NGS Analysis Using Galaxy
Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV
Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV
Galaxy is an open-source framework for integrating various computational tools and databases into a cohesive workspace.

- A web-based service we provide, integrating many popular tools and resources for comparative genomics.

- A completely self-contained application for building your own Galaxy style sites.
Data intensive biology for everyone.

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Get Galaxy

Learn Galaxy

Get Involved

Search all resources

The Galaxy Team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

http://galaxyproject.org/
Galaxy main web interface

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

Have mass spec data? Galaxy-P!

Three ways to Galaxy-P.
Public usegalaxyx.org
Local getgalaxy.org
Cloud biocloudcentral.msi.umn.edu
or maybe 4.

Tweets

Galaxy Project @galaxyproject
Storage system updates at TACC on Tues, Dec. 9 AM to 6 PM EST (UTC -0500). During this time, usegalaxy.org will be unavailable.

Galaxy Project @galaxyproject
Dec 9 Workshop: An Introduction to De Novo Genome Assembly using Galaxy, UQ St Lucia, Mark Crowe (@markcrowe) bit.ly/1cV6y7M

Galaxy Project @galaxyproject
December 2013 Galaxy Update: 50+ public servers (3 new), 63 new

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Biology and Biostatistics & Consumer Science departments at Emory University.

This instance of Galaxy is utilizing infrastructure generously provided by the iPlant Collaborative at the Texas Advanced Computing Center, with support from the National Science Foundation.
Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV
Galaxy: the instant web-based tool and data resource integration platform

- Open Source downloadable package that can be deployed in individual labs

- Modularized
  - Add new tools
  - Integrate new data sources
  - Easy to plug in your own components

- Straightforward to run your own private galaxy server
Welcome to IIGB’s Galaxy Server!

Overview
Galaxy is an open, highly customizable, web-based platform for the analysis of next generation sequence data and many other biological data types. It enables users to run computationally demanding next generation sequencing analysis tasks on powerful server hardware from a graphical web browser-based user interface rather than the Linux command-line. A subset of application supported by Galaxy is given in the left pane. Much more detailed descriptions of Galaxy’s basic functionalities including user tutorials are available here.

Why Local Galaxy Service?
There are many advantages of using a local Galaxy server here at UCR rather than public test instances of Galaxy available on the internet. The most important are: (1) shorter waiting queues for analysis tasks; (2) elimination of time consuming uploads of large data sets; (3) support for analyzing much larger data sets than this is possible on public services; (4) the ability to customize software tools and database collections.

How to Gain Access?
This instance of Galaxy runs on IIGB’s high performance compute (HPC) infrastructure, called Biocluster. As such its usage is covered by the annual registration fee for this infrastructure (see here for details). Users with an active Biocluster account can access this Galaxy service using their existing user name and password without any extra cost. New account requests for this service can be sent to support@biocluster.ucr.edu.

Additional Databases and Software Tools
Support requests for including additional reference genomes and software tools on IIGB’s Galaxy server can be sent to support@biocluster.ucr.edu.

Workshops on Galaxy
Past and future UCR workshop events on using Galaxy are listed here. The user manual from previous workshops can be accessed here.

Enter IIGB’s Galaxy Service
To enter this service, click here.
Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV
Galaxy – the one stop shop for Genome Analysis

- **Analyze**
  - Retrieve data directly from popular data resources or upload your own.
  - Interactively manipulate genomic data with a comprehensive and expanding best-practices toolset.
  - Galaxy is designed to work with many different datatypes. (Link)

- **Visualize**
  - Trackster is Galaxy’s visualization and visual analysis environment.
  - See more details (Link)

- **Publish and Share**
  - Results and step-by-step analysis record (Data Libraries and Histories)
  - Customizable pipelines (Workflows)
  - Complete protocols (Pages)
Tools and Datasources

- **Datasources**
  - Upload file from your computer
  - UCSC table browser
  - BioMart, modENCODE fly server

- **Tool Suites**
  - Text manipulation
  - Format converters
  - NGS
  - Graphing plotting
  - More
Datasets are accessible from within Galaxy or for download.

