Understanding Cancer Genomes (and Transcriptomes!) using Galaxy

Jeremy Goecks

Department of Biology
Department of Math and Computer Science
Emory University
Advancing Computing and Genomics

Research model
1. find computing challenge while doing genomics
2. invent new computing technology to address challenge
3. demonstrate usefulness of new technology via genomics investigation
Advancing Computing and Genomics

Realizing research model
1. computing challenge: analyzing cancer genomes
2. new computing technology: Galaxy tools, workflows, and visual analysis
3. genomics investigation: pancreatic cancer transcriptome
Roadmap

Galaxy
Analyzing Cancer Genomes and Transcriptomes
Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and collaborative computational genomics
A User Perspective of Galaxy

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

Customizable open-source software on various HPC resources
- public website — http://usegalaxy.org
- local instance
- on the cloud
Module (plug-in) architecture
- {attributes + behaviors} define a module, and module implementations are written as needed

There are Galaxy module definitions/support for:
- tools
- data types (file formats)
- data sources (e.g., sequencers)
- data stores (file systems)
- job scheduling engines (e.g., DRMAA, Condor)
- visualizations/visual analysis (e.g., genome browser, Circos plot, scatterplot)
Roadmap

Galaxy

Analyzing Cancer Genomes and Transcriptomes
New Prostate Cancer Tests Could Reduce False Alarms

By Andrew Pollack

Sophisticated new prostate cancer tests are coming to market that might supplement the unreliable P.S.A. test, potentially saving tens of thousands of men each year from unnecessary biopsies, operations and radiation treatments.

Some of the tests are aimed at reducing the false alarms, and accompanying anxiety, caused by elevated P.S.A. readings. Others, intended for use after a definitive diagnosis, examine the genetic workings of the cancer to distinguish dangerous tumors that need treatment from slow-growing ones that might be left alone.

Cancer Centers Racing to Map Patients’ Genes

By Anemona Hartocollis

The promise of whole genome sequencing can be seen in trials like one for bladder cancer at Memorial, where the effects of a drug normally used for breast cancer were disappointing in all but one of about 40 patients, whose tumor went away, Dr. Baselga said. Investigators sequenced the patient’s whole genome. “The patient had a mutation in one gene that was right on the same pathway as the therapy,” Dr. Baselga said. “And that explained why this worked.”
Using Galaxy for Analysis of Cancer Genomes/Transcriptomes

New tools
- e.g. variant calling, fusion detection, variant annotation and filtering, VCF manipulation

New workflows
- workflows are understandable and extendable

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

- **Tophat2**
  - RNA-Seq FASTQ file, forward reads
  - RNA-Seq FASTQ file, reverse reads
  - Gene Model Annotations
    - fusions (tabular)
    - insertions (bed)
    - deletions (bed)
    - junctions (bed)
    - accepted_hits (bam)

- **Tophat Fusion Post**
  - BAM file of aligned RNA-Seq reads
  - Tabular file of potential fusions
    - results.txt (txt)
    - results.html (html)

- **Cufflinks**
  - SAM or BAM file of aligned RNA-Seq reads
  - Reference Annotation
  - Global model (for use in Tracks file)
    - genes_expression (tabular)
    - transcripts_expression (tabular)
    - assembled_isoforms (gtf)
    - total_map_mass (txt)

- **Mark Duplicate reads**
  - SAM/BAM dataset to mark duplicates
    - out_file (bam)
    - html_file (html)

- **VarScan**
  - BAM file 1 > BAM file
  - output_vcf (vcf)
    - output_vcf (vcf)
    - output_log (txt)

**Gene fusions**

**Transcripts + expression levels**

**Variants**
Advantages of Galaxy Workflows

Not a black box, so can swap out tools and modify parameters
- recomputable

Human readable, especially for non-programmers

“-able”: import, export, share, publish, embed
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>HP MIA (84)</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>HP PC (1769)</td>
<td>21</td>
<td>29</td>
<td>23</td>
<td>14</td>
<td>29</td>
<td>15</td>
<td>49</td>
</tr>
</tbody>
</table>

Cell line does not appear very similar to tumors

OM = OncoMap, HP = hybrid capture with probes
Using Mutations for Characterizing Tumors

<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>HP MIA (84)</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>HP PC (1769)</td>
<td>21</td>
<td>29</td>
<td>23</td>
<td>14</td>
<td>29</td>
<td>15</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
(See ISMB talk for discussion of clustering patients via gene expression data from RNA-seq.)
Variants + Gene Expression + Annotation: Targeted eQTL Analysis

Finds and annotates variants in differentially-expressed genes or isoforms
Galaxy

Tools

... ...

Traditional Analysis

Datasets

... ...
Galaxy

Datasets

Tools

... 

Visual Applications

Visualization

Traditional Analysis

Visual Analysis
Web-based Visualization for High-throughput Genomic Datasets

State-of-the-art data management
- automatic indexing for aggregate data and individual data points
- data on demand + multi-level caching

Can share and publish fully-functional visualizations

Framework for adding new visualizations
- similar to tool config in Galaxy
Sweepster
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
- workflows soon!
Thanks! And More:

http://galaxyproject.org
http://usegalaxy.org
http://bitbucket.org/galaxy/galaxy-central
http://wiki.galaxyproject.org

1. Longer Talk, more Biology
   Sunday, 14:10-14:35

2. Integrated Visualization & Computing Workshop
   Tuesday, 14:10-16:05