
Mapping with Tophat: Max # of Alignments Allowed

Some reads align to more than one place equally well.
For such reads, how many should Tophat include?
If more than the specified number, Tophat will pick those with the best mapping score.
Tophat breaks ties randomly.
Tophat assigns equal fractional credit to all $n$ mappings

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

9:00 Welcome
9:30 Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 RNA-Seq Example Part II
2:45 Break
3:00 Sharing, Publishing, and Reproducibility
3:25 Setting up your own Galaxy Cluster on Amazon
4:00 Done
Galaxy Community Resources
Galaxy Community Resources: Galaxy Biostar

Tens of thousands of users leads to a lot of questions. Absolutely have to encourage community support.

Project traditionally used mailing list

Moved the user support list to Galaxy Biostar, an online forum, that uses the Biostar platform.

Want help? Get answers.

https://biostar.usegalaxy.org/
Galaxy Community Resources: Mailing Lists
http://wiki.galaxyproject.org/MailingLists

Galaxy-Dev
Questions about developing for and deploying Galaxy
High volume (5200 posts in 2013, 900+ members)
(3246 posts in 2014, 1000+ members)

Galaxy-Announce
Project announcements, low volume, moderated
Low volume (47 posts in 2013, 3400+ members)
(34 posts in 2014, 4400+ members)

Galaxy-User (discontinued 2014/05)
Questions about using Galaxy and usegalaxy.org
High volume (1328 posts in 2013, 2600+ members)
(358 posts in 2014, 2600+ members)
Unified Search: http://galaxyproject.org/search

**Find**
- Everything on …
- Tools for …
- Email about …
- Source code for …
- Published Histories, Pages, Workflows, about …
- Related feature requests
- Papers using Galaxy for …
- Documentation on …
Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

- **Accessible**: Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible**: Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent**: Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

This is the Galaxy Community Wiki. It describes all things Galaxy.

**Use Galaxy**

Galaxy's public web server usegalaxy.org makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive user documentation applicable to any public or local Galaxy instance is available.

**Community & Project**

Galaxy has a large and active user community and many ways to get involved.
- **Community**
- **Users**: Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.

**Deploy Galaxy**

Galaxy is a free and open source project available to all. Local Galaxy servers can be set up by downloading the Galaxy application.
- **Admin**
- **Cloud**

**Contribute**

- **Users**: Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.
**Events**

*Galaxy Event Horizon*

Events with Galaxy-related content are listed here.

Also see the Galaxy Events Google Calendar for a listing of events and deadlines that are not on the Galaxy Community. This is also available as an RSS feed.

If you know of any event that should be added to this page and/or to the Galaxy Event Calendar, please send it to outreach@galaxyproject.org.

For events prior to this year, see the Events Archive.

**Upcoming Events**

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic/Event</th>
<th>Venue/Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>December 12</td>
<td><strong>Introduction to Galaxy Workshop</strong></td>
<td>Virginia State University, Petersburg, Virginia</td>
</tr>
<tr>
<td>December 16-19</td>
<td><strong>RNA-Seq and ChIP-Seq Analysis with Galaxy</strong></td>
<td>UC Davis, California, United States</td>
</tr>
<tr>
<td>January 10-14</td>
<td><strong>Galaxy for SNP and Variant Data Analysis</strong></td>
<td>Plant and Animal Genome XXIII (PAG2014), States</td>
</tr>
<tr>
<td>January 19-20</td>
<td><strong>NGS pipelines with Galaxy</strong></td>
<td>e-Infrastructures for Massively Parallel Sequencing, Sweden</td>
</tr>
<tr>
<td>February 9-13</td>
<td><strong>Analyse bioinfomatic de séquences sous Galaxy</strong></td>
<td>Montpellier, France</td>
</tr>
<tr>
<td>February 16-18</td>
<td><strong>Accessible and Reproducible Large-Scale Analysis with Galaxy</strong></td>
<td>Genome and Transcriptome Analysis, Genome, Conference, San Francisco, California</td>
</tr>
<tr>
<td>February 16-18</td>
<td><strong>Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic</strong></td>
<td>Genomics &amp; Sequencing Data Integration, of Molecular Medicine Tri-Conference, San Francisco, California</td>
</tr>
</tbody>
</table>

