Galaxy Workflows

Workflows Workshop

August 9-10, 2016
Online

Dave Clements & Enis Afgan
Galaxy Team
Johns Hopkins University
http://galaxyproject.org/

#usegalaxy    @galaxyproject
Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turning that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides
Launching Galaxy in Jetstream

bit.ly/gxyjets

(https://wiki.galaxyproject.org/Cloud/Jetstream)
Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turning that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides
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"


http://galaxyproject.org
Galaxy was created for genomics.
Now used across a spectrum of disciplines

Metabolomics, Natural Language, Image Analysis, Climate Change, Social Science, Cosmology, ...
Galaxy is available several ways ...

bit.ly/ww_gxy_slides
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

(CIGA)<n>Galaxy by CBIIT

Integrated publishing of workflows from GIGA in SCIENCE

Cistrome

A Galaxy Server dedicated to ChIP-* analysis

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

http://aws.amazon.com/education
http://globus.org/
http://wiki.galaxyproject.org/Cloud
We'll use this now

bit.ly/ww_gxy_slides
Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turning that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides
Specific Question: **Repeats in Genes?**

Which genes physically overlap with large numbers of repeats?
Genes: Exons, Introns, and alternative versions

Figure 7-21 Essential Cell Biology 3/e (© Garland Science 2010)

http://oregonstate.edu/instruction/bi314/fall12/geneexpression.html
Genes: Exons, Introns, and alternative versions

What to remember:

**Exons** become **proteins** and **proteins** become **you**.

**Exons** can be **combined** to create **different proteins**.

**Warning:**

"Biology is a system of exceptions." Tom Conlin, 2000
Repeats: Simple and Complex

**Simple** repeats:
DNA stuttering

**Complex** repeats:
DNA that can replicate and insert copies of themselves

Half the human genome is repetitive
1/50th of the human genome is exons
Galaxy: Should be active:

Copy & paste IP address into a new browser tab.
Welcome to Galaxy on the Jetstream Cloud

Galaxy on the Jetstream Cloud is ready for use!

To learn how to use Galaxy please see the wiki. To install new tools to your Galaxy follow the tutorial.

Thank you for using Galaxy.

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and Computer Science at Johns Hopkins 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 Johns Hopkins.
Create a login on your server

Click Return to the home page.

Note: Connection is not encrypted
Repeats in Genes: A General Plan

• Get data about Genes, and about Repeats
• Identify which genes have overlapping repeats
• Count repeats per gene

bit.ly/ww_gxy_slides
Get Data: Genes and Repeats

Import 3 data files into your Galaxy instance
Get Data: Genes and Repeats

Paste these 2 URLs into the paste box:

Get Data: Genes and Repeats

Set the **Genome** to **hg38** (the most recent human)

Set the data **type** to **bed**
Get Data: Genes and Repeats

Paste one more dataset. Click **Paste/Fetch data** again, and then paste this URL into the new paste box:

Get Data: Genes and Repeats

Set the data type to **tabular** and the **Genome** to **hg38**

Then click **Start**, and then **Close**.
Get Data: Genes and Repeats

The three datasets show up in your history, first as queued, and then as done. The datasets are automatically uncompressed by Galaxy.
Get Data: Datasets

The three datasets are:

3. Transcript to Gene mapping

2. Repeats - as identified by RepeatMasker software

1. Transcripts - Gene and Exon info
Datasets: Take a peek

Preview a dataset in the history by clicking on the dataset's name.

