Using Galaxy for High-throughput Sequencing (HTS) Analysis

The Galaxy Team
http://UseGalaxy.org
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq, RNA-seq, Variant Detection
HTS Data

From the Sequencer:
• reads and quality scores (FASTQ)

In the Analysis Pipeline / Workflow:
• alignments against reference genome (SAM, BAM)
• annotations (GFF, BED)
• genome Assemblies (FASTA)
• quantitative tracks, e.g. conservation (WIG)
FASTQ Quality Scores

@UNIQUE_SEQ_ID
GATTTGGGGTTCAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAAGTTT
+
! ''* (((***++))%%%++) (%%%%) .1***--**' '')**55CCF >>>>>> CCCCCCCC655

Galaxy tools generally use Sanger format
• Need to convert quality scores to Sanger using Groomer tool

http://en.wikipedia.org/wiki/FASTQ_format
Getting Your Data into Galaxy

Cannot upload any file larger than 2GB via Web browser
  * Galaxy does not currently support compressed files

Use FTP client, e.g. FileZilla: http://filezilla-project.org/
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq and RNA-seq
Prepare and Quality Check

SOLiD
- SOLiD2fastq

Illumina
- Groomer
  - Quality Statistics
  - Generic Box Plot Tool
  - Read Trimmer
  - Quality Filter

454
- low quality splitter
- 454fastq

Combining Sequences and Qualities

This tool joins a FASTA file to a Quality Score file, creating a single FASTQ block for each read. Specifying a set of quality scores is optional; when not provided, the output will be fastq-sanger or fastq-ssanger (when a csfasts is provided) with each quality score being the maximal allowed value (93).

The tool can be used, for example, to convert 454-type output to FASTQ.
Grooming --> Sanger

[Image of Galaxy software interface showing FASTQ Groomer tool]
Quality Statistics and Box Plot Tool

NGS TOOLBOX BETA
NGS: QC and manipulation
ILLUMINA DATA
- 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

Graph/Display Data
- Histogram of a numeric column
- Scatterplot of two numeric columns
- Plotting tool for multiple series and graph types
- Boxplot of quality statistics

Box plot in Galaxy
FastQC Report

Mon 20 Jun 2011
dataset_1750787.dat

Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per sequence quality scores
- ✔ Per base sequence content
- ✔ Per base GC content
- ✔ Per sequence QC content
- ✔ Per base N content
- ✔ Sequence Length Distribution
- ✔ Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✗ Kmer Content

Basic Statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>dataset_1750787.dat</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Encoding</td>
<td>Sanger / Illumina 1.9</td>
</tr>
</tbody>
</table>
Read Trimming

FASTQ Trimmer

FASTQ File:
2: imported: GM12878..ple Dataset

Define Base Offsets as:
Absolute Values

Offset from 5' end:
0
Values start at 0, increasing from the left

Offset from 3' end:
16
Values start at 0, increasing from the right

Keep reads with zero length:

Execute

FASTQ Quality Trimmer

FASTQ File:
7: FASTQ Trimmer on data 2

Keep reads with zero length:

Trim ends:
5' and 3'

Window size:
1

Step Size:
1

Maximum number of bases to exclude from the window during aggregation:
0

Aggregate action for window:
min score

Trim until aggregate score is:
>=

Quality Score:
0.0

Execute
Filter FASTQ

FASTQ File:
7: FASTQ Trimmer on data 2
Requires groomed data: if your data does not appear here try using the FASTQ groomer.

Minimum Size:

Maximum Size:
A maximum size less than 1 indicates no limit.

Minimum Quality:

Maximum Quality:
A maximum quality less than 1 indicates no limit.

Maximum number of bases allowed outside of quality range:

This is paired end data:

Quality Filter on a Range of Bases

Add new Quality Filter on a Range of Bases

Execute

Quality Filter on a Range of Bases

Define Base Offsets as:

Absolute Values
Use Absolute for fixed length reads (Illumina, SOLiD)
Use Percentage for variable length reads (Roche/454)

Offset from 5' end:

Values start at 0, increasing from the left

Offset from 3' end:

Values start at 0, increasing from the right

Aggregate read score for specified range:

Keep read when aggregate score is:

Quality Score:

0.0

Remove Quality Filter on a Range of Bases

Add new Quality Filter on a Range of Bases

Execute
Manipulate FASTQ

FASTQ File:
7: FASTQ Trimmer on data 2
Requires groomed data: if your data does not appear here try using the FASTQ gromer.

