

Using Galaxy for NGS Analysis

Daniel Blankenberg

Postdoctoral Research Associate

The Galaxy Team

<http://UseGalaxy.org>

Overview

- NGS Data
- Galaxy tools for NGS Data
- Galaxy for Sequencing Facilities

Overview

- NGS Data
- Galaxy tools for NGS Data
- Galaxy for Sequencing Facilities

NGS Data

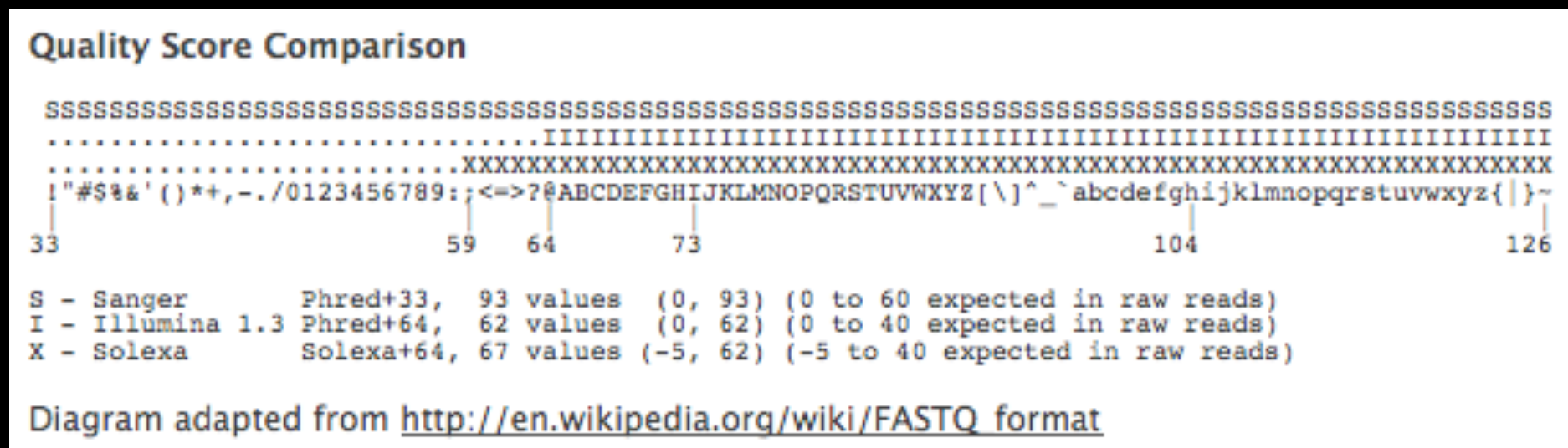
- Raw: Sequencing Reads (FASTQ)
- Derived
 - Alignments against reference genome (SAM/BAM)
 - Annotations
 - GFF
 - BED
 - Genome Assemblies

A Note on FASTQ

- Contains Sequence data and quality data

```
@UNIQUE_SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
! ' ' * ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) ) * * 55CCF>>>>>CCCCCCCC65
```

- Several Variants exist



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.

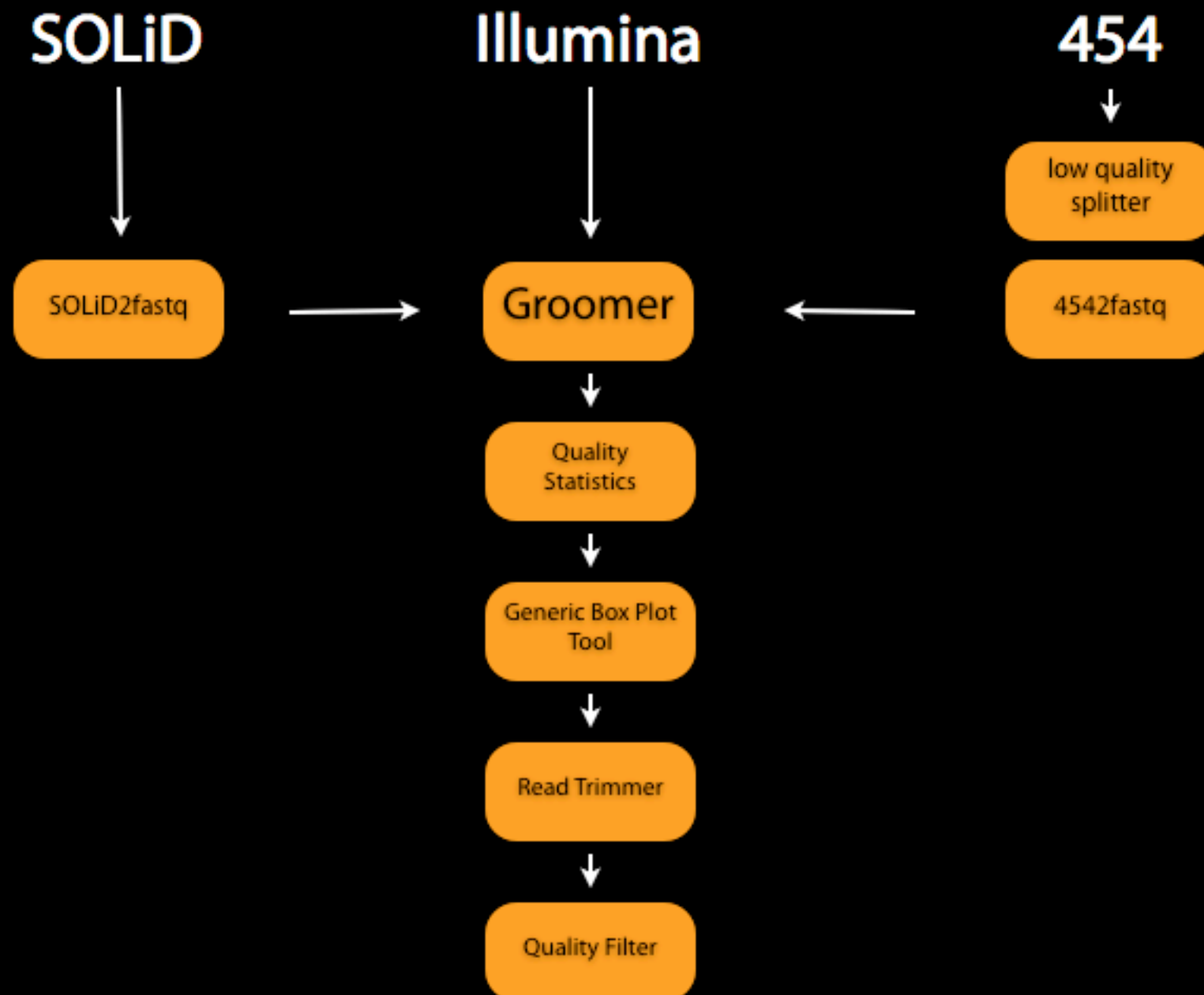
Overview

- NGS Data
- Galaxy tools for NGS Data
- Galaxy for Sequencing Facilities

Available NGS Analysis Toolsets

- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMTools
- SNP & INDEL analysis
- Peak Calling / ChIP-seq
- RNA-seq analysis

Prepare and Quality Check



Combining Sequences and Qualities

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User

Tools

- [FASTQ splitter on joined paired end reads](#)
- [FASTQ joiner on paired end reads](#)
- [FASTQ Summary Statistics by column](#)
- ROCHE-454 DATA**
- [Build base quality distribution](#)
- [Select high quality segments](#)
- [Combine FASTA and QUAL into FASTQ](#)
- AB-SOLID DATA**
- [Convert SOLiD output to fastq](#)
- [Compute quality statistics for SOLiD data](#)
- [Draw quality score boxplot for SOLiD data](#)
- GENERIC FASTQ MANIPULATION**
- [Filter FASTQ reads by quality score and length](#)
- [FASTQ Trimmer](#)
- [FASTQ Quality Trimming sliding window](#)
- [FASTQ Masker](#)

Combine FASTA and QUAL

FASTA File:
1: 454.fasta

Quality Score File:
2: 454.qual

Force Quality Score encoding:
ASCII

Execute

What it does

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 fastqsanger or fastqcassanger (when a csfasta is provided) with each quality score being the maximal allowed value (93).