<table>
<thead>
<tr>
<th>Data Library Name</th>
<th>Data Library Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes</td>
<td>Data from the 1000 Genomes Project FTP site</td>
</tr>
<tr>
<td>AC-exome</td>
<td>Data for Nature Letter &quot;Complete Khoisan and Bantu genomes from southern Africa&quot;</td>
</tr>
<tr>
<td>Bushman</td>
<td>Data for Nature Letter &quot;Complete Khoisan and Bantu genomes from southern Africa&quot;</td>
</tr>
<tr>
<td>ChIP-Seq Mouse Example</td>
<td>Data used in examples that demonstrate analysis of ChIP-Seq data</td>
</tr>
<tr>
<td>Chobi</td>
<td>Contains user guide, reference files, and configuration files for the Cloudmap WCS analysis pipeline</td>
</tr>
<tr>
<td>CloudMap</td>
<td>Contains user guide, reference files, and configuration files for the Cloudmap WCS analysis pipeline</td>
</tr>
<tr>
<td>Codon Usage Frequencies</td>
<td>Codon Usage Frequencies</td>
</tr>
<tr>
<td>Coleman</td>
<td>IonPCG</td>
</tr>
<tr>
<td>Denisovan sequences</td>
<td>Files from &quot;A high-coverage genome sequence from an archaic Denisovan Individual&quot; Meyer et al. Science 2012 and basic processed data.</td>
</tr>
<tr>
<td>Erythroid Epigenetic Landscape</td>
<td>Dynamics of the epigenetic landscape during erythroid differentiation after CATA1 restoration</td>
</tr>
<tr>
<td>Evolutionary Trajectories in a Phage</td>
<td>Experimental evolution (Illumina)</td>
</tr>
<tr>
<td>GATK</td>
<td>Consortium</td>
</tr>
<tr>
<td>GCAT</td>
<td>Genome Diversity</td>
</tr>
<tr>
<td>Genome Diversity</td>
<td>Nucleotide polymorphisms for several threatened species</td>
</tr>
<tr>
<td>gnu_1000GP</td>
<td>gnu_1000GP</td>
</tr>
<tr>
<td>He_2010</td>
<td>He_2010</td>
</tr>
<tr>
<td>Heteroplasmwy</td>
<td>Data for Genome Biology 2011 manuscript</td>
</tr>
<tr>
<td>iGenomes</td>
<td>Selected files from Illumina iGenomes collection</td>
</tr>
<tr>
<td>Illumina BodyMap 2.0</td>
<td>RNA-seq data for the Illumina BodyMap 2.0 project</td>
</tr>
<tr>
<td>Illumina iDEA Datasets (sub-sampled)</td>
<td>Sub-sampled versions of datasets used for the Illumina iDEA challenge</td>
</tr>
<tr>
<td>Irish whole genome</td>
<td>Irish whole genome sequence and analysis</td>
</tr>
</tbody>
</table>
Workflows

- Workflows specify the steps in a process.
- Workflows are analysis that are meant to be run, each time with different user-provided datasets.
Pages are documentation within that Galaxy that explain the steps and reasoning in a particular history or workflow.
History

- Histories are all steps in the process and the used setting.
- Histories can be imported into your session and rerun as it is or modified.
User Account

- An account is not required to access the Galaxy public Main or Test instances,
- But if used, the data quota is increased and full functionality across sessions opens up, such as naming, saving, sharing, and publishing Galaxy objects (Histories, Workflows, Datasets, Pages).
Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV
NGS Data

- Raw: Sequencing Reads (FASTQ)
- Derived
  - Alignments against reference genome (SAM/BAM)
  - Annotations
    - GFF/GTF
    - BED
FASTQ Format