Match Reads
Add new Match Reads

Manipulate Reads
Add new Manipulate Reads
Execute

Match Reads
Match Reads by:
Sequence Content
Sequence Match Type:
Regular Expression
Match by:
N
Remove Match Reads 1
Add new Match Reads
Manipulate Reads
Add new Manipulate Reads
Execute

Match Reads
Match Reads by:
Sequence Content
Sequence Match Type:
Regular Expression
Match by:
N
Remove Match Reads 1
Add new Match Reads
Manipulate Reads
Add new Manipulate Reads
Execute

Match Reads
Match Reads on:
Miscellaneous Actions
Miscellaneous Manipulation Type:
Remove Read
Remove Manipulate Reads 1
Add new Manipulate Reads
Execute
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq, RNA-seq, Variant Detection
Mapping HTS Data

Collection of interchangeable mappers
- accept fastq format, produce SAM/BAM

Mappers for
- DNA
- RNA
- Local realignment
Mappers

DNA
- short reads: Bowtie, BWA, BFAST, PerM
- longer reads: LASTZ

Metagenomics
- Megablast

RNA / gapped-reads mapper
- Tophat
Commonly Used/Default Parameters
### Default Best-Practices

- Fully customizable parameters

### Variable Levels of Settings

- **Full Parameter List**
  - Use Commonly used settings. If you want full control use Full List

### Selecting Settings

- **Seed hits require a 19 bp word with matches**
  - Allows you to set word size and number of mismatches

- **Select transition settings**
  - Allow one transition in each seed hit
  - Affects the number of allowed transition substitutions

### Performing Extension of Seed Hits to HSPs

- **Perform gap-free extension of seed hits to HSPs**
  - High-scoring segment pairs

- **Perform chaining of HSPs?**
  - **No**

### Gap Penalties

- **Gap opening penalty:** 400
- **Gap extension penalty:** 30

### X-drop and Y-drop Thresholds

- **X-drop threshold:** 910
- **Y-drop threshold:** 9370

### Setting HSP Thresholds

- **Set the threshold for HSPs (ungapped extensions scoring lower are discarded):** 3000
- **Set the threshold for gapped alignments (gapped extensions scoring lower are discarded):** 3000

### Entropy Filtering

- **Involves entropy when filtering HSPs?**
  - **No**

### Reference Name Modification

- **Do you want to modify the reference name?**
  - **No**

### Identity Filters

- **Do not report matches below this identity (%):** 0%
- **Do not report matches above this identity (%):** 100%
- **Do not report matches that cover less than this percentage of each read:** 0%

### Case Conversion

- **Convert lowercase bases to uppercase?**
  - **Yes**

### Execution

- **Execute**

### What it does

**LASTZ** is a high performance pairwise sequence aligner derived from BLASTZ. It is written by Bob Harris in Webb Miller's laboratory at Penn State University. Special scoring sets were derived to improve runtime performance and quality. This Galaxy version of LASTZ is geared towards aligning short (Illumina/Solexa, A/B/SOLiD) and medium (Roche/454) reads against a reference sequence. There is excellent, extensive documentation on LASTZ available [here](#).

### Input formats

LASTZ accepts reference and reads in FASTA format. However, because Galaxy supports implicit format conversion the tool will recognize fastq and other method specific formats.
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq, RNA-seq, Variant Detection
SNPs & INDELs

SNPs from Pileup
- Generate
- Filter
GATK Tools

Local re-alignment
Base re-calibration
Genotyping

Alpha status
* please try, report bugs
* available on test server: http://test.g2.bx.psu.edu/
Unified Genotyper

Inputs
- BAM files

Lots of possible parameters

Output
- VCF file(s)
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq, RNA-seq, Variant Detection
Peak Calling / ChIP-seq analysis

Punctate binding
- transcription factors

Diffuse binding
- histone modifications
- PolII
Punctate Binding --> MACS

Inputs
- Enriched Tag file
- Control / Input file (optional)

Outputs
- Called Peaks
- Negative Peaks (when control provided)
- Shifted Tag counts (wig, convert to bigWig for visualization)

Diffuse Binding

CCAT (Control-based ChIP-seq Analysis Tool)

I have Peaks, now what?

