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, a web-based genome analysis platform

- **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.
Galaxy Project Interface

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
- Use project's free server or other public servers

Get Galaxy
- Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy
- Screencasts, Galaxy 101, ...

Get Involved
- Mailing lists, Tool Shed, wiki

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.
Galaxy analysis 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.

https://usegalaxy.org/
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 Data Sources

- **Data Sources**
  - Upload file from your computer
  - UCSC table browser
  - BioMart, modENCODE, GrameneMart, WormBase servers.....

- **Tool Suites**
  - Text manipulation
  - Join, Subtract and Group
  - Format converters
  - NGS
  - Graph plotting
  - Motif tools
  - More......
Data Libraries

- Datasets are accessible from within Galaxy or for download.
Workflows specify the steps in a process.

Workflows are analysis that are meant to be run, each time with different user-provided datasets.
Pages

- Pages are documentation within the 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
    - VCF / BCF
  - 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.

```plaintext
@SRR064154.208 HWUSI-EAS627_1:8:1:2:1681 length=38
ANGANNNGGACTTTGAAAAAGAGAGTCAAAGAGTGCTTG
+
?!08!!!3C?BCBB<BCBB?BBACABBBBBBB@CABAB
```
FASTQ Quality Score

- Quality score represents the error probability of a given basecall.
- In a FASTQ file, quality scores 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.

---

### Quality Score Representation

<table>
<thead>
<tr>
<th>Score</th>
<th>ASCII Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>!</td>
</tr>
<tr>
<td>5</td>
<td>#</td>
</tr>
<tr>
<td>9</td>
<td>$</td>
</tr>
<tr>
<td>10</td>
<td>&amp;</td>
</tr>
<tr>
<td>15</td>
<td>'</td>
</tr>
<tr>
<td>20</td>
<td>(</td>
</tr>
<tr>
<td>26</td>
<td>)</td>
</tr>
<tr>
<td>31</td>
<td>+</td>
</tr>
<tr>
<td>33</td>
<td>/</td>
</tr>
<tr>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>53</td>
<td>4</td>
</tr>
<tr>
<td>54</td>
<td>5</td>
</tr>
<tr>
<td>55</td>
<td>6</td>
</tr>
<tr>
<td>56</td>
<td>7</td>
</tr>
<tr>
<td>57</td>
<td>8</td>
</tr>
<tr>
<td>60</td>
<td>9</td>
</tr>
<tr>
<td>63</td>
<td>:</td>
</tr>
<tr>
<td>64</td>
<td>;</td>
</tr>
<tr>
<td>65</td>
<td>&lt;</td>
</tr>
<tr>
<td>67</td>
<td>&gt;</td>
</tr>
<tr>
<td>73</td>
<td>`</td>
</tr>
<tr>
<td>91</td>
<td>^</td>
</tr>
<tr>
<td>104</td>
<td>_</td>
</tr>
<tr>
<td>126</td>
<td>`</td>
</tr>
</tbody>
</table>