Use this tool, for example, to convert 454-type output to FASTQ.

```
@EYKX4VC01B65GS length=54 xy=0784_1754 region=1 run=R_2007_11_07_16_15_57_
CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAACGAATTCGACTATGCCGAA
+
B8C:===A8C<?==@6<<=====B8-B9E<@6==B;B9<=====A8=C:
@EYKX4VC01BNCSP length=187 xy=0558_3831 region=1 run=R_2007_11_07_16_15_57_
CTTACCGGTCAACCACCGTGCCCTTCAGGATTGATCGCCAGATCGGTCGGTGCGTCAGGCGGGGTGACATCGCCCACCACGGTACTCACTGGCTGGCTCTGGTTCCCCGGCGGCATCGGAG
+
<D;:F=F=:<E=<E<====<E?4<=E=8E<<====<F><;<99E<;=E=9:6=9=;C::LE7*84====;=HA-<E==;F==;====<=;E<<<E=<<==E=E-HA-D=;F>====F>=E
@EYKX4VC01CD9FT length=115 xy=0865_1719 region=1 run=R_2007_11_07_16_15_57_
GGGGGCTTTGGCCTGTGCTCCGGCACCTCGCAAGAGCTACAGCAGGCGCGGCTGGCGATCATCGGCGGCACGCGCGGCTATATGTGCGCCGGAACACACCACCCGCACCCAACGCG
+
D91*#<HB.E<E<====<=B8F====E<====E<====<=====F====F=;=E<=====F=D;:<====<E<D:A7=====C:E<C:=<====E<=D>'====F?)B9=<<<
@EYKX4VC01B8FW0 length=95 xy=0799_0514 region=1 run=R_2007_11_07_16_15_57_
TAAATTTCAAGGAATGCAATTCAGGGTCGTGTGTTTAGACTTCGGCTTTAGAGACCTGAATACGTCAAAAACATAACTTCATGATATCTTGCAGT
+
=IC0D='<B8C9A7===JC2===F?*====<=F?)==<=D;<D;=F?*====C:==A7;=====LE8-'=6=<1=A8<====<====A7=;;<=
@EYKX4VC01BCGYW length=115 xy=0434_3926 region=1 run=R_2007_11_07_16_15_57_
GGCCAGCCGGGACAGCGTTGTTGGGCTGCATGGCGACGAGCTAAAAGTCGCCATCACC GCCCGCGCGGTTGATGGGCAGGCTAATGCCCATCTGGTAAAAACTTTCTCGCCAAAC
+
=';0<=F=JD2=6=86<E<9E=IC/7:=9<=F=;=<<====<LE7)=;=<;/:5=C9:IB3"4<1E=E=6<;JC17=F>;;D<=;JC1==<=F>;LE8-',HA-=25==2E>(9
@EYKX4VC01AZXC6 length=116 xy=0292_0280 region=1 run=R_2007_11_07_16_15_57_
GGGGGCGTTTGGCCTGTGCTCCGGCACCTCGCAAGAGCTACAGCAGGCGCGGCTGGCGATCATCGGCGGCACGCGCGGCTATATGTGCGCCGGAACACACCACCCGCACCCAACGCG
+
```

History

Combine QUAL and Sequence

2: 454.qual
52 lines
format: qual454, database: ?
Info: uploaded qual454 file

1: 454.fasta
18 sequences
format: fasta, database: ?
Info: uploaded fasta file

Grooming --> Sanger

[illegible]

Quality Statistics and Box Plot Tool

NGS TOOLBOX BETA

NGS: QC and manipulation

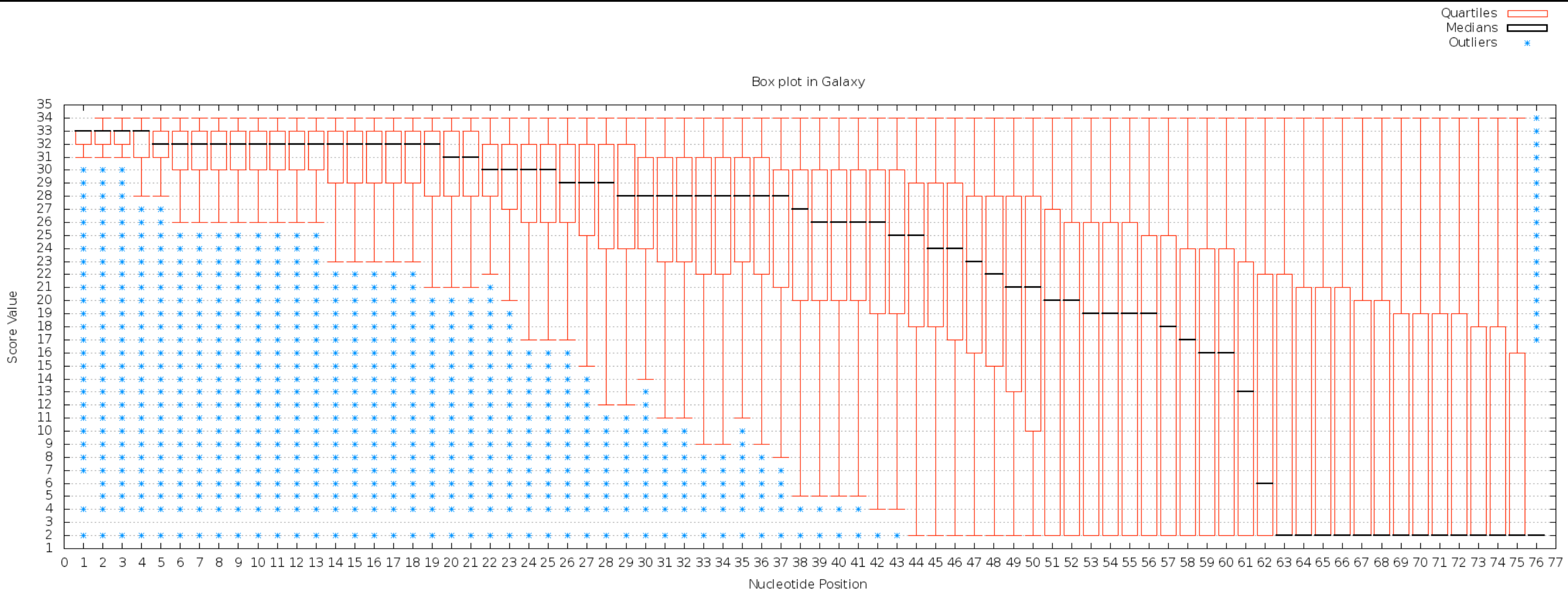
ILLUMINA DATA

- [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



Read Trimming

Tools

Options ▾

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](#)
- 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](#)

FASTQ Trimmer

FASTQ File:

2: imported: GM12878..ple Dataset ▾

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:

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

This tool allows you to trim the ends of reads.

You can specify either absolute or percent-based offsets to be trimmed. When using the percent-based method, offsets are given as a percentage of the read length.

For example, if you have a read of length 36:

```
@Some FASTQ Sanger Read  
CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA  
+  
=@@.0;B-378>CBA@>7@7BBCA4-48%<;%<B@
```

And you set absolute offsets of 2 and 0:

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

Quality Filtering

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:

0

Maximum Size:

0

A maximum size less than 1 indicates no limit.

Minimum Quality:

0.0

Maximum Quality:

0.0

A maximum quality less than 1 indicates no limit.