A Intersect
- First query
- Intervals to intersect with (second query)
- Overlapping intervals
- Overlapping pieces of intervals

B Subtract
- First query
- Intervals to subtract (second query)
- Intervals with no overlap
- Non-overlapping pieces of intervals

C Merge
- Query
- Merged Intervals

D Concatenate
- First query
- Second Query
- Concatenate

E Complement
- Query
- Complement

F Cluster
- Query
- Find clusters
- Merge clusters

Compare to other annotations using interval operations
A simple goal: determine number of peaks that overlap a) coding exons, b) 5-UTRs, c) 3-UTRs, d) introns and d) other regions

Get Data

Import Peak Call data

Retrieve Gene location data from external data resource

Extract exon and intron data from Gene Data (Gene BED To Exon/Intron/Codon BED expander x4)

Create an Identifier column for each exon type (Add column x4)

Create a single file containing the 4 types (Concatenate)

Complement the exon/intron intervals

Force complemented file to match format of Gene BED expander output (convert to BED6)

Create an Identifier column for the ‘other’ type (Add column)

Concatenate the exons/introns and other files

Determine which Peaks overlap the region types (Join)

Calculate counts for each region type (Group)
Secondary Analysis
Annotation Profiler

One click to determine base coverage of the interval (or set of intervals) by a set of features (tables) available from UCSC galGal3, mm8, panTro2, rn4, canFam2, hg18, hg19, mm9, rheMac2
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq, RNA-seq, Variant Detection
Transcriptome Analysis
(with a reference genome)

TopHat

Cufflinks/compare/diff

TopHat

Map RNA (FASTQ) to a reference Genome
  - gapped mapper

Outputs
  - BAM file of accepted hits
  - BED file of splice junctions
Cufflinks

Goal: transcript assembly and quantitation

Input: aligned RNA-Seq reads, usually from TopHat

Outputs
- assembled transcripts (GTF)
- genes’ and transcripts’ coordinates, expression levels
Cuffcompare

Goals
✦ generate complete list of transcripts for a set of transcripts
✦ compare assembled transcripts to a reference annotation

Inputs: assembled transcripts from Cufflinks

Outputs:
✦ Transcripts Combined File
✦ Transcripts Accuracy File
✦ Transcripts Tracking Files
Goals
✦ differential expression testing
✦ transcript quantitation

Inputs
✦ Combined set of transcripts
✦ mapped reads from 2+ samples

Outputs
✦ differential expression tests for transcripts, genes, splicing, promoters, CDS
✦ quantitation values for most elements

Cuffdiff
Next Steps

Filtering
- for differentially expressed elements
- combined transcripts (e.g. for those differentially expressed between samples)

Extract transcript sequences and profile sequences for function
Integrating Tools and Visualization
Working to add GATK Unified Genotyper (and more!) to Trackster as well
Working with HTS Tools

Often challenging
✦ many parameters
✦ time intensive
✦ evaluating results difficult

Good options
✦ filter early, filter often: easier to understand fewer results
✦ experimentation: can rerun tools, workflows
✦ visualization: use tools in Trackster when possible
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq, RNA-seq, Variant Detection
Sample information tracked in Galaxy, state changes through laboratory workflow are captured, data is linked back to sample in user’s workspace.
Sample Tracking System

Built-in system for tracking sequencing requests

Customizable interfaces
- Sequencing Facility Managers/Administrators
- Users/Biologists

Streamlines data delivery: sequencing runs to users
How does it all work?

Customer:
- Create Galaxy sequencing request
- Add samples
- Submit request

Facility Manager:
- Transfer datasets to Galaxy data library
- Update sample states
- Assign barcodes to samples
Sequencing Facility Managers

Setup the Galaxy sample tracking system according to the core facility workflow [Once per request type]

Create and submit a sequencing request on behalf of another user

Reject an incomplete or erroneous sequencing request

Receive samples and assign them tracking barcodes.

Setup data transfer from the sequencer
Sequencing Facility Users

Create and submit a sequencing request
Edit and resubmit a rejected sequencing request
Obtain datasets at the end of a sequencing run
Select Libraries and Histories, and Workflows to populate and run on sequenced samples.
Configure Available Request / Sample Options
Configurations can be
- custom-built
- loaded from provided configuration files
Configure the Sequencer
User Creates a Request
User Creates a Request

**Sequencing Requests**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Items</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For 0 selected requests: **Delete**  **Undelete**

**Create a new sequencing request**

Select a request type configuration:
- Atlantic Biosciences

Contact the lab manager if you are not sure about the request type configuration.

