Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20  RNA-Seq Analysis, Part I
2:50  Break
3:05  RNA-Seq Analysis, Part II
4:30  Launch your own Galaxy with AWS
5:00  Done

bit.ly/btigxy
Goals

Provide an 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.

This workshop does cover RNA-Seq but you won't be an expert at the end of the workshop. You will know enough to get started, and how to use Galaxy to learn more.
What is Galaxy?

Keith Bradnam's definition:

"A web-based platform that provides a simplified interface to many popular bioinformatics tools."

From

"13 Questions You May Have About Galaxy"

Galaxy is available several ways ...
As a free for everyone service on the web: usegalaxy.org
Galaxy is available as Open Source Software

Galaxy is installed in locations around the world.

http://getgalaxy.org
Explore the Galaxy with RNA-Rocket

The Microbiome Analysis Center
Life on a Smaller Scale

Welcome to the Metabiome Portal @ GNU

Integrated publishing of workflows from GIGA in Science

(GIGA)n Galaxy by CBIIT

Cistrome
A Galaxy Server dedicated to ChIP-* analysis

080 +
Public Galaxy Servers and still counting

The Genomic HyperBrowser

Powered by Galaxy

SCDE
STEM CELL DISCOVERY ENGINE

Experiments Connected

Whale Shark Galaxy!

South Green bioinformatics platform

Genomic analysis tools for southern and Mediterranean plants

bit.ly/gxyServers
Galaxy is available on the Cloud

We are using this today

http://aws.amazon.com/education
http://globus.org/
http://wiki.galaxyproject.org/Cloud
Galaxy on the Cloud: Galaxy CloudMan
http://usegalaxy.org/cloud

• Start with a **fully configured and populated** (tools and data) Galaxy instance.
• Allows you to scale up and down your compute assets as needed.
• Someone else manages the data center
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20 RNA-Seq Analysis, Part I
2:50 Break
3:05 RNA-Seq Analysis, Part II
4:30 Launch your own Galaxy with AWS
5:00 Done

bit.ly/btigxy
Quick Poll: Are you ...

1. A bioinformatics novice
2. A bioinformatics apprentice
3. A bioinformatics guru

Yes, those are your only choices.

http://galaxyproject.org
Basic Analysis
Which exons have most overlapping Repeats?
Use Human, HG38, GENCODE v24, Chromosome 22

cloud1.galaxyproject.org
cloud2.galaxyproject.org
cloud5.galaxyproject.org
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 )
Use this program to retrieve the data associated with a track in text format, to calculate intersections of DNA sequence covered by a track. For help in using this application see Using the Table Browser in this form, the User's Guide for general information and sample queries, and the OpenHelix Table, a presentation of the software features and usage. For more complex queries, you may want to use this tool.

To examine the biological function of your set through annotation enrichments, send the data to Galaxy for use with diverse computational tools. Refer to the Credits page for the list of contributors and for help using these data. All tables can be downloaded in their entirety from the Sequence and Annotation Download page.


region: genome  position: chr22  filter: create

output format: BED - browser extensible data  Send output to Galaxy  output file: (leave blank to keep output in browser)  file type returned: plain text

get output summary/statistics
Output knownGene as BED

Include custom track header:
- name = tb_knownGene
- description = table browser query on knownGene
- visibility = pack
- url =

Create one BED record per:
- Whole Gene
- Upstream by 200 bases
- Exons plus 0 bases at each end
- Introns plus 0 bases at each end
- 5' UTR Exons
- Coding Exons
- 3' UTR Exons
- Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream in order to avoid extending past the edge of the chromosome.

Send query to Galaxy
Cancel
(Identify which exons have Repeats)
Operate on Genomic Intervals \(\rightarrow\) Join

(Identify which exons have Repeats)
Published History: Exons with overlapping repeats, basic

Join, Subtract, and Group → Group

(Count Repeats per exon)

Exon overlap counts
Yay!

We have exon names and counts!

We are now going to extend that work.

Let's create a copy of this history that we will extend.
Exons & Repeats: Pick an Exercise

1. Report the number of overlapping repeats every exon has (including exons with 0 overlapping repeats.)

2. Output the list of exons that have overlapping repeats, in BED format. Set the score column be the number of overlapping repeats that exon has.

Everything you need will be in these toolboxes
• Text manipulation (cut is particularly useful)
• Operate on genomic intervals
• Join, subtract and group
• Filter and sort
But first, create a login

Don't need to login to use Galaxy, but do need one to use all its features

Use an email address you can remember.

Use a low security password.

This account will go away on Wednesday night.
Second, name your existing history

Give your existing history a meaningful name.
3rd, make a copy of your history

(cog) → Copy History
Name the copy based on the exercise you pick

Becomes your new current history.
All exons, even those with no overlap can take advantage of fact that scores are already 0. Join, subtract and group not a bad place to start.