Maximum number of bases allowed outside of quality range:

0

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

Quality Filter on a Range of Bases 1

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:

0

Values start at 0, increasing from the left

Offset from 3' end:

0

Values start at 0, increasing from the right

Aggregate read score for specified range:

min score

Keep read when aggregate score is:

>=

Quality Score:

0.0

Remove Quality Filter on a Range of Bases 1

Add new Quality Filter on a Range of Bases

Execute

Manipulate FASTQ

Manipulate FASTQ

FASTQ File:

7: FASTQ Trimmer on data 2

Requires groomed data: if your data does not appear here try using the FASTQ groomer.

Match Reads

Add new Match Reads

Manipulate Reads

Add new Manipulate Reads

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 groomer.

Match Reads

Match Reads 1

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

Manipulate FASTQ

FASTQ File:

7: FASTQ Trimmer on data 2

Requires groomed data: if your data does not appear here try using the FASTQ groomer.

Match Reads

Match Reads 1

Match Reads by:

Sequence Content

Sequence Match Type:

Regular Expression

Match by:

N

Remove Match Reads 1

Add new Match Reads

Manipulate Reads

Manipulate Reads 1

Manipulate Reads on:

Miscellaneous Actions

Miscellaneous Manipulation Type:

Remove Read

Remove Manipulate Reads 1

Add new Manipulate Reads

Execute

- Remove reads with N's?

Available NGS Analysis Toolsets

- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMTools
- SNP & INDEL analysis
- Peak Calling / ChIP-seq
- RNA-seq analysis

Mapping NGS Data

- Collection of **interchangeable** mappers
- Accept FASTQ Format
- Create SAM/BAM Format
 - SAMTools*
- Algorithms for
 - DNA
 - RNA
 - Local Re-alignment

*Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

Mappers

- Short Reads
 - Bowtie
 - BWA
 - BFAST
 - PerM
- Longer Reads
 - LASTZ
- Metagenomics
 - Megablast
- RNA
 - Tophat

Lastz

Align sequencing reads in:

Against reference sequences that are:

Using reference genome:

If your genome of interest is not listed, contact the Galaxy team

Output format:

Lastz settings to use:

For most mapping needs use Commonly used settings. If you want full control use Full List

Select mapping mode:

Roche-454 98% identity

Roche-454 95% identity

Roche-454 90% identity

Roche-454 85% identity

Roche-454 75% identity

Illumina 95% identity

Illumina 85% identity

Reference name?:

View this identity (%):

Do not report matches above this identity (%):

Do not report matches that cover less than this percentage of each read:

Convert lowercase bases to uppercase:

Align sequencing reads in:

53: FASTQ to FASTA on data 7

Against reference sequences that are:

locally cached

Using reference genome:

Aedes aegypti: AegL1

If your genome of interest is not listed, contact the Galaxy team

Output format:

SAM

Lastz settings to use:

Full Parameter List

Commonly used

use Commonly used settings. If you want full control use Full List

Full Parameter List

Which strand to search?:

Both

Select seeding settings:

Seed hits require a 19 bp word with matches in

allows you set word size and number of mismatches

Select transition settings:

Allow one transition in each seed hit

affects the number of allowed transition substitutions

Perform gap-free extension of seed hits to HSPs (high scoring segment pairs)?:

No

Perform chaining of HSPs?:

No

Gap opening penalty:

400

Gap extension penalty:

30

X-drop threshold:

910

Y-drop threshold:

9370

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

Involve entropy when filtering HSPs?:

No

Do you want to modify the reference name?:

No

- Variable Levels of Settings
- Default Best-Practices
- Fully customizable parameters

Do you want to modify the reference name?:

No

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

Convert lowercase bases to uppercase:

Yes

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, AB/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.

Available NGS Analysis Toolsets

- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- **SAMTools**
- **SNP & INDEL analysis**
- Peak Calling / ChIP-seq
- RNA-seq analysis

SNPs & INDELS

- SNPs from Pileup

- Generate

- Filter

INDELS

NGS: SAM Tools

- [Filter SAM](#) on bitwise flag values
- [Convert SAM](#) to interval
- [SAM-to-BAM](#) converts SAM format to BAM format
- [BAM-to-SAM](#) converts BAM format to SAM format
- [Merge BAM Files](#) merges BAM files together
- [Generate pileup](#) from BAM dataset
- [Filter pileup](#) on coverage and SNPs
- [Pileup-to-Interval](#) condenses pileup format into ranges of bases
- [flagstat](#) provides simple stats on BAM files

The screenshot shows the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, and User. The left sidebar lists various tools under categories like NGS: SAM Tools, NGS: QC and manipulation, NGS: Mapping, NGS: SAM Tools, NGS: Indel Analysis, NGS: Peak Calling, NGS: RNA Analysis, RGENETICS, and SNP/WGA: Data; Filters. The main panel displays the 'Indel Analysis' tool configuration. The 'Select sam file to analyze:' dropdown is set to '54: BAM-to-SAM on dat..nverted SAM'. The 'Frequency threshold:' is set to '0.015'. The 'Execute' button is visible. Below the configuration, the 'What it does' section explains the tool's purpose. At the bottom, a table shows the results of the analysis.

Indel Analysis

Select sam file to analyze:
54: BAM-to-SAM on dat..nverted SAM

Frequency threshold:
0.015
Cutoff

Execute

What it does

Given an input sam file, this tool provides analysis of the indels. It filters out matches that do not meet the frequency threshold. The way this frequency of occurrence is calculated is different for deletions and insertions. The CIGAR string's "M" can indicate an exact match or a mismatch. For SAM containing the following bits of information (assuming the reference "ACTGCTCGAT"):

CHROM	POS	CIGAR	SEQ
ref	3	2M1I3M	TACTTC
ref	1	2M1D3M	ACGCT
ref	4	4M2I3M	GTTCAAGAT
ref	2	2M2D3M	CTCCG
ref	1	3M1D4M	AACCTGG
ref	6	3M1I2M	TTCAT
ref	5	3M1I3M	CTCTGTT
ref	7	4M	CTAT
ref	5	5M	CGCTA
ref	3	2M1D2M	TGCC

The following totals would be calculated (this is an intermediate step and not output):

POS	BASE	NUMREADS	DELPROPCALC	DELPROP	INSPROPSTARTCALC	INSSTARTPROP	INSPROPENDCALC	INSENDPROP
1	A	2	2/2	1.00	---	---	---	---
2	A	1	1/3	0.33	---	---	---	---
	C	2	2/3	0.67	---	---	---	---
3	C	1	1/5	0.20	---	---	---	---
	T	3	3/5	0.60	---	---	---	---
	-	1	1/5	0.20	---	---	---	---
4	A	1	1/6	0.17	---	---	---	---