### Quality Score Ranges

- **Sanger** (Phred+33): raw reads typically (0, 40)
- **Soxlexa** (Phred+64): raw reads typically (-5, 40)
- **Illumina 1.3+** (Phred+64): raw reads typically (0, 40)
- **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).
- **Illumina 1.8+** (Phred+33): raw reads typically (0, 41)
**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) ([Link](#))

  ```
  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) ([Link](#))
  - 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** (Browser Extensible Data) [Link]

- 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**

- VCF: Variant Calling Format [Link]
- BCF: Binary version of VCF
Available NGS Analysis Toolsets

- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMtools
- SNP and INDEL analysis
- RNA-seq analysis
- Peak calling / ChIP-seq
- Many more.....
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
Use project's free server or other public servers

Get Galaxy
Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy
Screencasts, Galaxy 101, ...

Get Involved
Mailing lists, Tool Shed, wiki

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 Conceptual Framework

**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.
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 complex pipelines for processing sequencing analysis tasks on powerful server hardware from a graphical web browser-based user interface rather than the Linux command line. The connection to Galaxy is given in the left pane. Much more detailed descriptions of Galaxy's basic functionalities including user tutorials are available here.

Service?

Features 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) ease of use; (2) elimination of time consuming uploads of large data sets; (3) support for analyzing much larger data sets than is possible on public Galaxy server infrastructure. Users of the UCR Galaxy server have access to an annual registration fee. Users with an active BioCluster account can access this Galaxy service using their existing user name and password without any extra cost. More information on the BioCluster project is available here.

The center column is where the menus and data will appear. It shows you the history of analysis steps, allow to view data and results, and more.

User contains links to the downloading, preparation and analysis tools.
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

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.
Getting Data

Welcome to IIIGB’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 applications 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 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 IIIGB’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 IIIGB’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 IIIGB’s Galaxy Service
To enter this service, click here.
Import data from UCSC genome browser

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 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.
Import data from UCSC genome browser

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered 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. Send data to GenomeSpace for use with diverse computational tools. 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.

clade: Mammal

group: Genes and Gene Predictions

region: genome

clade: Mammal

output format: BED - browser extensible data

Send output to

Galaxy

GREAT

GenomeSpace

file type returned: plain text gzip compressed

To reset all user cart settings (including custom tracks), click here.
Send query to Galaxy from UCSC genome browser

Output knownGene as BED

Include custom track header:
- name=
- description=
- visibility=
- url=

Create one BED record per:
- Whole Gene
- Upstream by 200 bases
- Exons plus 0 bases at each end
- Introns plus 0 bases at each end
- 5' UTR Exons
- Coding Exons
- 3' UTR Exons
- Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Send query to Galaxy
Cancel

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.fastsanger
NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- Getting Data in Galaxy
- **Analyzing Data in Galaxy**
  - Lift–Over
  - Text manipulation tools
  - Filter and Sort
  - Operate on Genomic Intervals
  - Quality Control
  - Mapping Data
- History and workflow
- Galaxy Exercises
Lift-Over: convert genome coordinates

The image shows a screenshot of the Lift-Over tool in Galaxy, a web-based platform for interactive, web-based scientific workflows. The tool allows converting genome coordinates between assemblies and genomes. It has options to specify the minimum ratio of bases that must remap, allow multiple output regions, and a button to execute the conversion.

The tool provides instructions on how to use it, including a section titled 'What it does', which explains that the tool is based on the LiftOver utility and Chain track from the UC Santa Cruz Genome Browser. It converts coordinates and annotations between assemblies and genomes, producing 2 files: one containing all the mapped coordinates and the other containing the unmapped coordinates, if any.

Example:
Converting the following hg16 intervals to hg18 intervals:
- chr1 85178 112199 40002185 0 +
- chr1 110458 112199 40007346 0 +
- chr1 112283 121122 40074528 0 +

will produce the following hg18 intervals:
- chr1 122991 100020 40002185 0 +
- chr1 158279 160820 40007346 0 +
Text manipulation tools

- **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
- **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

**Paste (version 1.0.0)**

**Paste:**
- `a 1`
- `a 2`
- `a 3`

**Delimit by:**
- **Tab**

**What it does**

This tool merges two datasets side by side. If the first (left) dataset contains column assignments such as chromosome, start, end and strand, these will be preserved. However, if you would like to change column assignments, click the pencil icon in the history item.

**Example**

First dataset:

- `a 1`
- `a 2`
- `a 3`

Second dataset:

- `20`
- `30`
- `40`

Pasting them together will produce:

- `a 1 20`
- `a 2 30`
- `a 3 40`
Filter and Sort on Galaxy

Extract features (version 1.0.0)

Select GFF data:

- [ ]

From:
- Column 1 / Sequence name *

Extract features:

- [ ]

Multi-select list - hold the appropriate key while clicking to select multiple columns.

Execute

What it does
This tool extracts selected features from GFF data.

Example
Selecting promoter from the following GFF data:

```
chr22 GeneA enhancer 10000000 10001000 500 + , T6A
chr22 GeneA promoter 10010000 10010100 900 + , T6A
chr22 GeneB promoter 19020000 19025000 400 - , T6B
chr22 GeneC CCDS5220 19030000 19050000 800 - , T6B
```

will produce the following output:

