Galaxy
Web based platform for bioinformatics analysis

June 21, 2012

Local copy: https://galaxy.wi.mit.edu/
Joint project between BaRC and IT.
Main site: http://main.g2.bx.psu.edu/
Talk Outline

• The Galaxy interface
• Getting data into Galaxy
• Overview of the tools
• The Next Generation Sequencing tool box:
  – Preprocessing and quality control
  – Analysis of ChIP-seq
  – Analysis of RNA-seq
• Visualizing data on a genome browser and workflows available for analysis
Galaxy Interface:
A web based platform for analysis of large genomic datasets

- Type “https://galaxy.wi.mit.edu/” in your browser address.
- You will be prompted for your name and password (these are the same that you use for your email)

- No need of programming experience.
- Integrates many bioinformatics tools within one interface.
- Keeps track of all the steps performed in an analysis. Even if you delete the datasets, the history keeps the tools used.
Hot Topics:

Galaxy Interface: Analyze Data

- **Data analysis**

**Tools window**

- Get Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA manipulation
- NGS: SAM Tools
- NGS: Peak Calling
- Workflows

**Data display and tool’s dialog window**

- Green: job is finished
- Yellow: job is running
- Gray: job is in queue
- Red: there is a problem

**History window:**

- datasets for each analysis are kept here
Galaxy Interface: Workflow
Galaxy Interface: Shared Data
Talk Outline

• The Galaxy interface
• Getting data into Galaxy
• Overview of the tools
• The Next Generation Sequencing tool box
  – Preprocessing and quality control
  – Analysis of ChIP-seq
  – Analysis of RNA-seq
• Visualizing data on a genome browser and workflows available for analysis
Getting Data: Upload File

Upload File

File Format

Upload or paste file

Execute

Genome Assembly
Getting Data: Uploading Large Files

Step 1: copy your file to
/nfs/galaxy/uploads/username@wi.mit.edu
using a sftp client
Getting Data: Uploading Large Files

Step 2: Select and upload the file within galaxy

Execute Genome Assembly
Getting Data from UCSC (local copy)

UCSC Main

Table Browser

- Retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve data by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

- table: refGene
- region: genome = Mouse, position: chr12:57715963-57815592
- identifiers (names/accessions): paste list, upload list
- output format: BED - browser extensible data
- output file: (leave blank to keep output in browser)
- file type returned: plain text, gzip compressed

Get Output
Getting Data from UCSC (local copy)

Send to Galaxy
Data Uploaded

View Summary

Adapted from OpenHelix tutorial
Edit Attributes

**Convert to new format**

- Convert BED to GFF
  - This will create a new dataset with the contents of this dataset converted to a new format.
  - **Convert**

**Change data type**

- **New Type:**
  - **bed**
  - This will change the data type of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.
  - **Save**
History

• All steps are saved.
• Every time we do a new operation a new dataset is created. Data is not overwritten.
• Can share history with other Galaxy users.
• Can create workflow to repeat an analysis.
History

Good Practices

✓ Rename the outputs of your jobs
✓ Make a new history for each analysis that you perform.
✓ Permanently delete data that you don’t need (or you will reach your quota of 250Gb).
History is not removed when datasets are removed
Talk Outline

• The Galaxy interface
• Getting data into Galaxy
• Overview of the tools
• The Next Generation Sequencing tool box
  – Preprocessing and quality control
  – Analysis of ChIP-seq
  – Analysis of RNA-seq
• Visualizing data on a genome browser and workflows available for analysis
Overview of the tools: Lift-Over
Text Manipulation

- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset
- Secure Hash / Message Digest on a dataset
- Filter on ambiguities in polymorphism datasets
- Arithmetic Operations on tables

Text Manipulation
- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns
Filter and Sort: Filter data on any column

Filter (version 1.1.0)
Filter:
3: MACS on data an.peaks: bed
Dataset missing? See TIP below.

With following condition:
c1=='chr22'
Double equal signs, --, must be used as shown above. To filter for an arbitrary string, use the Select tool.

TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a “Condition/data issue”.

TIP: If your data is not TAB delimited, use Text Manipulation->Convert

Syntax
The filter tool allows you to restrict the dataset using simple conditional statements. Columns are referenced with c and a number. For example, c1 refers to the first column of a tab-delimited file.
Convert Formats:
GFF-to-BED

What it does
This tool converts data from GFF format to BED format (scroll down for format description).