Available NGS Analysis Toolsets

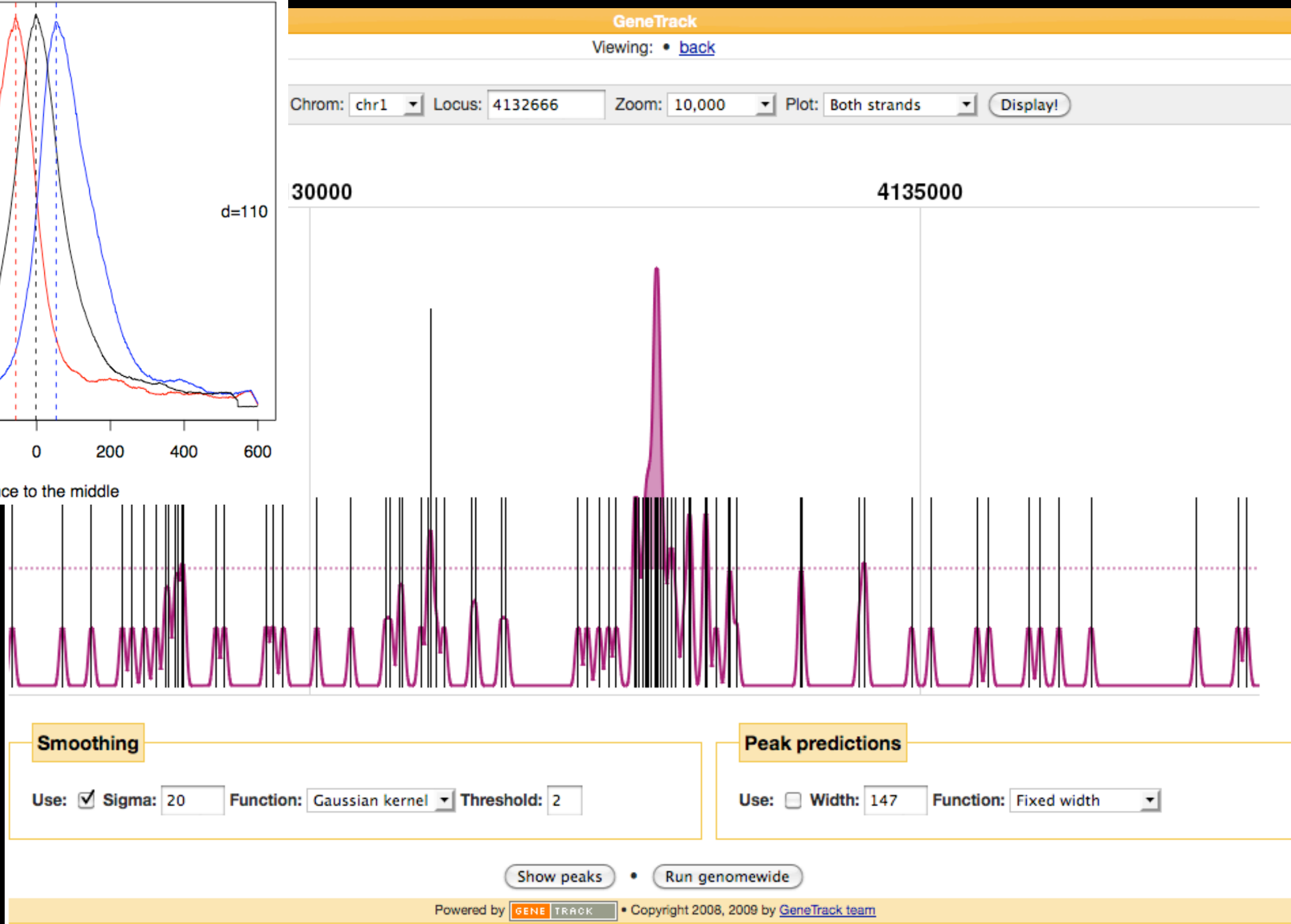
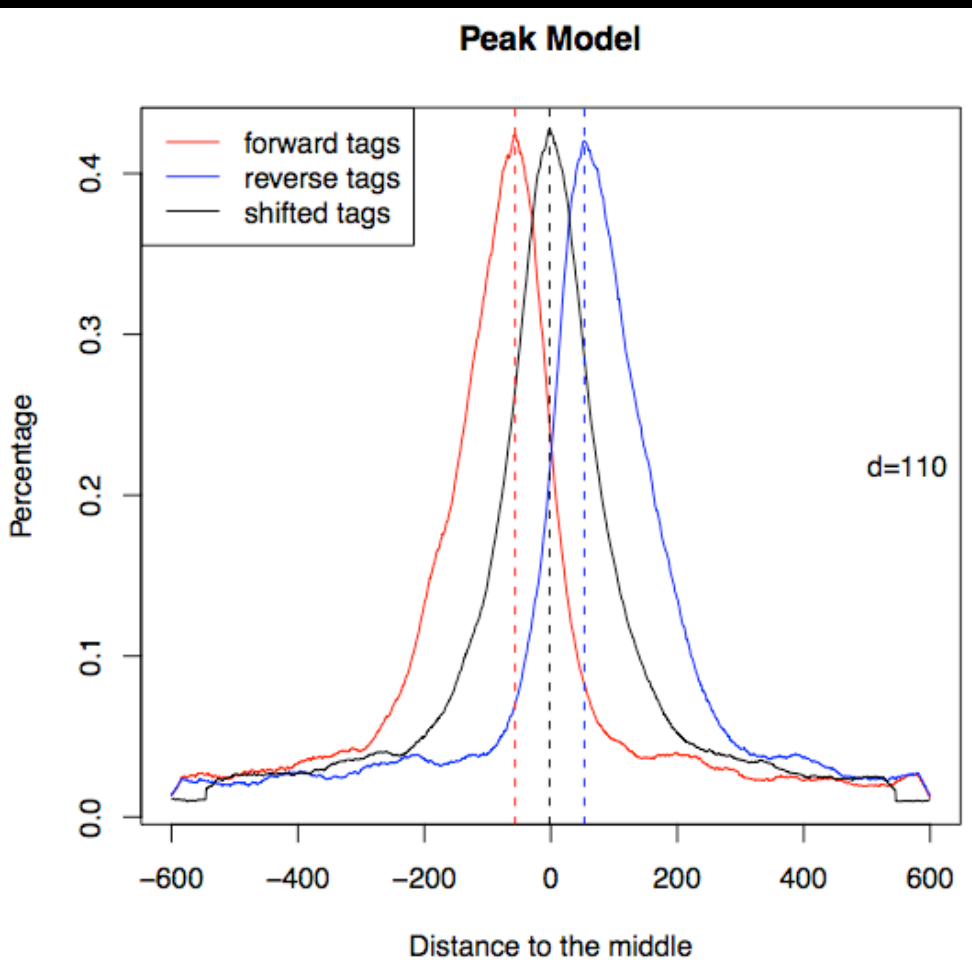
- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMTools
- SNP & INDEL analysis
- Peak Calling / ChIP-seq
- RNA-seq analysis

Peak Calling / ChIP-seq analysis

- Punctate Binding
 - Transcription Factors
- Diffuse Binding
 - Histone Modifications
 - PolII

Punctate Binding

GeneTrack



MACS

Zhang et al.
Model-based
Analysis of ChIP-
Seq (MACS).
Genome Biol
(2008) vol. 9 (9)
pp. R137

Albert I, Wachi S, Jiang C, Pugh BF. GeneTrack--a genomic data processing and visualization framework. Bioinformatics. 2008 May 15;24(10):1305-6. Epub 2008 Apr 3.

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)

