

# Using Galaxy for High-throughput Sequencing (HTS) Analysis

The Galaxy Team  
<http://UseGalaxy.org>

# Overview

## High-throughput Sequencing (HTS) Data

### Using Galaxy to Analyze HTS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

### Galaxy for Sequencing Facilities

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

# HTS Data

From the Sequencer:

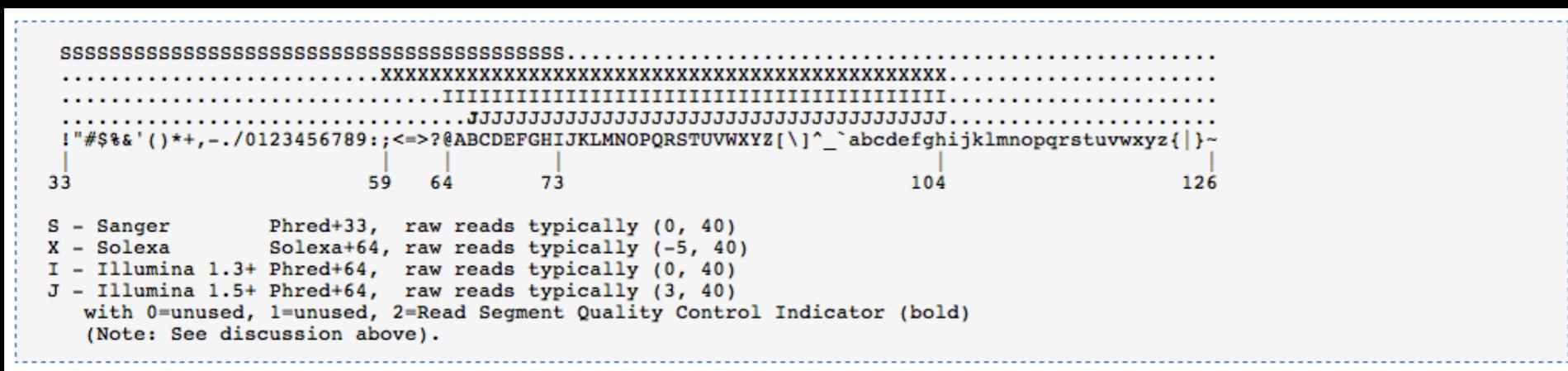
- reads and quality scores (FASTQ)

In the Analysis Pipeline / Workflow:

- alignments against reference genome (SAM, BAM)
- annotations (GFF, BED)
- genome Assemblies (FASTA)
- quantitative tracks, e.g. conservation (WIG)

# FASTQ Quality Scores

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



[http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format)