```
chr22 GeneA promoter 10010000 10010100 900 + , T6A
chr22 GeneB promoter 19020000 19025000 400 - , T6B
```
Operate on Genomic Intervals

- **Join, Subtract and Group**
  - Join, Subtract, and Group intervals of two datasets
  - Intersect the intervals of two datasets
  - Subtract the intervals of two datasets
  - Merge the overlapping intervals of a dataset
  - Concatenate two datasets into one dataset

- **Base Coverage**
  - Base Coverage of all intervals

- **Convert Features**
  - Coverage of a set of intervals on second set of intervals
  - Complement intervals of a dataset
  - Cluster the intervals of a dataset
  - Join the intervals of two datasets side-by-side
  - Get flanks returns flanking region/s for every gene
  - Fetch closest non-overlapping feature for every interval
  - Profile Annotations for a set of genomic intervals

- **Statistics**

- **Wavelet Analysis**

- **Graph/Display Data**

---

**Concatenate (version 1.0.1)**

**Concatenate:**
- First dataset
- Second dataset

**Both datasets are same filetype?:**
- If unchecked, Second dataset will be forced into format of First dataset

**TIP:** If your dataset does not appear in the pull-down menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

**Screeenshots!**
See Galaxy Interval Operation Screeenshots (right-click to open this link in another window).

**Syntax**
Both datasets are exactly the same filetype will preserve all extra fields in both files. Leaving this unchecked will force the second dataset to use the same column assignments for chrom, start, end, and strand, but will fill extra fields with a period('.'). In both cases, the output fields are truncated or padded with fields of periods to maintain a truly tabular output.

**Example**

![Graph showing concatenated intervals](image)
FASTA manipulation

This tool converts RNA FASTA files to DNA (and vice-versa).

In **RNA-to-DNA** mode, U's are changed into T's.
In **DNA-to-RNA** mode, T's are changed into U's.

**Example**

Input RNA FASTA file (from Sanger's mirBase):

```plaintext
>cel-let-7 MIMAT0000001 Caenorhabditis elegans let-7
UGAGAGAGAGAGAAUAAUAGU
>cel-lin-4 MIMAT0000002 Caenorhabditis elegans lin-4
UCUCUGAGAGCCAGUGUGA
>cel-miR-1 MIMAT0000003 Caenorhabditis elegans miR-1
UGGAUGUAAAGAGAUGU
```

Output DNA FASTA file (with RNA-to-DNA mode):

```plaintext
>cel-let-7 MIMAT0000001 Caenorhabditis elegans let-7
TGAGTGATAGTTCTGATA
>cel-lin-4 MIMAT0000002 Caenorhabditis elegans lin-4
TCCCTGAGACCTCCAGTTGA
>cel-miR-1 MIMAT0000003 Caenorhabditis elegans miR-1
TGGAATGAAAGAAGTATGTA
```
Analyzing Data: Next Generation Sequencing

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:
- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

FastQC is the best place to look for documentation - it’s very good. Some features of the galaxy wrapper you are using are described below.

This Galaxy Tool: You are using FastQC in Galaxy. This is easy because it has been packaged into a Galaxy tool by the Interplanetary Utilities Commission. It exposes the external package FastQC which is documented at FastQC. Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the Picard tools libraries for sam/bam processing.

The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by Jim Johnson.

Inputs and outputs:
- This wrapper will accept a Galaxy fastq, sam or bam as the input read file to check. It will also take an optional file containing a list of contaminants information, in the form of a tab-delimited file with 2 columns, name and sequence.
- FastQC produces a single HTML output file which is slightly adjusted so it looks good in Galaxy that contains all of the results. Including the following:
  - Basic Statistics
  - Per base sequence quality
  - Per sequence quality scores
  - Per base sequence content
  - Per base GC content
  - Per sequence GC content
  - Per base % content
  - Sequence Length Distribution
  - Sequence Duplication Levels
Analyzing Data: Next Generation Sequencing

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

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

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 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.
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 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 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.
NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- Getting Data in Galaxy
- Analyzing Data in Galaxy
  - Lift-Over
  - Text manipulation tools
  - Filter and Sort
  - Operate on Genomic Intervals
  - Quality Control
  - Mapping Data

- History and workflow
- Galaxy Exercises
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 (Integrative Genomics Viewer)
  - [http://www.broadinstitute.org/igv/](http://www.broadinstitute.org/igv/)