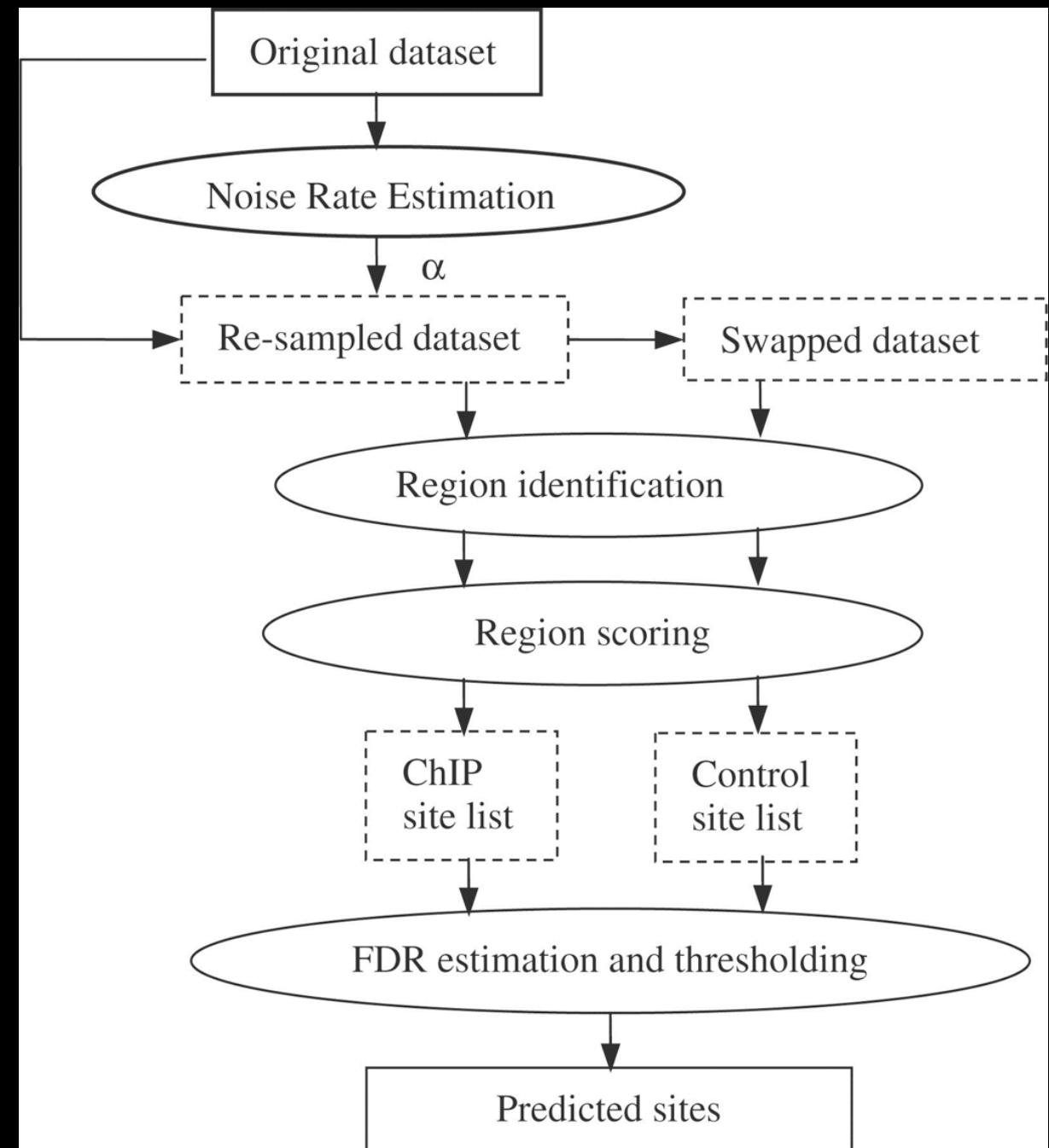
CCAT

ChIP-Seq Tag File:
5: Convert Genomic I..6 on data 3

ChIP-Seq Control File:
6: Convert Genomic I..6 on data 4

Advanced Options:
Hide Advanced Options

Select a pre-defined configuration file:
CCAT provided Histone Modification
CCAT provided Transcription Factor Binding
CCAT provided Histone Modification



ChIP-seq Exercise

<http://main.g2.bx.psu.edu/u/james/p/exercise-chip-seq>

ChIP-seq exercises

For this exercise we will use a ChIP-seq dataset for CTCF in the murine G1E_ER4 cell line. This dataset has been reduced to (mostly) contain only reads aligning to chr19:



Galaxy Dataset | G1E_ER4 CTCF (chr9)

A sample ChIP-seq dataset on CTCF in G1E_ER4 cells, reads have been reduced to those mapping to chr9 for demonstration use.



Click the 'import this dataset' button above to add this dataset to your analysis history to begin the analysis.

Mapping reads and peak calling

Step 1: First, for quality control, we will compute summary statistics on this dataset. Run the tool "NGS: QC and Manipulation > FASTQ Summary Statistics" on your dataset. When the job completes, inspect the results. How long are these reads? What is the median quality at the last position?

Step 2: Next we will map these reads to a reference genome. Use the "NGS: Mapping > Map with Bowtie for Illumina" tool. You will need to change the reference genome build you are mapping against to "mm9". Otherwise you can leave the default mapping options.

Step 3: Once reads are mapped, we will call peaks with MACS. Use the "NGS: Peak Calling > MACS" tool. You should also change the tag size to the read length you observed in Step 1. Otherwise the default values should be reasonable.

Step 4: Once MACS completes it will produce two datasets. One is a report on the peak calling process. The other contains the positions of the peaks. How many peaks were found? Click the link to "Display at UCSC main" and you will be able to see the positions of the peaks on the genome.

Calling peaks with a control sample

Next, we will incorporate an input DNA control, import the following dataset into your history:



Galaxy Dataset | G1E_ER4 input (chr19)
Reduced demo dataset, chr19 only



Step 1: Map the input DNA control to mm9 using Bowtie

Step 2: Load the MACS tool again. Select your previous CTCF dataset for ChIP-seq tag file, but now select the mapped input DNA for "ChIP-seq control file". How many peaks are called this time? What is the effect of using the input control?

Create a workflow and reuse

Step 1: At the top of the History panel, click "Options" and select "Extract Workflow". Here you have the chance to select which jobs will be included in the workflow. Click "Uncheck all" and then select the two "Map with Bowtie" jobs and the last "MACS" job.

Step 2: Import the following datasets -- CTCF ChIP and control for the G1E line:



Galaxy Dataset | G1E CTCF



Galaxy Dataset | G1E input



Step 3: At the bottom of the tools menu, select "Workflows > All Workflows", this will show the workflow list. Select the workflow you just created. You will be able to select input datasets for the two Bowtie steps, select the CTCF and input datasets. Click "Run Workflow".

Identify differential binding sites

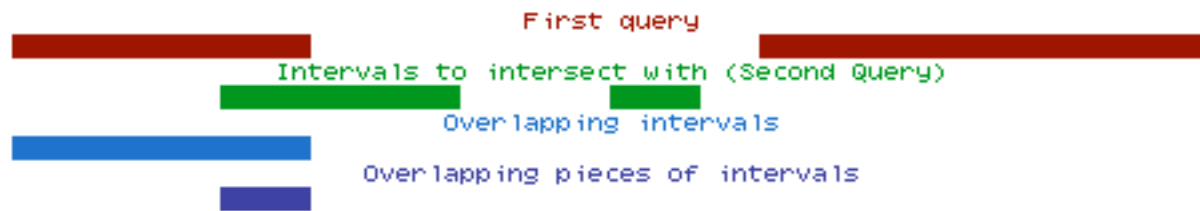
G1E is a model for erythropoiesis, the G1E line is a GATA1 null derived line which can be induced to differentiate by estradiol treatment (thus G1E-ER4). Here we will use Galaxy to identify sites that have differential binding across the two developmental stages.

I have Peaks, now what?

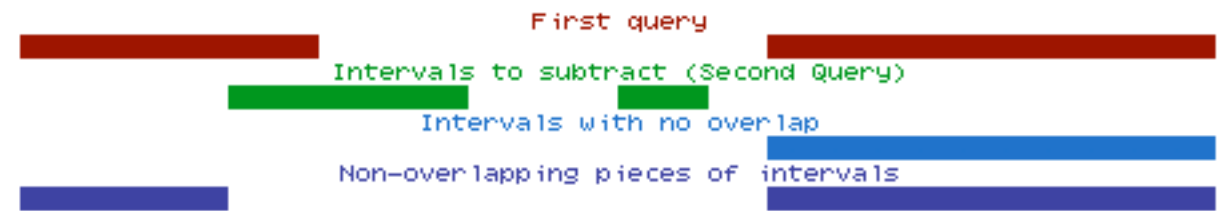
- Visualize
- Share
- Compare to existing Annotations
 - Interval Operations
 - Annotation Profiler