Tells us
- How big the dataset is
- The format (BED)
- The genome (hg38)
- Where it came from
- And a short preview of the data in it.
Datasets: See the whole thing (poke it in the eye)

The 6 column RepeatMasker dataset.
Repeats in Genes: A General Plan

- Get data about Genes, and about Repeats
- Identify which genes have overlapping repeats
- Count repeats per gene
Our first tool: Extract exons from genes

Open the **Extract Features** toolbox in the tool panel and select **Gene BED to Exon/...**

**Extract Coding Exons only** from the transcript dataset. Click **Execute**.
### Extracted Exons

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chr22</td>
<td>11065973</td>
<td>11066015</td>
<td>uc062bdq.1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>chr22</td>
<td>11067334</td>
<td>11067346</td>
<td>uc062bdq.1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>chr22</td>
<td>11066500</td>
<td>11066515</td>
<td>uc062bdr.1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>chr22</td>
<td>11067984</td>
<td>11068089</td>
<td>uc062bdr.1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>chr22</td>
<td>15528158</td>
<td>15529139</td>
<td>uc011agd.3</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>chr22</td>
<td>15690077</td>
<td>15690709</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>chr22</td>
<td>15695370</td>
<td>15695485</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>chr22</td>
<td>15695644</td>
<td>15695818</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>chr22</td>
<td>15698661</td>
<td>15698768</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>chr22</td>
<td>15700077</td>
<td>15700215</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>chr22</td>
<td>15702685</td>
<td>15702756</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>chr22</td>
<td>15708019</td>
<td>15708090</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>chr22</td>
<td>15709781</td>
<td>15709826</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>chr22</td>
<td>15710867</td>
<td>15711034</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>chr22</td>
<td>15719659</td>
<td>15719777</td>
<td>uc010gqp.3</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>chr22</td>
<td>15690077</td>
<td>15690314</td>
<td>uc062bej.1</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>chr22</td>
<td>15690425</td>
<td>15690709</td>
<td>uc062bej.1</td>
<td>0</td>
</tr>
</tbody>
</table>

6 column dataset, 1 record per exon

**Transcript name** has become the **exon name**
Identify which **exons** have overlapping repeats

Open the **Operate on Genomic Intervals** toolbox and select **Join**. **Join** the **exons** dataset **with** the **repeatMasker** dataset. Click **Execute**.
Result has 911 Exon-RepeatMasker pairings

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>chr22</td>
<td>11065973</td>
<td>11066015</td>
<td>uc062bdq.1</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>11064567</td>
<td>11067436</td>
<td>REP522</td>
<td>6038</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>chr22</td>
<td>11067334</td>
<td>11067346</td>
<td>uc062bdq.1</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>11064567</td>
<td>11067436</td>
<td>REP522</td>
<td>6038</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>chr22</td>
<td>11066500</td>
<td>11066515</td>
<td>uc062bdr.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>11067982</td>
<td>11068155</td>
<td>REP522</td>
<td>348</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>chr22</td>
<td>11067984</td>
<td>11068089</td>
<td>uc062bdr.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>11067982</td>
<td>11068155</td>
<td>REP522</td>
<td>348</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>chr22</td>
<td>15697373</td>
<td>15697532</td>
<td>uc062bek.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>15697355</td>
<td>15697497</td>
<td>AluJb</td>
<td>1340</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>chr22</td>
<td>15697373</td>
<td>15697532</td>
<td>uc062bek.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>15697947</td>
<td>15697530</td>
<td>(TTTTTA)n</td>
<td>32</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>15697373</td>
<td>15697532</td>
<td>uc062bek.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>15697530</td>
<td>15697651</td>
<td>AluJb</td>
<td>1340</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17105852</td>
<td>17105954</td>
<td>uc002zyi.5</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17109651</td>
<td>17109678</td>
<td>(GGA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17108306</td>
<td>17109820</td>
<td>uc002zyi.5</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17109651</td>
<td>17109678</td>
<td>(GGA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17108306</td>
<td>17109820</td>
<td>uc062bfr.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17109651</td>
<td>17109678</td>
<td>(GGA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120596</td>
<td>17120630</td>
<td>(GCA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120596</td>
<td>17120630</td>
<td>(GCA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120320</td>
<td>17120347</td>
<td>(GTA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120739</td>
<td>17120793</td>
<td>GA-rich</td>
<td>22</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120913</td>
<td>17120929</td>
<td>(TGCTGC)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120913</td>
<td>17120929</td>
<td>(TGCTGC)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120970</td>
<td>17120993</td>
<td>(TGCTGC)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zmb.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120970</td>
<td>17120993</td>
<td>(TGCTGC)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17534747</td>
<td>17534777</td>
<td>uc062bgi.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17534992</td>
<td>17534802</td>
<td>AluY</td>
<td>2037</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17550242</td>
<td>17550333</td>
<td>uc062bgm.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17550132</td>
<td>17550323</td>
<td>AluSx1</td>
<td>1189</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17619033</td>
<td>17619127</td>
<td>uc062bgu.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17619011</td>
<td>17619306</td>
<td>AluSc8</td>
<td>2210</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17726676</td>
<td>17726690</td>
<td>uc010ggy.4</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17726672</td>
<td>17726726</td>
<td>AmnSINE1</td>
<td>184</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>chr22</td>
<td>17726676</td>
<td>17727534</td>
<td>uc002zmz.4</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17726672</td>
<td>17726726</td>
<td>AmnSINE1</td>
<td>184</td>
<td>-</td>
</tr>
</tbody>
</table>