Galaxy tools generally use Sanger format

- ♦ Need to convert quality scores to Sanger using Groomer tool

# Getting Your Data into Galaxy

Cannot upload any file larger than 2GB via Web browser

- Galaxy does not currently support compressed files

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

# Overview

High-throughput Sequencing (HTS) Data

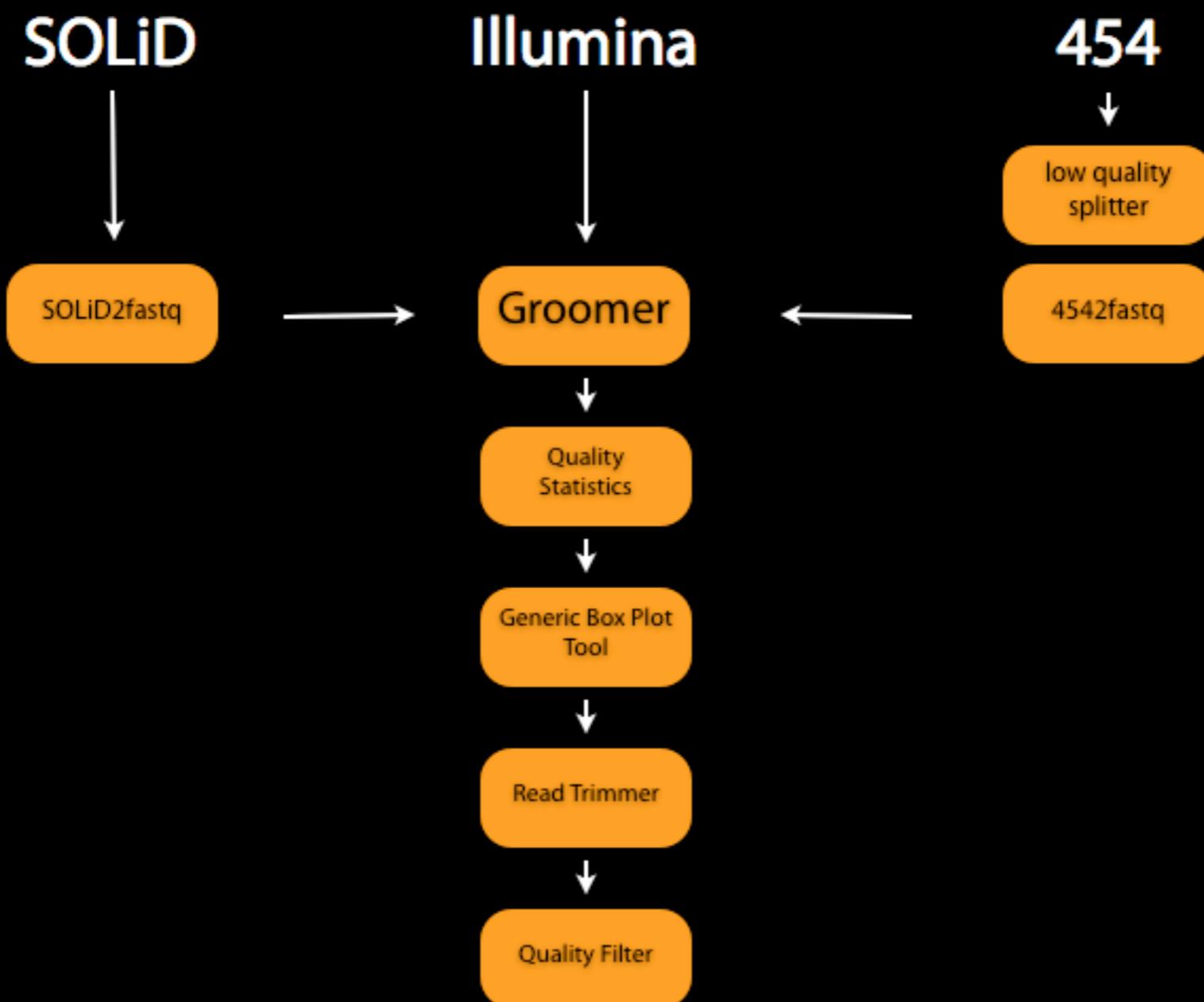
Using Galaxy to Analyze HTS Data

- ♦ Prepare, quality control and manipulate reads
- ♦ Read Mapping
- ♦ SNP & INDEL analysis
- ♦ Binding sites analysis and peak calling
- ♦ Transcriptome analysis