Published History: Exons with number of overlapping repeats, including 0
List of exons with overlaps, in BED

Can be done in two steps, one of them a Cut, plus an edit attributes step at the end:

Published History: Exons with overlapping repeats, in BED
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20  RNA-Seq Analysis, Part I
2:50  Break
3:05  RNA-Seq Analysis, Part II
4:30  Launch your own Galaxy with AWS
5:00  Done
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20  RNA-Seq Analysis, Part I
2:50  Break
3:05  RNA-Seq Analysis, Part II
4:30  Launch your own Galaxy with AWS
5:00  Done

bit.ly/btigxy
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
- And then rolling that up to genes

But, ...

is there anything inherent in the analysis about humans, exons or repeats?
Get back to the original history
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
Create a Workflow from a History: ...

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

### Workflow name

Workflow constructed from history 'Exons with overlapping repeats, basic'

[Create Workflow]  [Check all]  [Uncheck all]

<table>
<thead>
<tr>
<th>Tool</th>
<th>History items created</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCSC Main</td>
<td>1: Exons, chr22</td>
</tr>
<tr>
<td></td>
<td>- Treat as input dataset</td>
</tr>
<tr>
<td>UCSC Main</td>
<td>2: Repeats, chr22</td>
</tr>
<tr>
<td></td>
<td>- Treat as input dataset</td>
</tr>
<tr>
<td>Join</td>
<td>3: Join on data 2 and data 1</td>
</tr>
<tr>
<td></td>
<td>- Include &quot;Join&quot; in workflow</td>
</tr>
<tr>
<td>Group</td>
<td>4: Exons with overlapping repeats</td>
</tr>
<tr>
<td></td>
<td>- Include &quot;Group&quot; in workflow</td>
</tr>
</tbody>
</table>
Workflow editor

Published Workflow: Feature Overlap Counting
Workflow editor: save your changes

Published Workflow: Feature Overlap Counting
Guided: rerun with same inputs
Workflow $\rightarrow$ Run
Did that work?

On your own:
Count # of exons overlapping each repeat
Did that work? Why not?
Edit workflow: doc assumptions

Published Workflow: Feature Overlap Counting
Workflows: Sweet spots

Short, well-defined tasks, with well-defined inputs and outputs.

Analysis pipelines for large experiments with many samples where sample and data preparation protocols are the same throughout.
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20 RNA-Seq Analysis, Part I
2:50 Break
3:05 RNA-Seq Analysis, Part II
4:30 Launch your own Galaxy with AWS
5:00 Done

bit.ly/btigxy
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20 RNA-Seq Analysis, Part I
2:50 Break
3:05 RNA-Seq Analysis, Part II
4:30 Launch your own Galaxy with AWS
5:00 Done

bit.ly/btigxy
Quick Poll: Are you ...

1. An RNA-Seq **novice**
2. An RNA-Seq **apprentice**
3. An RNA-Seq **guru**

Yes, those are your only choices.

http://galaxyproject.org
RNA-Seq Analysis: Get the Data

Shared Data → Data Libraries → Training → RNA-Seq*
→ UC-Davis → Raw Reads
  Select first two
  MeOH_REP1_R1
  MeOH_REP1_R2
  Import into a new history

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
GATTTGGGGTTCAAGCAGTATCGATCAAAATAGTAAATCCATTTGTTCAACTCAGTTT
+
'!'*(((****))%+%+(%%%).1***-+''))**55CCF>))))))))CCCCCCCCCC65
```

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