Left 6 columns are Exon info, right 6 columns are RepeatMasker info. Each record represents an overlap pairing.
Repeats in Genes: A General Plan

- Get data about Genes, and about Repeats
- Identify which genes have overlapping repeats
- **Count repeats per gene**

bit.ly/ww_gxy_slides
Group results by Transcript name to get counts

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr22</td>
<td>11065973</td>
<td>11066015</td>
<td>uc062bdq.1</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>11064567</td>
<td>11067436</td>
<td>REP522</td>
<td>6038</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>11067334</td>
<td>11067346</td>
<td>uc062bdq.1</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>11064567</td>
<td>11067436</td>
<td>REP522</td>
<td>6038</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>11066500</td>
<td>11066515</td>
<td>uc062bd.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>11064567</td>
<td>11067436</td>
<td>REP522</td>
<td>6038</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>11067984</td>
<td>11068089</td>
<td>uc062bd.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>11067982</td>
<td>11068155</td>
<td>REP522</td>
<td>348</td>
<td>-</td>
</tr>
<tr>
<td>chr22</td>
<td>15697373</td>
<td>15697532</td>
<td>uc062bek.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>15697355</td>
<td>15697497</td>
<td>Alujb</td>
<td>1340</td>
<td>-</td>
</tr>
<tr>
<td>chr22</td>
<td>15697373</td>
<td>15697532</td>
<td>uc062bek.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>15697497</td>
<td>15697530</td>
<td>(TTTTTA)n</td>
<td>32</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>15697373</td>
<td>15697532</td>
<td>uc062bek.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>15697530</td>
<td>15697651</td>
<td>Alu jb</td>
<td>1340</td>
<td>-</td>
</tr>
<tr>
<td>chr22</td>
<td>17105852</td>
<td>17105954</td>
<td>uc002zly.5</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17105952</td>
<td>17106169</td>
<td>MIRb</td>
<td>424</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17108306</td>
<td>17109820</td>
<td>uc002zly.5</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17109651</td>
<td>17109678</td>
<td>(GGA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17108306</td>
<td>17109820</td>
<td>uc062brf.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17109651</td>
<td>17109678</td>
<td>(GGA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zym.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120596</td>
<td>17120630</td>
<td>(GCA)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zym.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120320</td>
<td>17120347</td>
<td>(GTG)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zym.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120739</td>
<td>17120793</td>
<td>GA-rich</td>
<td>22</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zym.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120913</td>
<td>17120929</td>
<td>(TGCTG)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zym.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120929</td>
<td>17120970</td>
<td>(GCC)n</td>
<td>19</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17119390</td>
<td>17121127</td>
<td>uc002zym.3</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17120970</td>
<td>17120993</td>
<td>(TGCTG)n</td>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17534743</td>
<td>17534747</td>
<td>uc062bgi.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17534499</td>
<td>17534802</td>
<td>AluY</td>
<td>2037</td>
<td>-</td>
</tr>
<tr>
<td>chr22</td>
<td>17550242</td>
<td>17550333</td>
<td>uc062bgm.1</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17550132</td>
<td>17550323</td>
<td>AluSx1</td>
<td>1189</td>
<td>-</td>
</tr>
<tr>
<td>chr22</td>
<td>17619033</td>
<td>17619127</td>
<td>uc062bgu.1</td>
<td>0</td>
<td>-</td>
<td>chr22</td>
<td>17619011</td>
<td>17619306</td>
<td>AluSc8</td>
<td>2210</td>
<td>+</td>
</tr>
<tr>
<td>chr22</td>
<td>17726676</td>
<td>17726690</td>
<td>uc010gyy.4</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17726672</td>
<td>17726726</td>
<td>AmnSINE1</td>
<td>184</td>
<td>-</td>
</tr>
<tr>
<td>chr22</td>
<td>17726676</td>
<td>17727534</td>
<td>uc002zmz.4</td>
<td>0</td>
<td>+</td>
<td>chr22</td>
<td>17726672</td>
<td>17726726</td>
<td>AmnSINE1</td>
<td>184</td>
<td>-</td>
</tr>
</tbody>
</table>

