Galaxy as a Platform for High-throughput Genomics

Jeremy Goecks

EMORY UNIVERSITY
Topics

Galaxy

Analyzing Cancer Genomes & Transcriptomes

Web-based Visual Analysis
Genomic Analyses are Difficult

Investigators unfamiliar with computation

Creating and reproducing workflows (pipelines) hindered by complexity: systems, scripts, tools, parameters

Collaboration and reuse difficult because current approaches do not support computational artifacts well
Galaxy Project: Fundamental Questions

When genomics (or any other biomedical science) becomes dependent on computational methods, how to:

✦ make tools and workflows **accessible** to scientists?
✦ ensure that analyses are **reproducible**?
✦ enable transparent **communication and reuse** of analyses?
Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and collaborative computational genomics
Galaxy Demo
What is Galaxy?

Platform for high-throughput genomics
1. get and integrate public, private data
2. analyze data and create workflows
3. visualization, sharing, publication

Customizable open-source software on various HPC resources
- public website — http://usegalaxy.org
- local instance
- on the cloud
Galaxy platform

- run tools, workflows on HPC resources
- create workflows, visualizations, pages
- share everything
Cloud Launch
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
AKIAI3YLUWEIRJOVADAA
This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Console.

Secret Key
sra1cyCFkLOIxvsrsGMjycVlyYFgs3L55x2SAs
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
ccluster1
This is the name for your cluster. You'll use this when you want to restart.

Cluster Password

Cluster Password - Confirmation

Key Pair
Create New - cloudman_keypair

Instance Type
Large

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

Submit
Welcome to CloudMan. This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to manage services provided by the application.

Status

Cluster name: cluster1
Disk status: 2.9G / 10G (29%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications • Data •

Autoscaling is off. Turn on?
Cloud Features

Resource configuration

Autoscaling

Snapshottting
Galaxy is Very Popular

Public Website (http://usegalaxy.org), anybody can use:
- ~500 new users per month, ~200 TB of user data, ~130,000 analysis jobs per month

Used and cited in more than 1000 publications
Galaxy is Very Popular

Local installations all over the world

http://bioteam.net/slipstream/galaxy-edition/
Topics

Galaxy

Analyzing Cancer Genomes & Transcriptomes

Web-based Visual Analysis
Preliminary Data

6 patients, whole transcriptome sequencing (RNA-seq) of primary tumor
- mixed populations!
- 3 +ERCC, 3 -ERCC (via IHC)

MiaPaCa2 cell line
- whole transcriptome
- targeted exome

Total sequencing data: ~70 GB

http://en.wikipedia.org/wiki/RNA-Seq
The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity


Affiliations | Contributions | Corresponding authors

Nature 483, 603–607 (29 March 2012) | doi:10.1038/nature11003
Received 25 July 2011 | Accepted 01 March 2012 | Published online 28 March 2012
Big Questions

How closely does MiaPaCa2 match to primary pancreatic tumors?

How to match patient to “best” CCLE cell line(s)?
Using Galaxy for Analysis of Cancer Transcriptomes

New tools
- complement existing transcriptome analysis tools

New workflows
- workflows are understandable, extendable, sharable

New visual analysis applications
- visualize and call variants in a Web browser
Single Sample Transcriptome Analysis

**gene fusions**

**transcripts + expression levels**

**variants**
Comparing Called Variants with Public Datasets
## Patient Mutations vs.

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>OM MIA (4)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>OM PC (11)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>OM ALL (114)</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>HP MIA (84)</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>HP PC (1769)</td>
<td>16</td>
<td>23</td>
<td>19</td>
<td>11</td>
<td>23</td>
<td>8</td>
<td>39</td>
</tr>
<tr>
<td>HP ALL (64,669)</td>
<td>110</td>
<td>180</td>
<td>143</td>
<td>97</td>
<td>136</td>
<td>65</td>
<td>87</td>
</tr>
</tbody>
</table>

OM = OncoMap, HP = hybrid capture with probes

http://www.broadinstitute.org/ccle/home
Patient Mutations vs. Cell line does not appear very similar to tumors