```plaintext
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)
```

NGS Data Quality: Assessment tools

NGS QC and Manipulation → **FastQC**
Generates summary quality information.

http://bit.ly/FastQCBoxPlot
NGS Data Quality: Assessment tools

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

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.

Common Trimming options

- **Drop the first n columns** from your reads
- **Drop the last n columns** from your reads
- **Sliding window** approach: only keep regions that are above a specified quality threshold
- **Keep or drop whole read** based on overall quality
Common Trimming Pitfalls

**Broken Pairs**

Often, one side of a pair passes QC, while the other does not. Broken pairings can affect results in subtle or drastic ways.

**Short short reads.**

QC may reduce reads to a length at which their mapping is no longer meaningful.
Need help with Trimming? (and anything else)

That's a whole lotta options...

Choices you make now have impact on downstream tools

NGS = a whole lotta options in general

What to do?
How to better understand bioinformatics & Galaxy

- **Experiment.** (You are already used to the idea and) Galaxy makes it easy

- **Read** tool documentation and tool and method review papers

- **Get Help!**
  - http://biostars.org/
  - http://seqanswers.com/
  - https://biostar.usegalaxy.org/
  - http://galaxyproject.org/search
Trimmomatic to the rescue

Trimmomatic preserves read pairing

Multiple filters can be run in arbitrary order

We'll use **sliding window**, followed by **minimum length**.
Run FastQC on post-Trimmmatic Datasets

NGS QC and Manipulation → FastQC

Now, let's see what changed

Shared History: RNA-Seq MeOH_REPI QC
Scratchbook: View multiple datasets

And the icon turns **yellow**!

Poke the **pre**-Trimmomatic reverse read FastQC report in the eye, and then poke the **post**-Trimmomatic FastQC report in the eye.
And after some resizing and scrolling you see this
NGS Data Quality Assessment

Now, just 10 more datasets to go!
Your Friend: The Multiple datasets button

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.32.3)

Paired end data?
- Yes
- No

Input Type
- Pair of datasets

Input FASTQ file (R1/first of pair)
- 1: MeOH_REP1_R1.fastq

Input FASTQ file (R2/second of pair)
- Multiple datasets
- 2: MeOH_REP1_R2.fastq

Perform initial ILLUMINACLIP step?
- Yes
- No

Cut adapter and other illumina–specific sequences from the read

Trimmomatic Operation
1: Trimmomatic Operation
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20 RNA-Seq Analysis, Part I
2:50 Break
3:05 RNA-Seq Analysis, Part II
4:30 Launch your own Galaxy with AWS
5:00 Done

bit.ly/btigxy
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20 RNA-Seq Analysis, Part I
2:50 Break
3:05 RNA-Seq Analysis, Part II
4:30 Launch your own Galaxy with AWS
5:00 Done

bit.ly/btigxy
RNA-seq Exercise: **Differential gene expression**

Take samples under multiple conditions
(MeOH and R3G exposure in our example)

Map them
Count them
Compare them
RNA-Seq Mapping: Get the Data

Import into a new history:

Shared Data → Data Libraries → Training → RNA-Seq
→ UC-Davis* → Post QC reads → Still paired reads
Select first two
  MeOH_REP1_R1 post QC
  MeOH_REP1_R2 post QC

Shared Data → Data Libraries → Training → RNA-Seq
→ UC-Davis → Reference
Select GTF for hg38, chr12 from Sanger

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 →
  
  GTF for hg38, chr12 from Sanger

You can also restrict Tophat to known annotations

- 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
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.
Mapping With Tophat: What to keep?

NGS BAM Tools → Filter

This shows two options for cleanup.

Use only the good stuff!
Only 5 more replicates to go!

Another way to avoid insanity is *Collections*
RNA-Seq Differential Expression: \textcolor{red}{Get the Data}

Import into a new history:

\textcolor{orange}{Shared Data} \rightarrow \textcolor{orange}{Data Libraries} \rightarrow \textcolor{orange}{Training} \rightarrow \textcolor{orange}{RNA-Seq*}  
\rightarrow \textcolor{orange}{UC-Davis} \rightarrow \textcolor{orange}{Mapped}  
Select all (OK, maybe just half of them)

\textcolor{orange}{Shared Data} \rightarrow \textcolor{orange}{Data Libraries} \rightarrow \textcolor{orange}{Training} \rightarrow \textcolor{orange}{RNA-Seq*}  
\rightarrow \textcolor{orange}{UC-Davis} \rightarrow \textcolor{orange}{Reference}  