Genomic Interval Operations

A Intersect



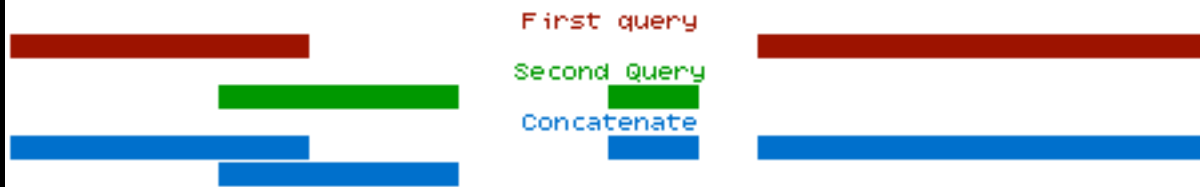
B Subtract



C Merge



D Concatenate



E Complement




F Cluster



Secondary Analysis

- 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

 **Galaxy**

Analyze DataWorkflowShared DataAdminHelpUser

ToolsOptions

[Get Data](#)
[Send Data](#)
[ENCODE Tools](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)

- Join two Queries side by side on a specified field
- Compare two Queries to find common or distinct rows
- Subtract Whole Query from another query
- Group data by a column and perform aggregate operation on other columns.
- Column Join

[Convert Formats](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Get Genomic Scores](#)
[Operate on Genomic Intervals](#)
[Statistics](#)
[Wavelet Analysis](#)
[Graph/Display Data](#)
[Regional Variation](#)
[Multiple regression](#)
[Multivariate Analysis](#)
[Evolution](#)

3 UTR	803	
5 UTR	574	
coding exons		2743
introns	13746	
other	12499	

HistoryOptions

2: MACS peak calls (broadPeak)

21,728 regions, format: interval, database: mm9

Info:

| display at UCSC [main test](#) | view in [GeneTrack](#) | display at Ensembl [Current](#)

1.Chrom	2.Start	3.End	4	5	6	7	8	9
chr1	4132666	4133002	.	0	.	16.04	14.366	0.
chr1	4322446	4323079	.	0	.	27.07	26.185	0.
chr1	4336241	4336651	.	0	.	23.06	18.736	0.
chr1	4406740	4407268	.	0	.	16.20	23.794	0.
chr1	4506655	4507162	.	0	.	20.30	21.868	0.
chr1	4758431	4758873	.	0	.	24.01	30.691	0.

1: UCSC Main on Mouse: refGene (genome)

28,108 regions, format: bed, database: mm9

Info: UCSC Main on Mouse: refGene (genome)

| display at UCSC [main test](#) | view in [GeneTrack](#) | display at Ensembl [Current](#)

1.Chrom	2.Start	3.End	4.Name	5	6
chr1	134212701	134230065	NM_028778	0	+
chr1	134212701	134230065	NM_001195025	0	+
chr1	33510655	33726603	NM_008922	0	-
chr1	58714963	58752833	NM_175370	0	-
chr1	25124320	25886552	NM_175642	0	-

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

Profile Annotations

Choose Intervals:
34: UCSC Main on Mous..na (genome) ▾

Keep Region/Table Pairs with 0 Coverage:
Discard ▾

Output per Region/Summary:
Per Region ▾

Choose Tables to Use:

- [+] ☒ Comparative Genomics
- [+] ☐ Genes and Gene Prediction Tracks
- [+] ☐ Mapping and Sequencing Tracks
- [+] ☐ Phenotype and Allele
- [+] ☐ Expression and Regulation
- [+] ☐ mRNA and EST Tracks
- [-] ☐ Variation and Repeats
 - ☒ Microsatellite
 - ☐ Simple Repeats
 - ☒ SNPs (128)
- [+] ☐ Uncategorized Tables

Selecting no tables will result in using all tables.

Execute

Available NGS Analysis Toolsets

- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMTools
- SNP & INDEL analysis
- Peak Calling / ChIP-seq
- RNA-seq analysis

RNA-seq

- TopHat
- Cufflinks
- Cuffcompare
- Cuffdiff

NGS: RNA Analysis

RNA-SEQ

- Tophat Find splice junctions using RNA-seq data
- 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
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

TopHat

- Map RNA (FASTQ) to a reference Genome
 - Uses Bowtie
 - BAM file of accepted hits
- Find Splice Junctions
 - File with two connected BED blocks

The screenshot shows the TopHat tool interface in a Galaxy environment. It includes several dropdown menus and buttons for configuring the analysis. The 'Will you select a reference genome from your history or use a built-in index?' dropdown is set to 'Use a built-in index'. Below it, a note states 'Built-ins were indexed using default options'. The 'Select a reference genome:' dropdown is set to 'Human (Homo sapiens): hg18 Canonical', with a note below it: 'If your genome of interest is not listed, contact the Galaxy team'. The 'Is this library mate-paired?:' dropdown is set to 'Single-end'. The 'RNA-Seq FASTQ file:' dropdown is set to '1: imported: h1-hESC..ple Dataset', with a note below it: 'Must have Sanger-scaled quality values with ASCII offset 33'. The 'TopHat settings to use:' dropdown is set to 'Use Defaults', with a note below it: 'You can use the default settings or set custom values for any of Tophat's parameters.' At the bottom, there is an 'Execute' button.

TopHat

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Built-ins were indexed using default options

Select a reference genome:
Human (Homo sapiens): hg18 Canonical
If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Single-end

RNA-Seq FASTQ file:
1: imported: h1-hESC..ple Dataset
Must have Sanger-scaled quality values with ASCII offset 33

TopHat settings to use:
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.

Execute

Cufflinks

- Input: aligned RNA-Seq reads (SAM/BAM; e.g. from TopHat)
- assembles transcripts
- estimates relative abundance
- tests for differential expression and regulation
- Outputs
 - GTF: Assembled Transcripts
 - Tabular, with coordinates and expression levels
 - Transcripts
 - Genes


Cuffcompare

- Compare assembled transcripts to a reference annotation (2+ GTF files)
- Track Cufflinks transcripts across multiple experiments (e.g. across a time course)
- Outputs:
 - Transcripts Accuracy File
 - "accuracy" of the transcripts in each sample when compared to the reference annotation data
 - Transcripts Combined File
 - union of all transfrags in each sample
 - Transcripts Tracking Files
 - matches transcript structure that is present in one or more input GTF files

Cuffdiff

- Inputs
 - GTF from Cufflinks, Cuffcompare, other source
 - 2 SAM files from 2+ samples
- Changes in
 - transcript expression
 - splicing
 - promoter use

RNA-seq Tutorial

 Galaxy





Analyze DataWorkflowShared DataVisualizationAdminHelpUser

Published Pages | [jeremy](#) | Galaxy RNA-seq Analysis Exercise





RNA-seq Analysis Exercise

Galaxy provides multiple tools for performing RNA-seq analysis. This exercise introduces these tools and guides use of these tools on some example datasets; prominent RNA-seq tools include Tophat and Cufflinks. Familiarity with Galaxy and the general concepts of RNA-seq analysis are useful for understanding this exercise. This exercise should take 1-2 hours.

Below are small samples of datasets from the [ENCODE Caltech RNA-seq track](#); specifically, the datasets are single 75bp reads from the h1-hESC and GM12878 cell lines. The sampled reads map mostly to chr19. Import the datasets to your history by clicking on the green-plus icon labeled 'Import'.

 [Galaxy Dataset | h1-hESC Sample Dataset](#)   
First 100,000 chr19 reads for ENCODE h1-hESC cell line, Caltech RNA-seq track, 1x75bp.

and

 [Galaxy Dataset | GM12878 Sample Dataset](#)   
First 100,000 chr19 reads for ENCODE GM12878 cell line, Caltech RNA-seq track, 1x75bp.

Understanding and preprocessing the reads

You should understand the reads a bit before analyzing them. Preprocessing may be needed as well.

Step 1: Compute statistics and create a boxplot of base pair quality scores for each set of reads using the [\[NGS: QC and manipulation >\] FASTQ Summary Statistics](#) tool and then plot the output using the [\[Graph/Display Data >\] Boxplot](#). Often, it is useful to trim reads to remove base positions that have a low median (or bottom quartile) score. For this exercise, assume a median quality score of below 15 to be unusable. Given this criterion, is trimming needed for the datasets? If so, which base pairs should be trimmed?

