Galaxy Community Conference 2014: Visualization Workshop

Jeremy Goecks
Assistant Prof. of Comp. Biology

Sam Guerler
Research Software Engineer

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

JOHNS HOPKINS UNIVERSITY
Topics

Visualization history and introduction
Biological Visualizations
Numerical Visualizations
Adding your own visualizations
Why Visualize?
Why Visualize?

Quick check: did it work?

Exploration and hypothesis generation

Sharing/publishing
Anscombe’s Quartet

http://en.wikipedia.org/wiki/Anscombe’s_quartet
Timeline of Visualization in Galaxy

- 2005: 1st Galaxy paper published
- 2008: Display applications
- 2010: Visualization development started
- 2011: 1st visualization paper published
- 2014:
Timeline of Visualization in Galaxy

1. 2005: 1st Galaxy paper published
2. 2008: Display applications
3. 2010 & 2011: Visualization development started
4. 2014: 1st visualization paper published

1. **visualization in Galaxy is nascent**
2. **you will be working with awesome new features**
3. **there may be bugs — help us fix them!**
Workshop Goals

**Participants**: learn about how to visualize your data in Galaxy
- biological visualizations
- numerical visualizations
- what Galaxy is doing underneath the covers

**Instructors**: feedback from you about what you like, don’t like, and where to go next
Galaxy Visualizations

Visualizations are first-class objects in Galaxy, just like tools

A visualization can be added to Galaxy via a configuration file that specifies:
  • datasets that can be used
  • location of visualization code (client-side or on server)

Galaxy handles visualization integration and data management, so users can focus on analyzing data (and developers can focus on creating visualizations)
Visualizations are 1st class Galaxy objects

Can be saved and versioned for reproducibility

Have a human-readable URL for sharing a fully interactive visualization:
http://usegalaxy.org/u/jgoecks/v/tumor-mutations

Can embed interactive visualizations in online supplementary materials via Galaxy Pages
Visualization Architecture

Client-server architecture

Lots of moving pieces

- prepare/process data on server
- send to client
- render on client
Topics

Visualization history and introduction

Biological Visualizations

Numerical Visualizations

Adding your own visualizations
Analysis goal: what similarities and differences can be found in cancer cell lines using exome and transcriptome sequencing?
Sequencing and Analysis

Profiled 3 pancreatic cancer cell lines using a 26-gene panel targeting known oncogenic driver mutations
  • MiaPaCa2, HPAC, and PANC-1

Data available:
  • exome: mapped reads, removed dups, called variants
  • transcriptome: mapped reads, assembled transcripts, computed expression
Display Applications

11: MiaPaCa2: Varscan variants
~150,000 lines
format: vcf, database: hg19

Log: tool progress
Log: tool progress
Picked up JAVA_OPTIONS:
Djava.io.tmpdir=/tmp
Got the following sample list:
MiaPaCa2-exome
Only variants will be reported
Min coverage: 8
Min reads2: 2
Min var freq: 0.01
Min avg qual: 15
P-value thresh: 0.99
Read

display at UCSC main
display with IGV web current local
display at RViewer main
Display Applications

11: MiaPaCa2: Varscan variants
~150,000 lines
format: vcf, database: hg19

Log: tool progress
Log: tool progress
Picked up JAVA_OPTIONS: -Djava.io.tmpdir=/tmp
Got the following sample list:
MiaPaCa2-exome
Only variants will be reported
Min coverage: 8
Min reads2: 2
Min var freq: 0.01
Min avg qual: 15
P-value thresh: 0.99
Read

- display at UCSC main
  display with VCF view current local
display at RViewer main
Display Applications

Advantages
- use tool that is familiar to you
- easy to view your data alongside public datasets

Disadvantages
- cannot save/share/version visualization
- many more visualizations than display applications in Galaxy
- no data processing or visual analysis, only visualization
Trackster—Galaxy’s Genome Browser
Trackster—Galaxy’s Genome Browser

Genome browsers remain a (the most?) powerful genome visualization
  • foundational tool

Trackster is for the high-throughput sequencing era
  • very large datasets, numerous simultaneous tracks
  • maximum flexibility for customization (e.g. rainbow tracks)
  • 2-3 indices per dataset for fast visualization