Select \textcolor{orange}{GTF for hg38, chr12 from Sanger}

Dataset collections!

**Dataset Collections** give Galaxy semantic knowledge about dataset relationships.

Tools can then take advantage of this knowledge.
Dataset collections
Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names

<table>
<thead>
<tr>
<th>0 unpaired forward</th>
<th>0 unpaired reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6 filtered out)</td>
<td>(6 filtered out)</td>
</tr>
</tbody>
</table>

Choose filters

Choose from the following filters to change which unpaired reads are shown in the display:

- Forward: _1, Reverse: _2
- Forward: _R1, Reverse: _R2

---

Create a collection of paired datasets

<table>
<thead>
<tr>
<th>3 unpaired forward</th>
<th>3 unpaired reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3 filtered out)</td>
<td>(3 filtered out)</td>
</tr>
</tbody>
</table>

Choose filters

Choose from the following filters to change which unpaired reads are shown in the display:

- _R1
- _R2

MeOH_REP1_R1
MeOH_REP2_R1
MeOH_REP3_R1

Pair these datasets

MeOH_REP1_R2
MeOH_REP2_R2
MeOH_REP3_R2

Pair these datasets
Create a collection of paired datasets

3 pairs created: all datasets have been successfully paired

0 unpaired forward – (0 filtered out)

0 unpaired reverse – (0 filtered out)

Choose filters Clear filters

_R1

_R2

3 paired Unpair all

<table>
<thead>
<tr>
<th>MeOH_REP1_R1</th>
<th>MeOH_REP1</th>
<th>MeOH_REP1_R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeOH_REP2_R1</td>
<td>MeOH_REP2</td>
<td>MeOH_REP2_R2</td>
</tr>
<tr>
<td>MeOH_REP3_R1</td>
<td>MeOH_REP3</td>
<td>MeOH_REP3_R2</td>
</tr>
</tbody>
</table>

Remove file extensions from pair names? □

Name: MeOH

Create list
Dataset collections
Dataset collections Created
Differential expression with 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.

Now supports geometric normalization, the same model used by DESeq (and in fact, it's now the default). Less prone to distortion from highly expressed genes.
Cuffdiff: Which transcript definitions to use?

We'll use the official genome annotations.

But there are a world of options out there for discovering and using novel transcripts.

StringTie, Cufflinks, Cuffmerge, ...
Cuffdiff

- Running with 2 Groups: MeOH and R3G
- Each group has 3 replicates each
- Can take advantage of collections
### Transcripts
- 13: GTF for hg38, chr12 from Sanger

### Omit Tabular Datasets
- Yes
- No

**Discard the tabular output.**

### Generate SQLite
- Yes
- No

**Generate a SQLite database for use with cummeRbund.**

### Input data type
- SAM/BAM

**CuffNorm supports either CXB (from cuffquant) or SAM/BAM input files. Mixing is not supported. Default: SAM/BAM**

### Condition
1: Condition
- Name: MeOH
- Replicates: 62: HISAT2 on R3G

2: Condition
- Name: R3G
- Replicates: 58: HISAT2 on MeOH
Cuffdiff

Execute it
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>
</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 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>0.90251</td>
<td>0</td>
<td>1</td>
<td>1 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>0</td>
<td>1</td>
<td>1 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 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 yes</td>
</tr>
<tr>
<td>AACS</td>
<td>AACS</td>
<td>AACS</td>
<td>chr12:12554992-125627871</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>29.2933</td>
<td>35.0339</td>
<td>0.258178</td>
<td>0</td>
<td>1</td>
<td>1 no</td>
</tr>
<tr>
<td>ABCB9</td>
<td>ABCB9</td>
<td>ABCB9</td>
<td>chr12:123404597-123451056</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 no</td>
</tr>
<tr>
<td>ABCCC9</td>
<td>ABCCC9</td>
<td>ABCCC9</td>
<td>chr12:29150323-2208626</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>553.247</td>
<td>487.261</td>
<td>-0.1323</td>
<td>-2.02806</td>
<td>0.0004</td>
<td>0.00162143 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>4.3436</td>
<td>0.000246739 yes</td>
<td></td>
</tr>
<tr>
<td>ACAB</td>
<td>ACAB</td>
<td>ACAB</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>0</td>
<td>1</td>
<td>1 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>0</td>
<td>1</td>
<td>1 no</td>
</tr>
<tr>
<td>ACAD5</td>
<td>ACAD5</td>
<td>ACAD5</td>
<td>chr12:112163350-12177811</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 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>0</td>
<td>1</td>
<td>1 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>0</td>