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OM MIA (4)</strong></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td><strong>OM PC (11)</strong></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td><strong>OM ALL (114)</strong></td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><strong>HP MIA (84)</strong></td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td><strong>HP PC (1769)</strong></td>
<td>16</td>
<td>23</td>
<td>19</td>
<td>11</td>
<td>23</td>
<td>8</td>
<td>39</td>
</tr>
<tr>
<td><strong>HP ALL (64,669)</strong></td>
<td>110</td>
<td>180</td>
<td>143</td>
<td>97</td>
<td>136</td>
<td>65</td>
<td>87</td>
</tr>
</tbody>
</table>

OM = OncoMap, HP = hybrid capture with probes
## Patient Mutations to Predict Tumor Attributes

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>OM MIA (4)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OM PC (11)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OM ALL (114)</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HP MIA (84)</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>HP PC (1769)</td>
<td>16</td>
<td>23</td>
<td>19</td>
<td>11</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>HP ALL (64,669)</td>
<td>110</td>
<td>180</td>
<td>143</td>
<td>97</td>
<td>136</td>
<td>65</td>
</tr>
<tr>
<td>Tumor %</td>
<td>90%</td>
<td>90%</td>
<td>100%</td>
<td>0%?</td>
<td>60%</td>
<td>40%</td>
</tr>
</tbody>
</table>

OM = OncoMap, HP = hybrid capture with probes
Gene Expression Clustering

Spearman Correlation

-0.14
0.31
0.46
0.48
0.62
0.77
Gene Expression Clustering

Spearman Correlation

P1 P2 P3
0.48

P5 P4 P6
0.62
0.77

-0.14
0.31
0.46
0.62
0.77

CL
Matching Patients to Cell Lines
## Matching Patients to Cell Lines

<table>
<thead>
<tr>
<th>Patient</th>
<th>Best Match (#, % muts)</th>
<th>Best Match (#, % muts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>KP3, KP2, KP4, PANC0327, PANC1005, QGP1 (4)</td>
<td>KP3 (4, 4.8%)</td>
</tr>
<tr>
<td>P2</td>
<td>PANC0327 (8)</td>
<td>CAPAN2 (7, 9.1%)</td>
</tr>
<tr>
<td>P3</td>
<td>SNU410, QGP1 (6)</td>
<td>KP3 (5, 6.0%)</td>
</tr>
<tr>
<td>P4</td>
<td>CAPAN2, PANC0403, MIAPACCA2, PANC0327 (5)</td>
<td>CAPAN (5, 6.5%)</td>
</tr>
<tr>
<td>P5</td>
<td>T3M4 (8)</td>
<td>T3M4 (8, 8.7%)</td>
</tr>
<tr>
<td>P6</td>
<td>CAPAN2, MIAPACCA2 (3)</td>
<td>CAPAN2 (3, 4.8%)</td>
</tr>
</tbody>
</table>

*38 pancreatic cell lines in CCLE*
Topics

Galaxy
Analyzing Cancer Genomes & Transcriptomes

Web-based Visual Analysis
Mutation Calling from RNA-seq

Variant calling from 6 patients, 700GB pileup file requires 48 hours to complete
Why Visual Analysis?

Challenges for bioinformatics tools
- time and compute intensive
- many parameters
- parameter sensitive and unpredictable

Approach: repeatedly run tools and visualize outputs to compare and find best parameters/approach
Galaxy

Tools

Traditional Analysis

Datasets
Web-based Visualization for High-throughput Genomic Datasets

State-of-the-art data management
- automatic indexing for aggregate data and individual data points
- retrieve data on demand and cache

Render in browser for speed and flexibility

Can share and publish fully-functional visualizations

Goecks et al. (2011) IEEE BioVis
Demo: Visual Analysis
Real-time Visual Analysis

Interactive use of production tool to call and visualize variants for multiple patients using parameter sweeps

A general approach for interactive visual analysis on very large genomics datasets
  - any Galaxy visual application, many tools (original application: transcript assembly)
  - can decide what data to analyze on the fly

Goecks et al. (2012) *Nature Biotechnology*
Concluding Thoughts

Galaxy is a very useful platform for high-throughout genomics
✦ accessible, reproducible, collaborative
✦ public, local, cloud

New tools, workflows, and visual analysis tools for analyzing high-throughput cancer sequencing data
✦ match patients to drug-profiled cancer cell lines via variants
✦ and soon variants + gene expression

Visualization/visual analysis are first-class objects in Galaxy
✦ visual analysis affords rapid experimentation with tool parameters
Thanks!
Questions?

http://galaxyproject.org

http://jeremygoecks.com
jeremy.goecks@emory.edu