- A FASTQ file normally uses four lines per sequence.
- Line 1 begins with a ‘@’ character and is followed by a sequence identifier.
- Line 2 is the raw sequence letters.
- Line 3 begins with a ‘+’ character, is optionally followed by the same sequence identifier.
- Line 4 encodes the phred quality values for the sequence in line 2, each value represents the error probability of a given base call.

```bash
@SRR064154.208 HWUSI-EAS627_1:8:1:2:1681 length=38
ANGANNNGGACTTTTGAAAAGAGAGTCAAAGAGTGCTTG
+
?!08!!3C?BCBB<BCBB?BBACABBBBBBBBBB@CABAB
```
Quality score represents the error probability of a given basecall.

In fastq file, quality score are often represented using the ASCII alphabet.

For example, a Phred score of 40 can be represented as the ASCII char “I” (40+33= ASCII #73), and an Illumina score of 40 as “h” (40+64=ASCII #104).

The range of scores will depend on the technology and the base caller used, but will typically be up to 40.

| S | Sanger, Illumina 1.8+ | Phred+33, raw reads typically (0, 41) |
| X | Solexa | Solexa+64, raw reads typically (-5, 40) |
| I | Illumina 1.3+ | Phred+64, raw reads typically (0, 40) |
| J | Illumina 1.5+ | Phred+64, raw reads typically (3, 40) |

with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
SAM Format

- SAM stands for Sequence Alignment/Map format.
- For more details: [http://samtools.sourceforge.net/SAM1.pdf](http://samtools.sourceforge.net/SAM1.pdf)
- Consists of header and alignment section
- 11 mandatory fields
GFF and GTF format

- General feature format (GFF)
  
  ```
  browser position chr22:10000000-10025000
  browser hide all
  track name=regulatory description="TeleGene(tm) Regulatory Regions"
  visibility=2
  chr22 TeleGene enhancer 10000000 10001000 500 + . touch1
  chr22 TeleGene promoter 10010000 10010100 900 + . touch1
  chr22 TeleGene promoter 10020000 10025000 800 - . touch2
  ```

- Gene Transfer format (GTF)
  
  - The list attribute must begin with 2 mandatory attributes.
  - `Gene_id_value, transcript_id_value`

  ```
  gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
  ```
**BED format**

- Flexible way to define the data lines in the annotation track.

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

**BCF/VCF format**

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=RH,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=Q0,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=Q0,Number=2,Type=Integer,Description="Haplotype Quality">