SAM/BAM, BED, GFF/GTF, VCF, Wiggle, BigWig, BigBed, BedGraph
Let’s visualize our data in Trackster

1. Create visualization
2. Add gene annotation (RefSeq)
3. Save visualization
4. Exit
5. Reopen visualization
Let’s visualize our data in Trackster

1. Create visualization
Let’s visualize our data in Trackster

1. Create visualization

2. Add gene annotation (RefSeq)
Let’s visualize our data in Trackster

1. Create visualization
2. Add gene annotation (RefSeq)
3. Save visualization
Let’s visualize our data in Trackster

1. Create visualization
2. Add gene annotation (RefSeq)
3. Save visualization
4. Exit
Let’s visualize our data in Trackster

1. Create visualization
2. Add gene annotation (RefSeq)
3. Save visualization
4. Exit
5. Reopen visualization
Behind the Scenes

Galaxy is indexing datasets for
- viewing large genomic regions (coverage plots)
- viewing small genomic regions (getting individual data points)
- feature names and locations

Indexes is the primary way that big datasets are visualized quickly
Modes and Searching

Tracks can be displayed differently
- coverage
- individual features

Let’s try different modes
- this is fast because data is sent from Galaxy server and rendered in your Web browser

Let’s try searching for a gene: ERBB2
Let’s Call Variants

VarScan
• Sample names: MiaPaCa2, PANC1, HPAC
• Run

Rename output: “Cell line variants”
Let’s Assemble Transcripts

Cufflinks
- input dataset is #7
- run

Rename output: “MiaPaCa2 Assembled Transcripts”
Let’s add data to Trackster

Add exome data for all cell lines...

...but where is our data?
Circster

Interactive Circos plot

Whole genome view with structural variation
Let’s view our data in Circster

Double-click or use trackpad to zoom in

change track min/max

what do we see?
Let’s add data to Circster and adjust options

1. Add transcriptome coverage data
Let’s add data to Circster and adjust options

1. Add transcriptome coverage data

2. Change arc dataset height
Let’s add data to Circster and adjust options

1. Add transcriptome coverage data

2. Change arc dataset height

3. Change max for tracks
Let’s add data to Circster and adjust options

1. Add transcriptome coverage data

2. Change arc dataset height

3. Change max for tracks

4. Save visualization
Back to Trackster: Rainbow Track for Coverage

1. Navigate to ERBB2 gene
2. Create group
3. Add transcriptome coverage tracks to group
4. Create composite track
5. Adjust max
6. what do we see?
Add More Data!

Add RNA-seq mapped reads, variants, and assembled transcripts

Look at ERBB2
  * bookmark!

Look at STK11
  * bookmark!

Look at KRAS —> LYRM5
  * bookmark!
Visual Analysis

KRAS and Variants
Sweepster
Topics

Visualization history and introduction
Biological Visualizations
Numerical Visualizations
Adding your own visualizations
What is Galaxy Charts?

Use Galaxy  
Create Tabular Results  
Visualize with Galaxy Charts
Import data files

Click on **Shared Data** and select **Data Libraries**. Navigate to the **Chart** library and import it into your history (data reference: http://dna.cs.byu.edu/treesaap and bacteriome.org).
Make a new chart (1 of 4)

Wait for the upload to complete. Select your **Dataset** and click on the **Visualization Icon** then select **Charts**.
Give your chart a name

Name your chart **Unclustered Heatmap**.
Select a chart type

Double click on the **Heatmap** icon.
Select data columns

At first click on **Row labels** and select **Column 2**. Then, click on **Draw**.
Unclustered Heatmap
Select your **Dataset** and click on the **Visualization Icon** then select **Charts**.
Give your chart a name

Name your chart **Clustered Heatmap**.
Select a new chart type

- Area charts
  - Regular (NVD3)
  - Expanded (NVD3)
  - Stream (NVD3)
  - Pie chart (NVD3)

- Data processing (requires 'charts' tool from Toolshed)
  - Histogram (NVD3)
  - Discrete Histogram (jqPlot)
  - Box plot (jqPlot)
  - Clustered Heatmap (Custom)

Double click on the **Clustered Heatmap** icon.
Select data columns

At first click on **Row labels** and select **Column 2**. Then, click on **Draw**.
Clustered Heatmap