**News**

*News Items*

**Opening at McMaster University**

The McArthur Lab in the McMaster University Department of Biochemistry & Biomedical Sciences is seeking a Systems Administrator / Information Technologist to help establish a new bioinformatics laboratory at McMaster, plus develop the next generation of the Comprehensive Antibiotic Resistance Database (CARD).

From the job announcement on EvoDir:

The candidate will configure BLADE and other hardware for general bioinformatics analysis, development of a GIT version control system, construction of an in house Galaxy server (usegalaxy.org), and development of a new interface, stand-alone tools, APIs, and algorithms for the CARD (based on Chado).

See the full announcement for details.

*December 2014 Galaxy Newsletter*

As always there’s a lot going on in the Galaxy this month. “Like what?” you say. Well, read the dang December Galaxy Newsletter we say! Highlights include:

- Galaxy Day! In Paris! This Wednesday!
- Near Richmond, Virginia? There’s a Galaxy Workshop at Virginia State U on December 12.
- GCC2015 needs sponsors!
- Other upcoming events on two continents
- **96 new papers**, including 6 highlighted papers, referencing, using, extending, and implementing Galaxy.
- Job openings at 7+ organizations
- A new mailing list: Galaxy-Training
- 15 new ToolShed repositories from 10 contributors
- And, 10 other juicy (well maybe not juicy, but certainly not crunchy) bits of news

Dave Clements and the crimp Galaxy Team

**Bioinformaticians, Freiburg**

Max Planck Institute of Immunobiology and Epigenetics in Freiburg, Germany has an opening for a Bioinformatician for an initial period of two years. The successful candidate will work at the interface between in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the institute. Main responsibilities include...
Cistrome
A Galaxy Server dedicated to ChIP-* analysis

deepTools

CoSSci
Galaxy for Complex Social Sciences

GWIS: Online exhaustive bivariate GWAS in minutes...

CloudBased Image Analysis & Processing Toolbox

The Microbiome Analysis Center
Community can create, vote and comment on issues

Slides, posters & videos now online
GCC 2015
Galaxy Community Conference
6-8th July 2015
The Sainsbury Laboratory
Norwich, UK

gcc2015.tsl.ac.uk
We also support community organized efforts and events.
Galaxy Resources & Community: Videos

“How to” screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Journal/Conference</th>
<th>DOI</th>
<th>Group Tags</th>
</tr>
</thead>
</table>

Over 2100 papers

Galaxy Training Network launched in October.

bit.ly/gxygtn
Agenda

9:00  Welcome
9:30  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35  RNA-Seq Example Part II
2:45  Break
3:00  Sharing, Publishing, and Reproducibility
3:25  Setting up your own Galaxy Cluster on Amazon
4:00  Done
# Tuxedo Suite: Some parts of the ensemble

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bowtie</strong></td>
<td>Short read mapper. Bowtie2 can do gapped alignments and emphasizes reads &gt; 50 bases</td>
</tr>
<tr>
<td><strong>Tophat</strong></td>
<td>Intron-aware mapper for RNA-Seq data. Works with Bowtie to find best mapping locations</td>
</tr>
<tr>
<td><strong>Cufflinks</strong></td>
<td>Construct transcript predictions from mapped reads (from Tophat output)</td>
</tr>
<tr>
<td><strong>Cuffmerge</strong></td>
<td>Merges multiple sets of transcript predictions into a unified set with one coherent set of IDs.</td>
</tr>
<tr>
<td><strong>Cuffdiff</strong></td>
<td>Differential expression analysis; Can work with Tophat output directly or Cufflinks/merge, if looking for novel genes/transcripts</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Used already</th>
<th>Will not use today</th>
<th>Will use next</th>
</tr>
</thead>
</table>