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Sample1 Sample2
1 4370 . rs6057 G A 29 . NS=2;DP=13;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:18:1:52,51 1|0:48:8:51,51
2 7330 . T A 3 q10 NS=5;DP=12;AF=0.017 GT:GQ:DP:HQ 0|0:56:3:58,50 0|1:3:5:56,3
1 110695 rs6058 A,T 67 PASS NS=2;DP=10;AF=0.333;0.567;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
1 130237 . T . 47 . NS=2;DP=16;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51
2 134567 microsat1 GTCT G,GIAGT 50 PASS NS=2;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2
```
Available NGS Analysis Toolsets

- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMTools
- SNP and INDEL analysis
- RNAseq analysis
- Peak calling / CHIP-seq
NGS Analysis Using Galaxy

- Galaxy overview and Interface
- Getting Data in Galaxy
- Analyzing Data in Galaxy
  - Quality Control
  - Mapping Data
- History and workflow
- Sequences and Alignment Format
- Galaxy Exercises
Getting started with Galaxy

Data intensive biology for everyone.

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Get Galaxy

Learn Galaxy

Get Involved

Use [project's free server](http://galaxyproject.org/) or [other public servers](http://galaxyproject.org/).

Install locally or [in the cloud](http://galaxyproject.org/) or get Galaxy on SlipStream.

Screencasts, Galaxy 101, ...

Mailing lists, Tool Shed, wiki

Search all resources

The Galaxy Team is a part of BX at [Penn State](http://pennstate.edu/), and the Biology and Mathematics and Computer Science departments at [Emory University](http://emory.edu/). The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

http://galaxyproject.org/
Obtain data from many data sources including the UCSC Table Browser, BioMart, WormBase, or your own data.

Prepare data for further analysis by rearranging or cutting data columns, filtering data and many other actions.

Analyze data by finding overlapping regions, determining statistics, phylogenetic analysis and much more.
Galaxy Interface Sections

- **User**
  - Contains links to the downloading, preparation and analysis tools.

- **Register**
  - Show you the history of your analysis steps, allow you view data and results, and more.

- **Center Column**
  - The center column is where the menus and data will appear.
Contains links to the downloading, preparation and analysis tools.

The center column is where the menus and data will appear.

Register

show you the history of your analysis steps, allow you view data and results, and more.

User
NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- **Getting Data in Galaxy**
- Analyzing Data in Galaxy
  - Quality Control
  - Mapping Data
- History and workflow
- Galaxy Exercises
Getting Data

Click Get Data
Getting Data: Upload File

1. **Upload File**
2. **File Format**
3. **Upload or paste file**
4. **Execute**
5. **Species**
Getting Data: Upload File

Specify multiple URLs into the "URL / Text" box.

http://bx.mathcs.emory.edu/outgoing/data/phiX174_genome.fa
http://bx.mathcs.emory.edu/outgoing/data/phiX174_reads.fastasanger
NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- Getting Data in Galaxy
- **Analyzing Data in Galaxy**
  - Quality Control
  - Mapping Data
- History and workflow
- Galaxy Exercises
Analyzing Data: Next Generation Sequencing

What it does

Allows one to find sequence variants and/or sites covered by a specified number of reads with bases above a set quality threshold. The tool works on single FASTA files. However, it also allows you to specify columns in the input file manually. The tool assumes the following:

- the quality scores follow phred33 convention, where input qualities are ASCII characters equal to the Phred quality plus 33.
- the pileup dataset was produced by the `samtools pileup` command (although you can override this by setting column assignments manually).

Types of pileup datasets

The descriptions of the following pileup formats are largely based on information that can be found on the `SAMTools` documentation page. The 6- and 6 column pileup:

- 6 column pileup:
- 6 column pileup:
Analyzing Data: Next Generation Sequencing

FASTQ file manipulation, like format conversation, summary statistics, trimming reads, filtering reads by quality score...
Analyzing Data: Next Generation Sequencing

Input: sanger FASTQ
Output: SAM format
Analyzing Data: Next Generation Sequencing

Filter pileup

Select dataset:

which contains:

- Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

3

Pileup lines with coverage lower than this value will be skipped

Only report variants:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals:

No

See "Output format" below for explanation

Print total number of differences:

No

See "Example 3" below for explanation

Print quality and base string:

Yes

See "Example 4" below for explanation

Execute

What it does

Allows one to find sequence variants and/or sites covered by a specified number of reads with bases above a set quality threshold. The tool works on six and ten column pileup formats produced with samtools pileup command. However, it also allows you to specify columns in the input file manually. The tool assumes the following:

- the quality scores follow phred33 convention, where input qualities are ASCII characters equal to the Phred quality plus 33.
- the pileup dataset was produced by the samtools pileup command (although you can override this by setting column assignments manually).

Types of pileup datasets

The descriptions of the following pileup formats are largely based on information that can be found on the SAMTools documentation page. The 6- and 10-column variants are described below.

Six column pileup:

1. chromosome
2. position
3. read name
4. number of reads
5. total bases
6. quality scores
NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- Getting Data in Galaxy
- Analyzing Data in Galaxy
  - Quality Control
  - Mapping Data
- History and workflow
- Galaxy Exercises
History: History Options

List saved histories and shared histories. Work on Current History, create new, clone, share, create workflow, set permissions, show deleted datasets or delete history.
Creates a workflow, allows user to repeat analysis using different datasets.
What’s next?

- Galaxy exercises
  - SNP-Seq
  - RNA-Seq

- Visualization
  - IGV