<td>1</td>
<td>1 no</td>
</tr>
<tr>
<td>ACS53</td>
<td>ACS53</td>
<td>ACS53</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 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>-0.79581</td>
<td>0.1588</td>
<td>0.258406 no</td>
</tr>
<tr>
<td>ACVR1B</td>
<td>ACVR1B</td>
<td>ACVR1B</td>
<td>chr12:52354540-52398063</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 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>1.77731</td>
<td>2.16161</td>
<td>0.759201</td>
<td>0</td>
<td>1</td>
<td>1 no</td>
</tr>
<tr>
<td>ADAM1A</td>
<td>ADAM1A</td>
<td>ADAM1A</td>
<td>chr12:11233686-112339706</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 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 no</td>
</tr>
<tr>
<td>ADCY6</td>
<td>ADCY6</td>
<td>ADCY6</td>
<td>chr12:49153974-49198280</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>9.32722</td>
<td>17.6743</td>
<td>0.922135</td>
<td>0</td>
<td>1</td>
<td>1 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 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 no</td>
</tr>
<tr>
<td>AGAP2</td>
<td>AGAP2</td>
<td>AGAP2</td>
<td>chr12:58118075-58135944</td>
<td>MeOH</td>
<td>R3G</td>
<td>OK</td>
<td>98.2358</td>
<td>116.302</td>
<td>0.243511</td>
<td>0.935119</td>
<td>0.11475</td>
<td>0.198086 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 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 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>7.13174</td>
<td>8.11722</td>
<td>0.190377</td>
<td>0</td>
<td>1</td>
<td>1 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 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 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 no</td>
</tr>
<tr>
<td>ALKH2</td>
<td>ALKH2</td>
<td>ALKH2</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>0.000246739 yes</td>
<td></td>
</tr>
<tr>
<td>ALX1</td>
<td>ALX1</td>
<td>ALX1</td>
<td>chr12:85674035-85685561</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 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</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

9:00  Welcome

9:20 Basic Analysis with Galaxy

10:45 Break

11:00 Basic Analysis into Reusable Workflows

12:20 Lunch (on your own)

1:20 RNA-Seq Analysis, Part I

2:50 Break

3:05 RNA-Seq Analysis, Part II

4:30 Launch your own Galaxy with AWS

5:00 Done

bit.ly/btigxy
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20  RNA-Seq Analysis, Part I
2:50  Break
3:05  RNA-Seq Analysis, Part II
4:30  Launch your own Galaxy with AWS
5:00  Done

bit.ly/btigxy
Acknowledgements

You
Surya Saha
Lukas Mueller
John Ashton

AWS
NIH
Johns Hopkins University
Penn State University

Boyce Thompson Institute
Cornell University

bit.ly/btgxy_feedback
Thanks
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:45 Break
11:00 Basic Analysis into Reusable Workflows
12:20 Lunch (on your own)
1:20  RNA-Seq Analysis, Part I
2:50  Break
3:05  RNA-Seq Analysis, Part II
4:30  Launch your own Galaxy with AWS
5:00  Done

bit.ly/btigxy_feedback
bit.ly/btigxy_feedback
2016 Galaxy Community Conference (GCC2016)

June 25-29, 2016
Bloomington, Indiana

galaxyproject.org/GCC2016

Slides & posters are now online. Video will be shortly
26 - 30 June  France

GCC 2017 Montpellier

Social Dinner at Château du Pouget

Le Corum Conference centre
November 7-11

Salt Lake City, Utah
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/
Scaling Training

Galaxy Training Network: Trainer Locations
The Galaxy Training Network (https://wiki.galaxyproject.org/Teach/GTN)

Made with Google My Maps

Galaxy Training Network
bit.ly/gxygtn
Proteomics
Metabolomics
Natural Language
Image Analysis
Climate Change
Social Science
Cosmology
Galaxy Community Resources: Mailing Lists
http://wiki.galaxyproject.org/MailingLists

**Galaxy-Dev**
Questions about developing for and deploying Galaxy
High volume (2336 posts in 2015, 1000+ members)

**Galaxy-Announce**
Project announcements, low volume, moderated
Low volume (36 posts in 2015, 6500+ members)

Also **Galaxy-UK, -France, -Proteomics, -Training,** ...
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](http://usegalaxy.org) makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive [user documentation](http://usegalaxy.org) applicable to any public or local Galaxy instance is available.

**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

**Community & Project**

Galaxy has a large and active user community and many ways to get involved.

