Using Galaxy to Understand Cancer Genomes

Jeremy Goecks

Computational Biology Institute

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC
Topics

The High-throughput Sequencing Era
Galaxy
Cancer Genomics with Galaxy
A Revolution in Biology

Low-cost, high-throughput sequencing technologies have become widespread

50-100 Gb / day
World Sequencing Capacity > 15Pb/year
The real cost of sequencing: higher than you think!

Andrea Booner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach and Mark B Gerstein

Pre-NGS (Approximately 2000)
Now (Approximately 2010)
Future (Approximately 2020)

Sample collection and experimental design
Sequencing
Data reduction
Data management
Downstream analyses
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
Topics

The High-throughput Sequencing Era

Galaxy

Cancer Genomics with Galaxy
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.

---

Goecks et al. (2010) *Genome Biology*
Galaxy Demo
Accessibility

All tools looks the same

No command line or programming

Easy to chain tools together into larger analyses
Reproducibility

Workflows enable reuse and provide precise reproducibility.

Users can add tags and annotations for additional context.
Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change


Summary of the paper

Polar bears (PBs) are usually adapted to the extreme Arctic environment and have become emblematic of the threat to biodiversity from global climate change. Their divergence from the lower-latitude brown bear provides a textbook example of rapid evolution of distinct phenotypes. However, limited mitochondrial and nuclear DNA evidence conflicts in the timing of PB origin as well as placement of the species within varus sister to the brown bear lineage. We gathered the first genomic data from contemporary polar, brown, and American black bear samples, in addition to a 110,000– to 1,600,000-year-old PB, to examine this problem from a genomic wide perspective. Nuclear DNA markers reflect a species tree consistent with expectations, showing polar and brown bears to be sister species. However, for the enigmatic brown bears native to Alaska’s Alexander Archipelago, we estimate that not only their mitochondrial genomes, but also 5–10% of their nuclear genomes, is most closely related to PBs, indicating ancient admixture between the two species. Explicit admixture analyses are consistent with ancient splits among PBs, brown bears and black bears that were later followed by occasional admixture. We also provide paleodemographic estimates that suggest bear evolution has tracked key climate events, and that PB is peculiarly experienced a prolonged and dramatic decline in its effective population size during the last ca. 100,000 years. We demonstrate that brown bears and PBs have had sufficiently independent evolutionary histories over the last 4–5 million years to leave imprints in the PB nuclear genome that likely are associated with ecological adaptation to the Arctic environment.

Datasets

Many of the analyses reported in the paper were based on the five datasets given here. (You can also find them under Shared Data -> Data Libraries -> Genome Diversity. Then under bear and dog.)

The first consists of 12,011,142 base-pair “SNPs”, i.e., positions in the dog genome where we detected two distinct nucleotides in the corresponding bear locations among the three brown bear species, polar bear, brown bear, and American black bear. Each row in the table corresponds to a SNP, and has 112 entries.

The bear assembly SNPs table contains 13,038,705 putative SNPs that were identified using a de novo assembly of the polar bear genome (rather than the dog assembly). Each row of the table corresponds to a SNP, and has 117 columns.

The bear mitochondrial SNPs table contains 8,949 positions where not all 20 individuals had the same nucleotide. Each row represents one of these SNPs, and has 11 columns.

One of the workflows (bear sweep table) uses a streamlined file with the locations of 13,014 dog genes (basically, each one is the longest of a set of overlapping splice variants). Each gene corresponds to a row of the table, which has 5 columns.

Workflows

This page shows three workflows that produce results presented in the polar bear paper. Almost all of the commands that they use are from the “Genome Diversity” tool set. (See the left panel under “Analysis Data.”)

The first workflow generates the data for Figure 4A of the paper. (These data were used to produce a more interactive PCA plot that includes other information.) The workflow needs to be applied to the “bear SNP” data set as follows: (1) Under “Analysis Data” in the black box create an empty history. (2) Under “Shared Data” -> “Published Pages”, use this page. (3) Import the “bear SNP” data set (labeled “1” in the green circle near the right of the green box), then click on the “Export page,” (4) Import the “bear SNP” workflow and click on “Start using this workflow.” (5) You will be taken to your Workflow page, which will have a workflow called “Imported bear PCA,” click on it and select “run.” (6) You will be taken to a history that includes the bear SNPs and the PCA workflow, and at the bottom of the workflow timeline panel and press “Run workflow.” (7) After the commands run, which takes a couple of minutes, click on the “eye” icon for the PCA and look at the three Outputs. (Currently, the PCA workflow exposes an internal error – a so-called “race condition” – in Galaxy, which may cause the PCA command to fail. If that happens, you can re-run the PCA, not the entire workflow by clicking on the line that says something like “7: PCA on data R,” clicking on the blue eye icon, and clicking on “Detect.” You also may need to give Galaxy a minute after the workflow finishes so it can put the output files in the correct places.)

Goecks et al., Genome Biology , 2010
Galaxy Datasets

Datasets

Galaxy Datasets in bear assembly SNP table contains 1,689 SNPs where not all 25 individuals had the same nucleotide. Each row represents one of these SNPs, and each column represents a unique SNP. Each row includes 12 columns.

Galaxy Datasets in bear assembly table contains 13,686 SNPs for the bear assembly, each column represents a unique SNP. Each row includes 12 columns.

Galaxy Workflow in bear assembly map

Galaxy Workflow in bear admixture map

Step 7: Admixture

Create admixture plots of the SNP dataset on the "bear assembly" page.