Name of the Experiment
- My first ChIP-seq Experiment

Description
- This is Experiment was performed using the protoc

Name
- (Optional)

Scientific Contact
- dan@bx.psu.edu office address
  - office
  - Penn State University
  - Wartik Lab
  - University Park PA 16803
  - United States
  - Phone: 867-5309

(Optional)

**Save**  **Add samples**
User Adds a Sample

### Add Samples to Sequencing Request "My first ChIP-seq Experiment"

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>Data Library</th>
<th>Folder</th>
<th>History</th>
<th>Workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample_1</td>
<td></td>
<td>Dan's Sequencing Requests</td>
<td>ChIP-seq</td>
<td>My own ChIP-seq Experiment!</td>
<td>Dan's ChIP-seq Workflow</td>
</tr>
</tbody>
</table>

For each sample, select the data library and folder in which you would like the run datasets deposited. To automatically run a workflow on run datasets, select a history first and then the desired workflow.

#### Layout1

**Copy 1 samples from sample** [None]

Select the sample from which the new sample should be copied or leave selection as None to add a new "generic" sample.

- [Add sample](#)  [Save]  [Cancel]

Click the Add sample button for each new sample and click the Save button when you have finished adding samples.

#### Import samples from csv file

History
Workflow to run
Samples Added, Submit Request

Add Samples to Sequencing Request "My first ChIP-seq Experiment"

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>Data Library</th>
<th>Folder</th>
<th>History</th>
<th>Workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample_1</td>
<td>Unsubmitted</td>
<td>Dan's Sequencing Requests</td>
<td>ChIP-seq</td>
<td>My own ChIP-seq Experiment</td>
<td>Dan's ChIP-seq Workflow</td>
</tr>
</tbody>
</table>

For each sample, select the data library and folder in which you would like the run datasets deposited. To automatically run a workflow on run datasets, select a history first and then the desired workflow.

Layout1

Copy 1 samples from sample None

Select the sample from which the new sample should be copied or leave selection as None to add a new "generic" sample.

Add sample

Click the Add sample button for each new sample.

Import samples from csv file
Samples enter “New” state
Sequencing Facility is informed of Request

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Samples</th>
<th>Type</th>
<th>Last Updated</th>
<th>State</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>My first ChIP-seq Experiment</td>
<td>This is Experiment was performed using the protocol ...</td>
<td>1</td>
<td>Atlantic Biosciences</td>
<td>26 minutes ago</td>
<td>In Progress</td>
<td><a href="mailto:dan@bx.psu.edu">dan@bx.psu.edu</a></td>
</tr>
<tr>
<td>new request</td>
<td></td>
<td>1</td>
<td>Atlantic Biosciences</td>
<td>3 days ago</td>
<td>Complete</td>
<td><a href="mailto:customer@corp.com">customer@corp.com</a></td>
</tr>
<tr>
<td>some experiment test</td>
<td>a test description</td>
<td>1</td>
<td>Atlantic Biosciences</td>
<td>3 days ago</td>
<td>Complete</td>
<td><a href="mailto:customer@corp.com">customer@corp.com</a></td>
</tr>
</tbody>
</table>
Sequencing Facility Receives Samples

- Facility assigns a barcode to sample tubes
- Scans barcode at each step to change state

User can watch progress of sequencing request
Sequencing Finished

Datasets are transferred from sequencer into Galaxy
- library
- user’s history

Galaxy Workflow is executed on Dataset

User is automatically emailed
Extending Sample Tracking with ngLims

An add-on written by community contributor Brad Chapman

http://bitbucket.org/chapmanb/galaxy-central

https://bitbucket.org/galaxy/galaxy-central/wiki/LIMS/nglims

Sample tracking is completely extensible

Track manually, with barcodes, or integrate with an existing LIMS

Everything is configuration driven, capture whatever data and support whatever workflow you want

Interaction with sequence instruments and secondary analysis is completely pluggable
  • For services that provide a web / REST API even easier
Example: extensions from Brad Chapman for flowcell layout, multiplexing, ...
Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq, RNA-seq, Variant Detection
Supported by the **NHGRI (HG005542, HG004909, HG005133), NSF (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health**
Using Galaxy

Use public Galaxy server: UseGalaxy.org
Download Galaxy source: GetGalaxy.org
Galaxy Wiki: GalaxyProject.org
Screencasts: GalaxyCast.org
Public Mailing Lists
• galaxy-bugs@bx.psu.edu
• galaxy-user@bx.psu.edu
• galaxy-dev@bx.psu.edu
ChIP-seq and RNA-seq exercises

RNA-seq
  - start Tophat mapping first (second section), then look at QC (first section)

Chip-seq
- http://usegalaxy.org/u/james/p/exercise-chip-seq

Variant Detection
- Using GATK, Picard Tools

http://ec2-50-16-165-27.compute-1.amazonaws.com/