Example
The following data in GFF format:

```
chr22 GeneA enhancer 1000000 1000100 500 +.
chr22 GeneB promoter 1001000 1001100 900 +.
```

Will be converted to BED (note that 1 is subtracted from the start coordinate):

```
chr22 1000000 1000100 enhancer 500 +.
chr22 1001000 1001100 promoter 900 +.
```

About formats
BED format Browser Extensible Data format was designed at UCSC for displaying data tracks in the Genome Browser. It has three required fields and several additional optional ones:
The first three BED fields (required) are:
Operate on Genomic Intervals:
Intersect the intervals of two datasets
Operate on Genomic Intervals: Intersect the intervals of two datasets
Other tools

Join, Subtract and Group

FASTA manipulation

Fetch Sequences and Fetch Alignments
Talk Outline

• The Galaxy interface
• Getting data into Galaxy
• Overview of the tools
• The Next Generation Sequencing tool box
  – Preprocessing and quality control
  – Analysis of ChIP-seq
  – Analysis of RNA-seq
• Visualizing data on a genome browser and workflows available for analysis
NGS Tools

Next Generation Sequencing Tools
NGS: QC and manipulation

Galaxy / WIBR

**Tools**

**NGS: QC and manipulation**

- **FASTQC: FASTQ/SAM/BAM**
  - Fastqc: Fastqc QC using FastQC from Babraham
  - ILLUMINA FASTQ
  - FASTQ Groomer: convert between various FASTQ quality formats
  - FASTQ splitter: on joined paired end reads
  - FASTQ joiner: on paired end reads
  - FASTQ Summary Statistics by column

- **GENERIC FASTQ MANIPULATION**
  - Filter FASTQ reads by quality score and length
  - FASTQ Trimmer by column
  - FASTQ Quality Trimmer by sliding window
  - FASTQ Masker by quality score
  - Manipulate FASTQ reads on various attributes
  - FASTQ to FASTA converter
  - FASTQ to Tabular converter
  - Tabular to FASTQ converter

Galaxy / WIBR

**Tools**

**FASTX-TOOLKIT FOR FASTQ DATA**

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement
- Trim sequences
Illumina data format

- Fastq format:

@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh

/1 or /2 paired-end

@seq identifier
seq
+any description
seq quality values
Sequence quality values on different FASTQ formats

http://en.wikipedia.org/wiki/FASTQ_format

<table>
<thead>
<tr>
<th>Character</th>
<th>Sanger</th>
<th>Solexa</th>
<th>Illumina 1.3+</th>
<th>Illumina 1.5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phred+33</td>
<td>raw reads typically (0, 40)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solexa+64</td>
<td>raw reads typically (-5, 40)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phred+64</td>
<td>raw reads typically (0, 40)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phred+64</td>
<td>raw reads typically (3, 40)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To discriminate between Solexa and Illumina 1.3+ check if your sequences' quality scores have any of the characters ;<==>?
FASTQ formats and FASTQ Groomer

**FASTQ Groomer**
- **FASTQ Groomer** convert between various FASTQ quality formats.
- **FASTQ Groomer** can convert between various FASTQ formats.
- **FASTQ Groomer** supports input formats like Illumina 1.3-1.7 and Solexa.

**What it does**
This tool offers several conversions options relating to the FASTQ format.

- When using Basic options, the output will be **sanger** formatted or **cssanger** formatted (when the input is Color Space Sanger).
- When converting, if a quality score falls outside of the target score range, it will be coerced to the closest available value (i.e., the minimum or maximum).
- When converting between Solexa and the other formats, quality scores are mapped between Solexa and PHRED scales using the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009 Dec 16.
- When converting between color space (csSanger) and base/sequence space (Sanger, Illumina, Solexa) formats, adapter bases are lost or gained; if gained, the base ‘G’ is used as the adapter. You cannot convert a color space read to base space if there is no adapter present in the color space sequence. Any masked or ambiguous nucleotides in base space will be converted to ’N’s when determining color space encoding.

**Quality Score Comparison**


1. Output from Illumina 1.8+ pipelines are Sanger encoded.
NGS: Quality Control

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Talk Outline

• The Galaxy interface
• Getting data into Galaxy
• Overview of the tools
• The Next Generation Sequencing tool box
  – Preprocessing and quality control
  – Analysis of ChIP-seq
  – Analysis of RNA-seq
• Visualizing data on a genome browser and workflows available for analysis
Analysis of ChIP-seq experiments

Preprocessing
Convert to FASTQ Sanger format

FASTQ Groomer

 Step 1
Map reads

Bowtie

 Step 2
Call peaks bound

MACS
Mapping Reads with **Bowtie**

![Galaxy / WIBR interface](image)

**Tools**
- Get Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- Map with Bowtie for Illumina
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- Workflows

**Map with Bowtie for Illumina (version 1.1.2)**

- **Select a reference genome:**
  - Mouse (mm9 Canonical)
  - Arabidopsis thaliana (TAIR9)
  - C. elegans (WS210)
  - D. melanogaster (dm3)
  - Human (hg18 Canonical)
  - Human (hg18 Full)
  - Human (hg19 Canonical)
  - Human (hg19 Full)
  - Mouse (mm10 Canonical)
  - Mouse (mm10 Full)
  - Mouse (mm8 Canonical)
  - Mouse (mm8 Full)

- **FASTQ file:**
  - 58: FASTQ Groomer on data 36
  - Must have ASCII encoded quality scores

- **Bowtie settings to use:**
  - **Full parameter list**
  - **Use the full parameter list for most mapping needs**
  - **Skip the first n reads (-s):**
    - 0
  - **Only align the first n reads (-u):**
    - -1
    - -1 for off
  - **Trim n bases from high-quality (left) end of each read before alignment (-S):**
    - 0