- Community

**Contribute**

- **Users**: Share your histories, workflows, visualizations, data libraries, and [Galaxy Pages](http://galaxyproject.org/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 filtered for 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>Introduction to Galaxy Workshop</td>
<td>Virginia State University, Petersburg, Virginia</td>
</tr>
<tr>
<td>December 16-19</td>
<td>RNA-Seq and ChIP-Seq Analysis with Galaxy</td>
<td>UC Davis, California, United States</td>
</tr>
</tbody>
</table>

2015

| January 10-14 | Galaxy for SNP and Variant Data Analysis | Plant and Animal Genome XXIII (PAG2014) States |
| January 19-20 | NGS pipelines with Galaxy | e-Infrastructures for Massively Parallel Sequencing, Sweden |
| February 9-13 | Analyse bioinformatique de séquences sous Galaxy | Montpellier, France |
| February 16-18 | Accessible and Reproducible Large-Scale Analysis with Galaxy | Genome and Transcriptome Analysis, 1st Galaxy Conference, San Francisco, California |
|               | Large-Scale NGS data Analysis on Amazon Web Services Using Globus Genomic | Genomics & Sequencing Data Integration, of Molecular Medicine Tri-Conference, Santa Barbara, California |

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 EviDir:

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 crisp 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 an in-house deep-sequencing facility (HiSeq-2500) and the various research groups at the institute. Main responsibilities include primary analysis of deep sequencing data and quality control.
Galaxy Resources & Community: Videos

“How to” screencasts on using and deploying Galaxy

Talks from previous meetings.

http://vimeo.com/galaxyproject
Galaxy Resources & Community: CiteULike Group

Now almost 3000 papers

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
A free for everyone web service:

http://usegalaxy.org

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.*
What is Galaxy?

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

http://galaxyproject.org
Yay! We have a list of genes and overlap counts!*

Now, what can we do with that?

All sorts of things.

* Technically, we have a list of gene symbols, and the maximum number of overlapping repeats from any of its transcripts. We also haven't done things like normalize the scores based on gene length. Your mileage may vary. Let's not sweat the details.
GO Term Enrichment

Do genes with particular functions tend to occur in this list more often than they would by random chance?
GO: Create a list of just the gene symbols

Remember how?
(Stop or) GO: Can do this step, or just watch

http://geneontology.org/
### GO: Results from whole genome, 1 or more overlapping repeats (8969 genes)

Displaying only results with P<0.05; [click here to display all results](#)

<table>
<thead>
<tr>
<th>GO biological process complete</th>
<th>Homo sapiens (REF)</th>
<th>upload_1 ( ▼ Hierarchy NEW! )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>chromatin modification</td>
<td>289</td>
<td>196</td>
</tr>
<tr>
<td>chromatin organization</td>
<td>636</td>
<td>376</td>
</tr>
<tr>
<td>chromosome organization</td>
<td>984</td>
<td>555</td>
</tr>
<tr>
<td>organelle organization</td>
<td>3133</td>
<td>1636</td>
</tr>
<tr>
<td>cellular component organization</td>
<td>5133</td>
<td>2606</td>
</tr>
<tr>
<td>cellular process</td>
<td>14559</td>
<td>6671</td>
</tr>
<tr>
<td>cellular component organization or biogenesis</td>
<td>5288</td>
<td>2688</td>
</tr>
<tr>
<td>macromolecular complex subunit organization</td>
<td>1983</td>
<td>1021</td>
</tr>
<tr>
<td>peptidyl-lysine modification</td>
<td>314</td>
<td>194</td>
</tr>
<tr>
<td>peptidyl-amino acid modification</td>
<td>855</td>
<td>456</td>
</tr>
<tr>
<td>cellular protein modification process</td>
<td>2836</td>
<td>1397</td>
</tr>
<tr>
<td>protein modification process</td>
<td>2836</td>
<td>1397</td>
</tr>
<tr>
<td>protein metabolic process</td>
<td>4036</td>
<td>1908</td>
</tr>
<tr>
<td>macromolecule metabolic process</td>
<td>7359</td>
<td>3685</td>
</tr>
<tr>
<td>organic substance metabolic process</td>
<td>9032</td>
<td>4308</td>
</tr>
<tr>
<td>metabolic process</td>
<td>9443</td>
<td>4480</td>
</tr>
<tr>
<td>primary metabolic process</td>
<td>8601</td>
<td>4133</td>
</tr>
<tr>
<td>macromolecule modification</td>
<td>3007</td>
<td>1480</td>
</tr>
</tbody>
</table>

Published History: Gene-Repeat overlap, entire genome