There is a record for every time a repeat overlaps an exon in a transcript. The # of records with each transcript name is the # of overlaps.
Open **Join, Subtract and Group** toolbox and select **Group**. Set **Group by column** to **Column 4** (the transcript name).
Group results by Transcript name to get counts

Click + Insert Operation and set the Type to Count. Click Execute.
For each value of column 4, keep a count of # records with that value.
Group results by Transcript name returns:

Returns 628 transcripts that have one or more overlapping repeats.

But, biologists more often think in terms of Genes, rather than transcripts.

Let's associate these counts with the transcript's genes.
3rd imported dataset is a mapping from transcript names to genes.
Associate transcript counts with genes

Open **Join, Subtract and Group** toolbox and select **Join**. Join the transcript count dataset (**Group on**) with the dataset 3 (**transcriptGeneSymbol** mapping file). Transcript name is **Column 1** in both datasets.
For each transcript, we now have a count, and gene name.

For some genes, we have different overlap counts.
For each gene, get the max # of overlaps

Why column 4 and column 2?
230 Genes on chr22 have 1+ overlapping repeats

Sort the results so genes with most overlaps are sorted first.

Break ties by gene name.
Maximum # overlaps on chr22 is 6 in 3 different genes

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CECR6</td>
<td>6</td>
</tr>
<tr>
<td>SCARF2</td>
<td>6</td>
</tr>
<tr>
<td>SHANK3</td>
<td>6</td>
</tr>
<tr>
<td>MYH9</td>
<td>5</td>
</tr>
<tr>
<td>TCF20</td>
<td>5</td>
</tr>
<tr>
<td>AC007326.1</td>
<td>4</td>
</tr>
<tr>
<td>BAIAP2L2</td>
<td>4</td>
</tr>
<tr>
<td>CACNA1I</td>
<td>4</td>
</tr>
<tr>
<td>CRELD2</td>
<td>4</td>
</tr>
<tr>
<td>DRICH1</td>
<td>4</td>
</tr>
<tr>
<td>MAPK8IP2</td>
<td>4</td>
</tr>
<tr>
<td>MED15</td>
<td>4</td>
</tr>
<tr>
<td>RIMBP3</td>
<td>4</td>
</tr>
<tr>
<td>RIMBP3C</td>
<td>4</td>
</tr>
<tr>
<td>SH3BP1</td>
<td>4</td>
</tr>
<tr>
<td>TBX1</td>
<td>4</td>
</tr>
<tr>
<td>TMEM191B</td>
<td>4</td>
</tr>
<tr>
<td>TNRC6B</td>
<td>4</td>
</tr>
<tr>
<td>CABIN1</td>
<td>3</td>
</tr>
<tr>
<td>CARD10</td>
<td>3</td>
</tr>
<tr>
<td>CCDC157</td>
<td>3</td>
</tr>
<tr>
<td>CFI5 SR1</td>
<td>3</td>
</tr>
</tbody>
</table>
Before we move on, a word about naming

Unnamed history and Sort on data 8 are both, um, accurate, but not informative.