Step 2: If necessary, trim the reads based on your answers to step A using [\[NGS: QC and manipulation >\] FASTQ Trimmer](#)

Map processed reads

The next step is mapping the processed reads to the genome. The major challenge when mapping RNA-seq reads is that the reads, because they come from RNA, often


Overview

- NGS Data
- Galaxy tools for NGS Data
- Galaxy for Sequencing Facilities

Sample Tracking System

- Built-in system for tracking sequencing requests
- Customizable interfaces
 - Sequencing Facility Managers/Administrators
 - Customers/Users/Biologists
- Streamlines delivery of data from sequencing runs to customers

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
 - Transfer the datasets from the sequencer to Galaxy at the end of the sequence run.
- 
- Can be automated

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

The screenshot displays the Galaxy web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Lab', 'Admin', 'Help', and 'User'. The left sidebar contains a menu with categories: 'Administration' (Security: Manage users, Manage groups, Manage roles; Data: Manage data libraries; Server: Reload a tool's configuration, Profile memory usage, Manage jobs), 'Form Definitions' (Manage form definitions), and 'Sample Tracking' (Manage sequencers and external services, Manage request types, Sequencing requests, Find samples). The main content area is titled 'Forms' and features a 'Create new form' button, a search bar with the text 'search', and a link to 'Advanced Search'. Below this is a table listing forms:

<input type="checkbox"/>	Name	Description	Type
<input type="checkbox"/>	Analysis Portal run details ▼		Sample run details template
<input type="checkbox"/>	Atlantic Biosciences Analysis Portal Form ▼		External Service Information Form
<input type="checkbox"/>	Atlantic Biosciences request ▼		Sequencing Request Form
<input type="checkbox"/>	Atlantic Biosciences sample ▼		Sequencing Sample Form

At the bottom of the table area, it says 'For 0 selected forms:' followed by 'Delete' and 'Undelete' buttons.

- Uses Galaxy forms, do this once

[View](#)

Edit form definition "Atlantic Biosciences request" (Sequencing Request Form)

Name

Description

Form definition fields

1. Name (TextField)

2. Scientific Contact (AddressField)

[Add field](#)

[Save](#)

- Configurations can be
 - custom-built
 - loaded from provided configuration files

[Edit](#)

Form definition "Atlantic Biosciences sample" (Sequencing Sample Form)

Layout1

Run type	Read length	Number of Lanes	Alignment target	Processing time	Comments
SelectField: <input style="width: 50px;" type="text"/> - (optional) Options: SR PE	SelectField: <input style="width: 50px;" type="text"/> - (optional) Options: 36 50 75 100	TextField: <input style="width: 100px;" type="text"/> - (optional)	TextField: <input style="width: 100px;" type="text"/> - (optional)	SelectField: <input style="width: 50px;" type="text"/> - (optional) Options: Std Rush option3	TextField: <input style="width: 100px;" type="text"/> - (optional)

Configure the Sequencer

Galaxy Analyze Data Workflow Shared Data Lab Admin Help User

Administration

- Security
 - Manage users
 - Manage groups
 - Manage roles
- Data
 - Manage data libraries
- Server
 - Reload a tool's configuration
 - Profile memory usage
 - Manage jobs
- Form Definitions
 - Manage form definitions
- Sample Tracking
 - Manage sequencers and external services
 - Manage request types
 - Sequencing requests
 - Find samples

External Services Reload external service types Create new external service

search Advanced Search

<input type="checkbox"/>	Name	Description	External Service Type	Last Updated
<input type="checkbox"/>	Analysis Portal service ▼		Atlantic Biosciences Analysis Portal	3 minutes ago

For 0 selected externalservices: Delete Undelete

Edit external service

Name: Analysis Portal service

Description:

Version: 1

Hostname or IP address: 192.168.56.101 (Required)

User name: administrator (Required)

Password: (Required)

Data directory:

- Available Services are now configured
- Ready for customers

How does it all work?

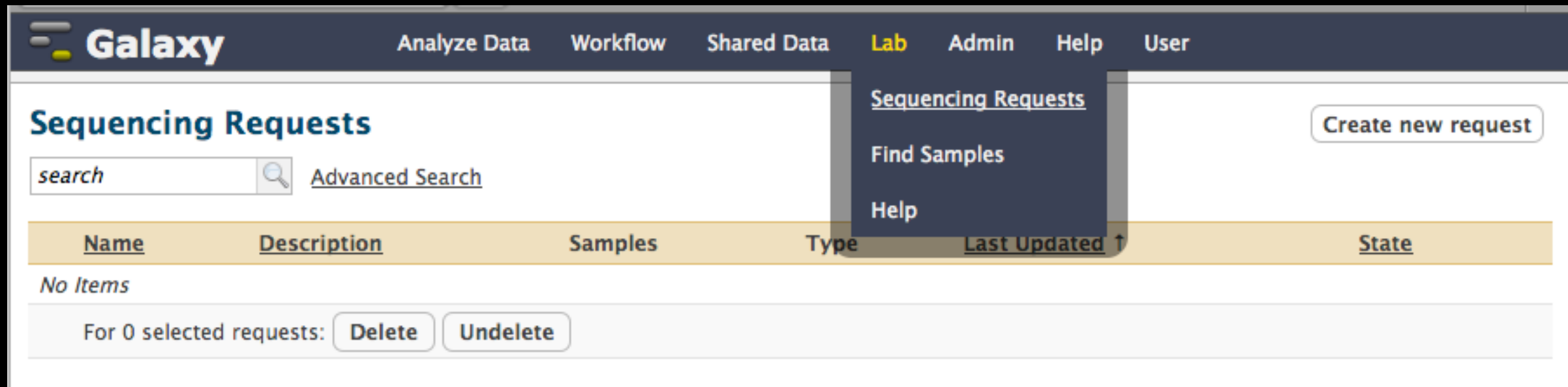
customer



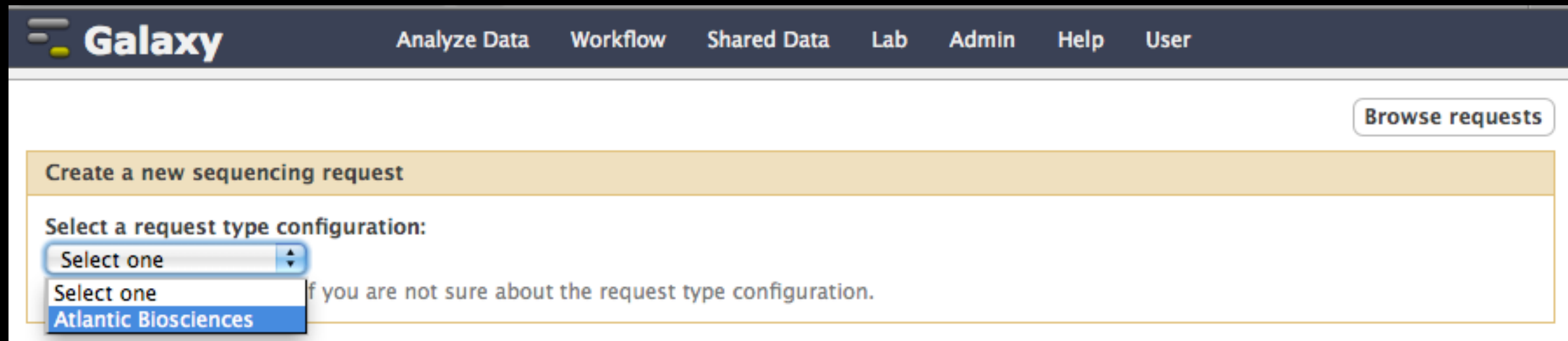
facility manager



Customer Creates a Request



The screenshot shows the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Lab (highlighted), Admin, Help, and User. The 'Lab' dropdown menu is open, showing options for Sequencing Requests, Find Samples, and Help. The main content area is titled 'Sequencing Requests' and features a search bar with the placeholder text 'search' and a magnifying glass icon, followed by a link to 'Advanced Search'. A 'Create new request' button is located in the top right corner. Below the search bar is a table with columns: Name, Description, Samples, Type, Last Updated (with an upward arrow), and State. The table currently displays 'No Items'. At the bottom, there is a message 'For 0 selected requests:' followed by 'Delete' and 'Undelete' buttons.