Use the **mouse wheel or your touch pad** to zoom into the **highlighted area**.
Enlarged view

Tooltips popup if you move the mouse pointer over a box. Here the interaction between B4143 and B3295 is highlighted. Click on Editor again to further customize this chart.
Go to the **Configuration** tab.
Chart settings

<table>
<thead>
<tr>
<th>X axis:</th>
</tr>
</thead>
</table>
| Axis label                  | X-axis
Provide a label for the axis. |
| Axis value type             | Auto
Select the value type of the axis. |

<table>
<thead>
<tr>
<th>Y axis:</th>
</tr>
</thead>
</table>
| Axis label                  | Y-axis
Provide a label for the axis. |
| Axis value type             | Auto
Select the value type of the axis. |

<table>
<thead>
<tr>
<th>Others:</th>
</tr>
</thead>
</table>
| Show legend                 | Yes  No
Would you like to add a legend? |
| Color scheme                | Jet
Select a color scheme for your heatmap |
| Url template                | http://someurl.com?id=__LABEL__
Enter a url to link the labels with external sources. Use __LABEL__ as placeholder. |

Heatmap specific options are **highlighted**. Feel free to set **axis labels** or other options.
Define a URL template

Would you like to add a legend?

Jet

Select a color scheme for your heatmap


Enter a url to link the labels with external sources. Use __LABEL__ as placeholder

Paste a database URL into the template URL field and add the __LABEL__ tag. You may use http://www.ncbi.nlm.nih.gov or any other database. Click on Draw to redraw the chart.
Data points linked to web sources

Double click on a box and the browser will open two new tabs using the previously defined URL template.
Cluster selection and analysis

Select one element from each **highlighted row**. What are the corresponding **protein functions**?
Identified protein categories

- Chemotaxis
- RNA Polymerase
- Flagella
- Chaperone

Please return to the Editor.
Select your **Dataset** and click on the **Visualization Icon** then select **Charts**.
Give your chart a name

Name your chart **Score Histogram**.
Analyze the score distribution

Double click on the **Histogram** icon and click on **Draw**.
Give your chart a name

Click on **Draw**.
Click on **Screenshot** and select **Save as PNG**. Finally, return to the **Editor** again.
Make a new chart (4 of 4)

Select your **Dataset** and click on the **Visualization Icon** then select **Charts**.
Give your chart a name

Name your chart **Discrete Histogram**.
Analyze the protein distribution

- Area charts
  - Regular (NVD3)
  - Expanded (NVD3)
  - Stream (NVD3)
  - Pie chart (NVD3)

- Data processing (requires 'charts' tool from Toolshed)

Double click on the **Discrete Histogram** icon.
Add more data

Click on Add Data.
Select a second data group

At first click on **Observations** and select **Column 2**. Then, click on **Draw**.
Which proteins have most interactions?

Done with Part I.
Scratchbook
Activate the **Scratchbook** by clicking on the above icon.
Activate the Scratchbook

Click on **Saved Visualizations**.
Activate the Scratchbook

Select a Visualization and repeat the process by selecting **Saved Visualizations** again.
Scratchbook for multiple charts

Resize all visualizations so they fit into the screen.
More Examples
Create a pie chart

Select the imported datasets, create a new chart and select **Pie chart**. Then, click on **Add data**.
Add first data group

Configure the **Helix frequency** column.
Add second data group

Configure the **Beta frequency** column.
Configure the *pie chart* as shown above. Then, click on **Draw**.
Glutamic acids seem to fit much better into helices than beta sheets. In other words, "Aspartic and Glutamic Acids are Important for Alpha-helix Folding", JBSD 2007.
Create a bar diagram

Create data groups for the following features: **Hydrophobicity**, **Membrane frequency**, **Flexibility**, **Helix frequency** and **Beta frequency**.
Use the **tooltips** to identify the amino acids which are likely to be found within membrane proteins.
Topics

Visualization history and introduction
Biological Visualizations
Numerical Visualizations
Adding your own visualizations
Adding your own Visualizations

Go to config/plugins/visualizations/charts

ccharts/others/YOURVIZNAME

Add three files to this directory:

Logo (logo.png)
Configuration (config.js)
Wrapper (wrapper.js)

ccharts/types.js

Rebuild by typing ‘npm install’ and ‘grunt’