A best practice is to name your histories, your inputs, and outputs with informative names.

To name the history, click on Unnamed history, enter the new name, and then hit the Return key.

To rename a dataset click on the dataset's pencil (edit attributes) icon. (And see the next slide.)
Rename a dataset

Edit Attributes

Name:
Genes w # overlapping repeats, chr22;

Info:
Sort on data 8

Annotation / Notes:
Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:
Human Dec. 2013 (GRCh38/hg38) (hg38)

Number of comment lines:

Save
Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

**Turn that analysis into a reusable workflow**

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

[bit.ly/ww_gxy_slides]
Now, let's **rerun that analysis**

With

- entire genome
- different species
- repeats identified by other repeat software
- ...

Reselecting all those tools and parameters manually is error prone.

It's also a path to insanity.
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
Create a Workflow from a History

**Extract Workflow from history**
Create a workflow from this history. Edit it to make some things clearer.

![Cog icon] → Extract Workflow
Create a Workflow from a History

Give it a meaningful name and click Create Workflow.
Edit the workflow

After resizing the window, and collapsing the tool panel, this is the initial layout of the workflow.
Rearrange the steps to make the data flow clearer.
Give the input and output datasets meaningful names

Name the transcripts input dataset.
Give the input and output datasets meaningful names

Name the repeats input dataset.
Give the input and output datasets meaningful names

Name the Transcript Gene mapping input dataset.
Name the output dataset.

Gene with maximum number of overlapping repeats in any of the gene's transcripts.
Save your workflow edits!

Review your workflows
Start a new history

Click the **Analyze Data** tab and then click the **View all histories** icon
Start a new history

Click the **Create New** button to create a new history
Drag the transcript and transcript/GeneSymbol datasets from your old history to your new one.
Exit the all histories view
Import TandemRepeatFinder repeats for chr22
Paste this URL into the paste box:
and set the Type to bed and the Genome to hg38
Our history now has all 3 datasets needed to run the workflow

3. SimpleRepeats for chr22
2. Transcript-Gene Symbol mapping for chr22
1. Transcripts for chr22
Run our workflow using SimpleRepeats
Run our workflow using SimpleRepeats
All tasks are queued

Successfully ran workflow "Genes-Repeats overlaps". The following datasets have been added to the queue:

4. Gene BED To Exon/Intron/Codon BED on data 1
5. Join on data 3 and data 4
6. Group on data 5
7. Join two Datasets on data 2 and data 6
8. Group on data 7
9. Genes w max # of overlapping repeats
All tasks finish. Take a look at the results.
And finally give the output dataset and the history meaningful names.
Time allowing

Sharing and publishing
Exporting
Comparing Gene Lists
Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turn that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

bit.ly/ww_gxy_slides
Select the running Galaxy instance and then click the delete (x) icon.
Agenda

Launching Galaxy in Jetstream

A quick stroll through the Galaxy

Demonstrate Galaxy by addressing a specific question

Turn that analysis into a reusable workflow

Shutting down Galaxy in Jetstream

Galaxy Ecosystem

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

June 25-29, 2016
Bloomington, Indiana
galaxyproject.org/GCC2016

Slides & posters are now online. Video will be shortly
November 7-11

2016 Galaxy Admin Training

SLC

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/
Galaxy Training Network
bit.ly/gxygtn
Acknowledgements

You

Scott Lathrop
Michael Miller
Mike Pingleton
XSEDE
Jetstream

NIH
Johns Hopkins University
Penn State University