RNA-Seq: Differential Expression with Cuffdiff
RNA-Seq Differential Expression: Get the Data

Create new history

остоя (cog) → Create New

Import:

Shared Data → Data Libraries

→ RNA-Seq UCDavis 2013 Example Data*

→ Tophat Outputs

→ Select all accepted_hits datasets

Also select genes_chr12.gtf

And then Import to current history

Cuffdiff

• Part of the Tuxedo RNA-Seq Suite (as are Tophat and Bowtie)
• 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

- Running with 2 Groups: MeOH and R3G
- Each group has 3 replicates each
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

**NGS: RNA Analysis ➔ Cuffdiff**
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>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>A2ML1</td>
<td>A2ML1</td>
<td>A2ML1</td>
<td>chr8:8975149-9029381</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>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>A2MP1</td>
<td>A2MP1</td>
<td>A2MP1</td>
<td>chr9:9381128-9386803</td>
<td>MeOH</td>
<td>R3G</td>
<td>NOTEST</td>
<td>2.49656</td>
<td>0</td>
<td>-inf</td>
<td>0</td>
<td>1</td>
<td>1</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>-2.22857</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>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ABC9</td>
<td>ABC9</td>
<td>ABC9</td>
<td>chr12:12340549-123451056</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>4.68869</td>
<td>1.7732</td>
<td>-1.40283</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ABC9</td>
<td>ABC9</td>
<td>ABC9</td>
<td>chr12:21290323-22086286</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>553.247</td>
<td>487.261</td>
<td>-0.18323</td>
<td>-2.02806</td>
<td>0.0004</td>
<td>0.00162143</td>
<td>yes</td>
</tr>
<tr>
<td>AB2D</td>
<td>AB2D</td>
<td>AB2D</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>4.3436</td>
<td>5e-05</td>
<td>0.000246739</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>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ACAD10</td>
<td>ACAD10</td>
<td>ACAD10</td>
<td>chr12:11213856-112194911</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>1</td>
<td>1</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>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ACRBP</td>
<td>ACRBP</td>
<td>ACRBP</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>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ACS4</td>
<td>ACS4</td>
<td>ACS4</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>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ACS5</td>
<td>ACS5</td>
<td>ACS5</td>
<td>chr12:81471808-81649582</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</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>-0.797581</td>
<td>0.1588</td>
<td>0.258406</td>
<td>no</td>
</tr>
<tr>
<td>ACR1B</td>
<td>ACR1B</td>
<td>ACR1B</td>
<td>chr12:52345450-52390863</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>32.5737</td>
<td>38.3075</td>
<td>0.233922</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ACRVL1</td>
<td>ACRVL1</td>
<td>ACRVL1</td>
<td>chr12:52301201-52317145</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>1.77137</td>
<td>2.16161</td>
<td>0.759201</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ADAM1A</td>
<td>ADAM1A</td>
<td>ADAM1A</td>
<td>chr12:112336866-112339706</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>30.0162</td>
<td>55.2154</td>
<td>0.879332</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ADAMTS20</td>
<td>ADAMTS20</td>
<td>ADAMTS20</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>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ADCY6</td>
<td>ADCY6</td>
<td>ADCY6</td>
<td>chr12:49159974-49198280</td>
<td>MeOH</td>
<td>R3G</td>
<td>NOTEST</td>
<td>9.32772</td>
<td>17.6743</td>
<td>0.922135</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ADIPOR2</td>
<td>ADIPOR2</td>
<td>ADIPOR2</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>-1.02392</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>-0.688267</td>
<td>0.2254</td>
<td>0.344537</td>
<td>no</td>
</tr>
<tr>
<td>AGAP1</td>
<td>AGAP1</td>
<td>AGAP1</td>
<td>chr12:58118075-58135944</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>98.2385</td>
<td>116.302</td>
<td>0.243511</td>
<td>0.935119</td>
<td>0.11475</td>
<td>0.198086</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>0</td>
<td>1</td>
<td>1</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>7.89626</td>
<td>0.366731</td>
<td>0</td>
<td>1</td>
<td>1</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>8.11772</td>
<td>0.190377</td>
<td>0</td>
<td>1</td>
<td>1</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>0</td>
<td>1</td>
<td>1</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>0.11346</td>
<td>0</td>
<td>1</td>
<td>1</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>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>ALKBH2</td>
<td>ALKBH2</td>
<td>ALKBH2</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>-3.34255</td>
<td>5e-05</td>
<td>0.000246739</td>
<td>yes</td>
</tr>
<tr>
<td>ALX1</td>
<td>ALX1</td>
<td>ALX1</td>
<td>chr12:85674035-85695561</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</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?
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