Step 8: Filter

Create filters for the SNP dataset on the "bear assembly" page.

Step 9: Summary table

Create a summary table of the SNP dataset on the "bear assembly" page.

Galaxy Workflow in bear summary table

Workflow "bear admixture map" has been inserted. You can click using this workflow or return to the services page.
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
- minimizes data movement
- 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
AKIAI3YLUEIRJOVADAA
This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Console.

Secret Key
sra1cyCFkLOIKvrsGMjyfCvLYFGs3LSSx2SAr
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
cluster1
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

**Cluster status log**
Welcome to Galaxy on the Cloud
managed by CloudMan
Cloud Features

Resource configuration
- CPUs (read mapping) vs. Memory (assembly)

Autoscaling
- automatically scale cluster as needed

Snapshotting
- share a complete Galaxy that others can copy and use
Visualization Framework

Galaxy

Tools
- Tophat for Illumina: Find splice junctions using RNA-seq data
- Tophat2: Gapped-read mapper for RNA-seq data
- Tophat Fusion Post: post-processing of Tophat results
- Cufflinks: transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare: compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge: merge together several Cufflinks assemblies
- Cuffdiff: find significant changes in transcript expression, splicing, and promoter use

FILTERING
- Filter Combined Transcripts using tracking file

NGS: SAM Tools
- NGS: Variant Detection
- NGS: Peak Calling
- NGS: Simulation

History
- Small Sample/Treatment
  Differential Expression Analysis
  10.7 MB
  15: Differential Transcript Expression
  2,533 lines
  format: tabular, database: hg19

1: Visualize

TCONS_00000001 = NM_00105248 XLOC_9
TCONS_00000002 = NM_130760 XLOC_9
TCONS_00000003 = NM_130762 XLOC_9
TCONS_00000004 = NM_033513 XLOC_9
TCONS_00000005 = NM_04259 XLOC_9
TCONS_00000006 = NM_05537 XLOC_9

14: Cuffdiff on data 1, data 2, and data 3: transcript FPKM tracking
13: Cuffdiff on data 1, data 2, and data 3: transcript differential expression testing
12: Cuffdiff on data 1, data 2, and data 3: gene FPKM
Scatterplot of 'Differential Transcript Expression'

Data Controls
Chart Controls
Statistics
Chart

Control FPKM

Treatment FPKM

AES
15163.4
4994.14
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

The High-throughput Sequencing Era
Galaxy

Cancer Genomics with Galaxy
Personalized Oncology

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

3 pancreatic cancer cell lines
- whole transcriptome
- targeted exome (cancer genes)

Total sequencing data: ~70 GB

http://en.wikipedia.org/wiki/RNA-Seq
Big Questions

What drugs is a patient likely to respond to given his/her genomic profile?
- genomic profile = mutations, gene expression, copy number, ...

How best to combine private (patient) data and public data?
- match profile to potential drugs?
- match profile to cell lines, then drugs and/or high-throughput screening?
Using Galaxy for Cancer Genomics

New tools
- complement existing high-throughput sequencing analysis tools

New workflows
- workflows are understandable, extendable, sharable

New visual analysis applications
- visualize and call variants in a Web browser
Integrates private and public data
Understandable, editable, sharable
Validation Using Public Data

- Highly targeted exome sequencing of 500+ cancer cell lines
- Drug response curves

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity


Affiliations | Contributions | Corresponding authors

Nature 483, 603-607 (28 March 2012) | doi:10.1038/nature1003
Received 25 July 2011 | Accepted 01 March 2012 | Published online 28 March 2012
Validation Results Using MiaPaCa2 Cell Line

Able to recover all 23 mutations, includes short insertions and deletions in CCLE

Found 16 druggable genes, leading to 98 potential drugs
Most cancer drugs are inhibitors, so gene expression is important
Single Sample Transcriptome Analysis

gene fusions

transcripts + expression levels

variants
Cufflinks DE data to Gene Regions

Workflow Canvas | Convert Cuff* tabular dataset to Interval

Input dataset
output

Cut
From
out_file1 (tabular)

Convert
in Dataset
out_file1 (tabular)

Convert
in Dataset
out_file1 (tabular)

Paste
Paste
and
out_file1
Validation Results Using MiaPaCa2 Cell Line

Able to recover all 23 mutations, includes short insertions and deletions in CCLE

After filtering for only mutations in highly-expressed genes:

Found 16 6 druggable genes, leading to 98 62 potential drugs
Matching Patients and Cell Lines
Comparing Called Variants with Public Datasets
Patient Mutations vs. OM MIA (4)

<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
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><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>

Cell line does not appear very similar to tumors

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

49
Clustering via Differential Expression
Gene Expression Clustering

Spearman Correlation

0.77
0.62
0.48
-0.14
0.31
0.46
0.62
0.77

P1
P2
P3
P5
P4
P6
CL
Gene Expression Clustering

Spearman Correlation

-0.14

0.31

0.46

0.48

P1

P2

P3

P5

P4

P6

CL
Interactive Visual Analysis
Mutation Calling from RNA-seq

Variant calling from 6 patient, 700GB pileup file requires 48 hours to complete
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
- driven by personalized oncology project
- integration is key: genomic profile with mutations + gene expression, private + public data
Thanks!

Questions?

http://galaxyproject.org

Postdoc and software engineers positions available in
Interactive Genomics Lab @ GW

http://jeremygoecks.com
jgoecks@gwu.edu