### GO: Results from whole genome, 2 or more overlapping repeats (2759 genes)

<table>
<thead>
<tr>
<th>GO biological process complete</th>
<th>Homo sapiens (REF)</th>
<th>upload_1 (* Hierarchy NEW! (?)</th>
<th>#</th>
<th>#</th>
<th>expected</th>
<th>Fold Enrichment</th>
<th>+/-</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>membrane depolarization during action potential</td>
<td>39</td>
<td>19</td>
<td></td>
<td></td>
<td>5.22</td>
<td>3.64</td>
<td>+</td>
<td>2.01E-02</td>
</tr>
<tr>
<td>biological regulation</td>
<td>11384</td>
<td>1776</td>
<td></td>
<td></td>
<td>1523.15</td>
<td>1.17</td>
<td>+</td>
<td>2.20E-18</td>
</tr>
<tr>
<td>membrane depolarization</td>
<td>61</td>
<td>24</td>
<td></td>
<td></td>
<td>8.16</td>
<td>2.94</td>
<td>+</td>
<td>3.99E-02</td>
</tr>
<tr>
<td>regulation of histone methylation</td>
<td>59</td>
<td>25</td>
<td></td>
<td></td>
<td>7.89</td>
<td>3.17</td>
<td>+</td>
<td>7.12E-03</td>
</tr>
<tr>
<td>regulation of histone modification</td>
<td>129</td>
<td>43</td>
<td></td>
<td></td>
<td>17.26</td>
<td>2.49</td>
<td>+</td>
<td>9.63E-04</td>
</tr>
<tr>
<td>regulation of primary metabolic process</td>
<td>5720</td>
<td>1046</td>
<td></td>
<td></td>
<td>765.32</td>
<td>1.37</td>
<td>+</td>
<td>5.02E-27</td>
</tr>
<tr>
<td>regulation of metabolic process</td>
<td>6087</td>
<td>1096</td>
<td></td>
<td></td>
<td>814.43</td>
<td>1.35</td>
<td>+</td>
<td>2.49E-26</td>
</tr>
<tr>
<td>regulation of biological process</td>
<td>10767</td>
<td>1708</td>
<td></td>
<td></td>
<td>1440.60</td>
<td>1.19</td>
<td>+</td>
<td>1.66E-20</td>
</tr>
<tr>
<td>regulation of macromolecule metabolic process</td>
<td>5730</td>
<td>1052</td>
<td></td>
<td></td>
<td>766.66</td>
<td>1.37</td>
<td>+</td>
<td>6.06E-28</td>
</tr>
<tr>
<td>regulation of cellular metabolic process</td>
<td>5781</td>
<td>1058</td>
<td></td>
<td></td>
<td>773.48</td>
<td>1.37</td>
<td>+</td>
<td>1.21E-27</td>
</tr>
<tr>
<td>regulation of cellular process</td>
<td>10292</td>
<td>1651</td>
<td></td>
<td></td>
<td>1377.04</td>
<td>1.20</td>
<td>+</td>
<td>1.75E-21</td>
</tr>
<tr>
<td>regulation of chromatin organization</td>
<td>152</td>
<td>50</td>
<td></td>
<td></td>
<td>20.34</td>
<td>2.46</td>
<td>+</td>
<td>1.43E-04</td>
</tr>
<tr>
<td>regulation of chromosome organization</td>
<td>272</td>
<td>66</td>
<td></td>
<td></td>
<td>36.39</td>
<td>1.81</td>
<td>+</td>
<td>4.52E-02</td>
</tr>
<tr>
<td>regulation of organelle organization</td>
<td>1097</td>
<td>211</td>
<td></td>
<td></td>
<td>146.78</td>
<td>1.44</td>
<td>+</td>
<td>1.37E-03</td>
</tr>
<tr>
<td>regulation of cellular component organization</td>
<td>2246</td>
<td>409</td>
<td></td>
<td></td>
<td>300.51</td>
<td>1.36</td>
<td>+</td>
<td>1.22E-06</td>
</tr>
<tr>
<td>histone lysine methylation</td>
<td>64</td>
<td>27</td>
<td></td>
<td></td>
<td>8.56</td>
<td>3.15</td>
<td>+</td>
<td>2.90E-03</td>
</tr>
<tr>
<td>histone methylation</td>
<td>84</td>
<td>31</td>
<td></td>
<td></td>
<td>11.24</td>
<td>2.76</td>
<td>+</td>
<td>6.88E-03</td>
</tr>
<tr>
<td>histone modification</td>
<td>337</td>
<td>88</td>
<td></td>
<td></td>
<td>45.09</td>
<td>1.95</td>
<td>+</td>
<td>5.91E-05</td>
</tr>
<tr>
<td>covalent chromatin modification</td>
<td>346</td>
<td>92</td>
<td></td>
<td></td>
<td>46.29</td>
<td>1.99</td>
<td>+</td>
<td>1.16E-05</td>
</tr>
<tr>
<td>macromolecule metabolic process</td>
<td>7359</td>
<td>1260</td>
<td></td>
<td></td>
<td>984.62</td>
<td>1.28</td>
<td>+</td>
<td>4.66E-23</td>
</tr>
<tr>
<td>organic substance metabolic process</td>
<td>9032</td>
<td>1384</td>
<td></td>
<td></td>
<td>1208.46</td>
<td>1.15</td>
<td>+</td>
<td>1.23E-07</td>
</tr>
</tbody>
</table>
GO: Results from **whole genome, 3 or more overlapping repeats (986 genes)**

<table>
<thead>
<tr>
<th>GO biological process complete</th>
<th>Homo sapiens (REF)</th>
<th>upload 1 (Hierarchy, NEW)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td># expected</td>
</tr>
<tr>
<td>histone H3–K4 methylation</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td>histone lysine methylation</td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>histone methylation</td>
<td>84</td>
<td>19</td>
</tr>
<tr>
<td>histone modification</td>
<td>337</td>
<td>43</td>
</tr>
<tr>
<td>covalent chromatin modification</td>
<td>346</td>
<td>44</td>
</tr>
<tr>
<td>macromolecule metabolic process</td>
<td>7359</td>
<td>496</td>
</tr>
<tr>