---

**BarC Hot Topics** Galaxy
Mapping Reads with **Bowtie**

[Image of Bowtie configuration options]

- Seed length: 36
- Minimum value is 5
- Whether or not to round to the nearest 10 and saturating at 30 (`--nomaqround`): Round to nearest 10
- Number of mismatches for SOAP-like alignment policy (`--v`): -1
- Whether or not to try as hard as possible to find valid alignments when they exist (`--y`): Do not try hard
- Report up to n valid alignments per read (`--k`): 1
- Whether or not to report all valid alignments per read (`--a`): Do not report all valid alignments
- Suppress all alignments for a read if more than n reportable alignments exist (`--m`): -1
- Write all reads with a number of valid alignments exceeding the limit set with the `--m` option to a file (`--max`):
- Write all reads that could not be aligned to a file (`--un`):
- Whether or not to make Bowtie guarantee that reported singleton alignments are 'best' in terms of stratum and in terms of the quality values at the mismatched positions (`--best`): Use best

Removes all strand bias. Only affects which alignments are reported by Bowtie. Runs slower with best option.
Analysis of ChIP-seq experiments: **MACS**
**MACS output**

Additional output created by MACS (MACS_in_Galaxy)

**Additional Files:**
- MACS in Galaxy diag.xls
- MACS in Galaxy model.pdf
- MACS in Galaxy model.txt
- MACS in Galaxy model.log
- MACS in Galaxy negative peaks.xls
- MACS in Galaxy peaks.xls

**Messages from MACS:**
INFO @ Thu, 07 Jun 2012 10:20:12:
# ARGUMENTS LIST:
# name = MACS_in_Galaxy
# format = SAM
# ChIP-seq file = /afs/galaxy2/galaxy-dist/databases/files/0000/dataset
# control file = /afs/galaxy2/galaxy-dist/databases/files/0000/dataset
# effective genome size = 1.87e+09
# tag size = 36
# band width = 300
# model fold = 10
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are: pvalue
INFO @ Thu, 07 Jun 2012 10:20:12: #1 read tag file
INFO @ Thu, 07 Jun 2012 10:20:12: #1 read treatment tags...
MACS output

![Galaxy/WIBR interface](image)

This dataset is large and only the first megabyte is shown below.

- Bed file with peaks

---

BarC Hot Topics: Galaxy
Creating Workflows
Workflow for ChIP-seq analysis
Example of downstream analysis: Intersect intervals of two datasets

Talk Outline

• The Galaxy interface
• Getting data into Galaxy
• Overview of the tools
• The Next Generation Sequencing tool box
  – Preprocessing and quality control
  – Analysis of ChIP-seq
  – Analysis of RNA-seq
• Visualizing data on a genome browser and workflows available for analysis
Expression Profiling Workflow

1. Preprocessing
   - FASTQ Groomer

2. Align with TopHat

3. Quantify transcripts with Cufflinks

**Hot Topics:**
- Galaxy
- TopHat
- Cufflinks

**Preprocessing**
- FASTQ Groomer

**Aligning**
- TopHat

**Quantifying**
- Cufflinks
Other tools for expression profiling

**Cuffcompare**: compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments.

**Cuffdiff**: find significant changes in transcript expression, splicing, and promoter use.

Align sample A with TopHat
Align sample B with TopHat
Compare samples with Cuffdiff

Talk Outline

• The Galaxy interface
• Getting data into Galaxy
• Overview of the tools
• The Next Generation Sequencing tool box
  – Preprocessing and quality control
  – Analysis of ChIP-seq
  – Analysis of RNA-seq
• Visualizing data on a genome browser and workflows available for analysis
Visualizing data on UCSC
Visualizing data on UCSC
Visualizing data on UCSC: BAM files

NGS: SAM Tools
- Filter SAM on bitwise flag values
- Convert SAM to interval

SAM-to-BAM converts SAM format to BAM format

7: Map with Bowtie for Illumina on data 3: mapped reads
~12,000,000 lines, 37 comments
format: sam, database: mm9
Info: Sequence file aligned.

78: SAM-to-BAM on data 7: converted BAM
501.9 Mb
format: bam, database: mm9
Info: Samtools Version: 0.1.18 (r982:295)
SAM file converted to BAM

display at UCSC bitters.wi.mit.edu
display with IGV web current local

Binary bam alignments file
Workflows available inside Whitehead

1. Workflow for ChIP-seq analysis.
2. Workflow for sorting a SAM file.
Documentation and Tutorials

- OpenHelix tutorials and exercises
  http://www.openhelix.com/cgi/tutorialInfo.cgi?id=82
- Galaxy tutorials
  http://galaxy.psu.edu/screencasts.html
- References
  Galaxy developers: The Center for Comparative Genomics & Bioinformatics, Pennsylvania State University
Previous Hot Topics

- Previous Hot Topics in Galaxy
  http://jura.wi.mit.edu/bio/education/hot_topics/GalaxyNGS/Galaxy_NGS.pdf

- Previous Hot Topics in NGS
  http://jura.wi.mit.edu/bio/education/hot_topics/ChIPseq/ChIPSeq_HotTopics.pdf