9:00  Welcome
9:30  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 RNA-Seq Example Part II
2:45 Break
3:00 Sharing, Publishing, and Reproducibility
3:25 Setting up your own Galaxy Cluster on Amazon
4:00 Done
Agenda

9:00  Welcome
9:30  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35  RNA-Seq Example Part II
2:45  Break
3:00  Sharing, Publishing, and Reproducibility
3:25  Setting up your own Galaxy Cluster on Amazon
4:00  Done
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
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.
Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond1,2,6,9, Samir Wadhawan3,6,7,
Francesca Chiaromonte4, Guruprasad Ananda1,3, Wen-Yu Chung1,3,8,
James Taylor1,5,9, Anton Nekrutenko1,3,9 and The Galaxy Team1

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.]
Windshield splatter analysis with the Galaxy metagenonomic pipeline: A live supplement

SEVERI KOSAKOVSKY POND1,2,*, SAMIR WADHAWAN3,6,*, FRANCESCA CHIAROMONTE4, GURUPRASAD ANANDA1,3, WEN-YU CHUNG1,3,7, JAMES TAYLOR1,5, ANTON NEKRUTENKO1,3 and THE GALAXY TEAM1*

Correspondence should be addressed to SKP. IT. 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):
```

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From there they can be re-analyzed through Galaxy using the above workflows and downloaded.

http://usegalaxy.org/u/aun1/p/windshield-splatter
Basic Analysis: Further reading & Resources

http://usegalaxy.org/galaxy101

https://vimeo.com/76343659
Agenda

9:00 Welcome
9:30 Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 RNA-Seq Example Part II
2:45 Break
3:00 Sharing, Publishing, and Reproducibility
3:25 Setting up your own Galaxy Cluster on Amazon
4:00 Done
Galaxy is available ... 

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

We are using the cloud today.
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.
CloudLaunch: From UseGalaxy.org

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.

Submit
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/cloud 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.

### Status

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

![Autoscaling is off. Turn on?](image)
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
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

9:00  Welcome
9:30  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 RNA-Seq Example Part II
2:45 Break
3:00 Sharing, Publishing, and Reproducibility
3:25 Setting up your own Galaxy Cluster on Amazon
4:00  Done (almost)
We Need Your Feedback

bit.ly/N Ugxy201502
Galaxy is **hiring post-docs and software engineers**

Please help.

Also Thanks To

Pamela Shaw
Kristi Holmes

National Institutes of Health

bit.ly/N Ugxy201502
Agenda

9:00  Welcome
9:30  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35  RNA-Seq Example Part II
2:45  Break
3:00  Sharing, Publishing, and Reproducibility
3:25  Setting up your own Galaxy Cluster on Amazon
4:00  Done

bit.ly/N Ugxy201502
Thanks

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

Matt Schipma
NGS Core Facility
Center for Genetic Medicine (CGM)
Northwestern University
m-schipma@northwestern.edu

bit.ly/N Ugxy201502
Cuffmerge

- Each Cufflinks run creates a set of transcript predictions.
- **Cuffmerge** unifies all those predictions into a single set.
- Makes this incredibly tedious task easy.