<td>organic substance metabolic process</td>
<td>9032</td>
<td>524</td>
</tr>
<tr>
<td>metabolic process</td>
<td>9443</td>
<td>535</td>
</tr>
<tr>
<td>chromatin organization</td>
<td>636</td>
<td>80</td>
</tr>
<tr>
<td>chromosome organization</td>
<td>984</td>
<td>93</td>
</tr>
<tr>
<td>organelle organization</td>
<td>3133</td>
<td>214</td>
</tr>
<tr>
<td>cellular component organization</td>
<td>5133</td>
<td>319</td>
</tr>
<tr>
<td>cellular process</td>
<td>14559</td>
<td>773</td>
</tr>
<tr>
<td>cellular component organization or biogenesis</td>
<td>5288</td>
<td>329</td>
</tr>
<tr>
<td>macromolecular complex subunit organization</td>
<td>1983</td>
<td>148</td>
</tr>
<tr>
<td>primary metabolic process</td>
<td>8601</td>
<td>519</td>
</tr>
<tr>
<td>cellular macromolecule metabolic process</td>
<td>6693</td>
<td>475</td>
</tr>
<tr>
<td>cellular metabolic process</td>
<td>8525</td>
<td>505</td>
</tr>
</tbody>
</table>

Published History: Gene-Repeat overlap, entire genome
Yay! But, a wee challenge

We have exon names and counts

We really want genes (or transcripts) and counts across the whole gene (or transcript)
What we have: Computer generated Exon IDs

uc002zmb.3_cds_0_0_chr22_17119391_r

Transcript ID is embedded in Exon ID.*

How can we extract the Transcript ID from the Exon ID?

(With the transcript ID we can summarize counts for each transcript and/or get the gene ID.)

* How do we know that's a transcript ID?
Create another copy of your original history
Create another copy of your original history

Put the word Gene in the history name
Extract the transcript ID

Split the exon ID into its constituent parts.

<table>
<thead>
<tr>
<th>uc002zmb.3</th>
<th>cds</th>
<th>0</th>
<th>0</th>
<th>chr22</th>
<th>17119391</th>
<th>r</th>
<th>6</th>
</tr>
</thead>
</table>

Text Manipulation → Convert delimiters to TAB (convert underscores to tabs)
Sum the scores for all exons in each transcript

Join, Subtract and Group →
Group: by Transcript ID; Sum score
Get list of transcript IDs

Published History: Transcripts with # of overlapping repeats
Have Transcripts, now get Gene IDs

Save list of Transcript IDs to a file.

We'll upload it to Ensembl BioMart

Published History: Transcripts with # of overlapping repeats
Ensembl BioMart

www.ensembl.org/biomart/martview

Specify Ensembl Genes 84, GRCh38.p5
Ensembl BioMart:

Specify attributes to put in output report
Specify attributes to put in output report
Ensembl BioMart:

Specify which genes we want this information for.
Save the results to a file for uploading into Galaxy
Get Genes into Galaxy

Chose local file, then Start, then Close
Get Gene IDs into Galaxy

Upload file from BioMart. Note that we lost 4-5 transcripts

Do we care? Can we find out which were lost?
Unite our Transcript Scores with Biomart info

Join, Subtract and Group →
Join: Transcripts with score and Biomart dataset; join on UCSC transcript ID
Unite our Transcript Scores with Biomart info

Join two Datasets side by side on a specified field (Galaxy Version 2.0.2)

Join
- File: mart_export.txt

using column
- Column: 4

with
- File: Transcripts with # overlapping repeats

and column
- Column: 1

Keep lines of first input that do not join with second input
- No

Keep lines of first input that are incomplete
- No

Fill empty columns
- No

Execute

Join, Subtract and Group → Join
Assign scores to genes

Join, Subtract and Group →
Group: by gene symbol; Max score

Published History: Genes with overlapping repeats
Now have a list of genes with # overlapping repeats

<table>
<thead>
<tr>
<th>Gene</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC007326.1</td>
<td>4</td>
</tr>
<tr>
<td>AGR7</td>
<td>2</td>
</tr>
<tr>
<td>ADM2</td>
<td>1</td>
</tr>
<tr>
<td>ADRBK2</td>
<td>1</td>
</tr>
<tr>
<td>ADSL</td>
<td>2</td>
</tr>
<tr>
<td>ANKRD54</td>
<td>1</td>
</tr>
<tr>
<td>AP000049.2</td>
<td>1</td>
</tr>
<tr>
<td>APOBEC3B</td>
<td>2</td>
</tr>
<tr>
<td>APOBEC3F</td>
<td>1</td>
</tr>
<tr>
<td>APOBEC3H</td>
<td>1</td>
</tr>
<tr>
<td>APOL3</td>
<td>1</td>
</tr>
<tr>
<td>APOL4</td>
<td>2</td>
</tr>
<tr>
<td>APOL5</td>
<td>1</td>
</tr>
<tr>
<td>APOL6</td>
<td>1</td>
</tr>
<tr>
<td>ARFGAP3</td>
<td>1</td>
</tr>
<tr>
<td>ARHGAP8</td>
<td>2</td>
</tr>
<tr>
<td>ARSA</td>
<td>1</td>
</tr>
<tr>
<td>ASCC2</td>
<td>2</td>
</tr>
<tr>
<td>ASPHD2</td>
<td>1</td>
</tr>
<tr>
<td>ATP6V1E1</td>
<td>1</td>
</tr>
</tbody>
</table>

Published History: Genes with overlapping repeats

- 228 lines
- 623 lines