The screenshot shows the 'Create a new sequencing request' form in the Galaxy interface. The top navigation bar is the same as the previous screenshot. A 'Browse requests' button is in the top right corner. The form has a title 'Create a new sequencing request' and a section titled 'Select a request type configuration:'. This section contains a dropdown menu with 'Select one' selected. Below the dropdown, there is a link that says 'Select one if you are not sure about the request type configuration.' and a button labeled 'Atlantic Biosciences'.

Customer Describes Request

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

(Required)

Description

This is Experiment was performed using the protoc

(Optional)

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

Customer Adds a Sample

 **Galaxy**

Analyze Data Workflow Shared Data Lab Admin Help User

Request Actions ▾

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

Name	State	Data Library	Folder	History	Workflow
<input type="text" value="Sample_1"/> <small>(required)</small>		<input type="text" value="Dan's Sequencing Requests"/>	<input type="text" value="ChIP-seq"/>	<input type="text" value="My own ChIP-seq Experiment!"/>	<input type="text" value="Dan's ChIP-seq Workflow"/>

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 samples from sample

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

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**

- Provide a name for the sample
- Data Library and Folder
- History
- Workflow to run

Samples Added, Submit Request

[Analyze Data](#)[Workflow](#)[Shared Data](#)[Lab](#)[Admin](#)[Help](#)[User](#)[Edit samples](#)[Submit request](#)[Request Actions ▾](#)

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

Name	State	Data Library	Folder	History	Workflow
Sample_1	Unsubmitted	Dan's Sequencing Requests	ChIP-seq	My own ChIP-seq Experiment!	Dan's ChIP-seq Workflow

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 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




The page at <http://localhost:8080> says:

More samples cannot be added to this request after it is submitted. Click OK to submit.

[Cancel](#)[OK](#)

Samples enter “New” state

 **Galaxy**

Analyze Data Workflow Shared Data Lab Admin Help User

Request Actions ▾

✓ The sequencing request has been submitted.

Sequencing request "My first ChIP-seq Experiment"

Current state:
[In Progress](#)

Description:
This is Experiment was performed using the protocol ...

User:
dan@bx.psu.edu

Request type:
Atlantic Biosciences

▶ More

Samples Edit samples ▾

Name	Barcode	State	Data Library	Folder	History	Workflow	Run Datasets
Sample_1		New	Dan's Sequencing Requests	ChIP-seq	My own ChIP-seq Experiment!	Dan's ChIP-seq Workflow	0

▶ Layout1

- Customer sends samples to sequencing facility

Sequencing Facility is informed of Request

Galaxy

Analyze Data

Workflow

Shared Data

Lab

Admin

Help

User

Administration

Security

- Manage users
- Manage groups
- Manage roles

Data

- Manage data libraries

Server

- Reload a tool's configuration
- Profile memory usage
- Manage jobs

Form Definitions

- Manage form definitions

Sample Tracking

- Manage sequencers and external services
- Manage request types
- Sequencing requests
- Find samples

Sequencing Requests

Create new request

search

Advanced Search

<input type="checkbox"/>	Name	Description	Samples	Type	Last Updated ↑	State	User
<input type="checkbox"/>	<u>My first ChIP-seq Experiment</u>	▼ This is Experiment was performed using the protocol ...	<u>1</u>	<u>Atlantic Biosciences</u>	26 minutes ago	In Progress	dan@bx.psu.edu
<input type="checkbox"/>	<u>new request</u>	▼	<u>1</u>	<u>Atlantic Biosciences</u>	3 days ago	Complete	customer@corp.com
<input type="checkbox"/>	<u>some experiment test</u>	▼ a test description	<u>1</u>	<u>Atlantic Biosciences</u>	3 days ago	Complete	customer@corp.com

For 0 selected requests:

Delete

Undelete

Sequencing Facility Receives Samples

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

Edit Current Samples of Sequencing Request "My first ChIP-seq Experiment"

Name	Barcode	State	Data Library	Folder	History	Workflow	Run Datasets	Delete
<input type="checkbox"/> Sample_1 <small>(required)</small>	<input type="text"/>	New	Dan's Sequencing Requests	ChIP-seq	My own ChIP-seq Experiment!	Dan's ChIP-seq Workflow	0	

For selected samples: Select one

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](#)

Save Cancel

Click the **Save** button when you have finished editing the samples

- Customer can watch progress of sequencing request

Sequencing Requests

<input type="checkbox"/>	Name	Description	Samples	Type	Last Updated ↑	State
<input type="checkbox"/>	My first ChIP-seq Experiment ▼	This is Experiment was performed using the protocol ...	1	Atlantic Biosciences	35 minutes ago	Complete

For 0 selected requests:

Sequencing Finished

- Datasets are transferred from sequencer into Galaxy
 - Library
 - User's history
- Galaxy Workflow is executed on Dataset
- Customer 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>
- <http://bcbio.wordpress.com/2011/01/11/next-generation-sequencing-information-management-and-analysis-system-for-galaxy/>

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

Acknowledgments

- All Members of the Galaxy Team (see them at <https://bitbucket.org/galaxy/galaxy-central/wiki/GalaxyTeam>)
- Thousands of our users
- GMOD Team
- UCSC Genome Informatics Team
- BioMart Team
- FlyMine/InterMine Teams
- Funding sources
 - NSF-ABI
 - NIH-NHGRI
 - Beckman Foundation
 - Huck Institutes at Penn State
 - Pennsylvania Department of Public Health
 - Emory University

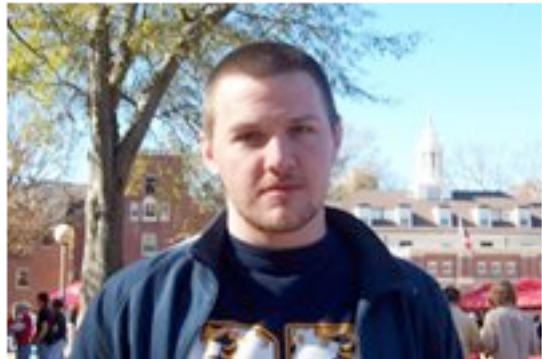
Galaxy Team



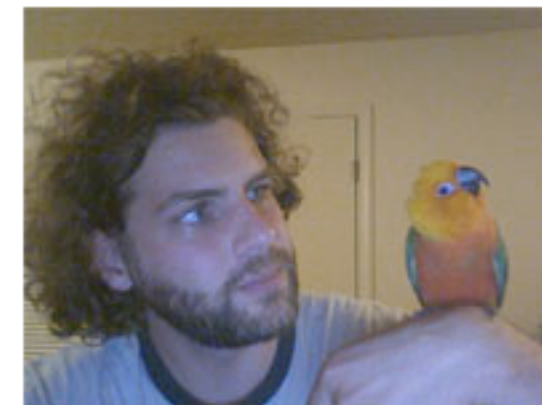
Enis Afgan | Emory



Guru Ananda | Penn State



Dannon Baker | Emory



James Taylor | Emory



Jeremy Goecks | Emory



Sergei Kosakovsky Pond | UCSD



Greg von Kuster | Penn State



Dave Clements | Emory



Nate Coraor | Penn State



Ross Lazarus | Harvard | BakerID



Kanwei Li | Emory



Anton Nekrutenko | Penn State



Kelly Vincent | Penn State

+ Jennifer
Jackson

Two full days of presentations,
workshops, and conversations by
and for Galaxy community members



<http://galaxy.psu.edu/gcc2011>