Galaxy for Sequencing Facilities

Galaxy exercises: ChIP-seq and RNA-seq

# Prepare and Quality Check



# Combining Sequences and Qualities

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools Options ▾

- 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 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 fastqcossanger (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_
CCGGTATCCGGGTGCCGTGATGACGCCACCGAACGAATTGACTATGCCAA
+
B8C:==A8C<%=<@6=<<=====B8=B9E<%=<B;B9<=====A8=C:
@EYKX4VC01BNCSP length=187 xy=0558_3831 region=1 run=R_2007_11_07_16_15_57_
CTTACCGGTACCACCGTCAGGATTGATGCCAGATCGTCGGTGCAGATGCCACCACGGTACTCACTGGCTGGCTCTGGTCCCGGGCATCGGAG
+
<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_
GGGGCCTTGGCTGCGTCCGGCACCTCGCAAGAGCTACAGCAGGCCGGCTGGCGATCATGGCGGACGCCGGCTATATGTCGCCGGAACACACCACCCGACCCAACCGC
+
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_
TAAATTTCAAGGAATGCAAATCAGGGTCGTGTAGACTCGGCTTAGAGACCTGAATACTCATGATATCTGCAGT
+
=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_
GGCCAGCCGGACAGCGTTGGCTGCATGGCGACGAGCTAAAGTCGCCATCACCGCCCCGGCTGATGGCAGGCTAATGCCATCTGGTAAAAACTTCTGCCAAAC
+
=';0<=F=JD2=6=86<E=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_
GGGGCGTTGGCTGCGTCCGGCACCTCGCAAGAGCTACAGCAGGCCGGCTGGCGATCATGGCGGACGCCGGCTATATGTCGCCGGAACACACCACCCGACCCAACCGC
+
D91*#<HB.E<E<=====B8F==E<=====E<=====F====F>;=E<=====F==D;<<<E<D:A7=====C:E<C:<=====E<D>'====F?)B9=<<
```

History Options ▾

Combine QUAL and Sequence

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

```
>EYKX4VC01B65GS length=54 xy=0784_1
33 23 34 25 28 28 32 23 34 27 4
>EYKX4VC01BNCSP length=187 xy=0558_
27 35 26 25 37 28 37 28 25 28 27 36
22 9 23 19 28 28 28 28 26 28 39 32
26 27 37 29 28 26 28 36 28 26 24 38
```

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

```
>EYKX4VC01B65GS length=54 xy=0784_1
CCGGTATCCGGGTGCCGTGATGACGCCACCGAACGAATTGACTATGCCAA
>EYKX4VC01BNCSP length=187 xy=0558_
CTTACCGGTACCACCGTCAGGATTGATGCCAGATCGTCGGTGCAGATGCCACCACGGTACTCACTGGCTGGCTCTGGTCCCGGGCATCGGAG
GGTGACATGCCACCACCGTACTCACTGGCTGGC
CACCACGTTGAGGGTATTCCCTCGGTTGGCTG
```

# Grooming --> Sanger

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools Options

NGS TOOLBOX BETA

NGS: QC and manipulation

- 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

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

**FASTQ Groomer**

File to groom:  
3: Combine FASTA and.. and data 2

Input FASTQ quality scores type:  
Sanger  
Solexa  
Illumina 1.3+  
**Sanger**  
Color Space Sanger  
Execute

What it does

This tool offers several conversions options relating to the FASTQ quality scores.

When using *Basic* options, the output will be *sanger* formatted (Sanger).

When converting, if a quality score falls outside of the target range, it will be converted to the minimum or maximum.

When converting between Solexa and the other formats, qual scores are converted using the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, Jansson P, Jonassen I, Lohman K, Mikkelsen T, Salzberg SL, Schatz MC, Stoeckert CJ, and Stoeckert CJ. Quality scores, and the Solexa/Illumina FASTQ variants. Nucleotide

When converting between color space (csSanger) and base space, adapter bases are lost or gained; if gained, the base 'G' is used as the adapter. You cannot convert a color space read to base space if there is no adapter present in the color space sequence. Any masked or ambiguous nucleotides in base space will be converted to 'N's when determining color space encoding.

**4: FASTQ Groomer on data 3**

18 sequences  
format: fastqsanger, database: ?  
Info: Groomed 18 sanger reads into sanger reads.  
Based upon quality and sequence, the input data is valid for: sanger  
Input ASCII range: '!'(33) – 'L'(76)  
Input decimal range: 0 – 43

@EYKX4VC01B65GS length=54 xy=0784\_1  
CCGGTATCCGGGTGCCGTGATGAGGCCACCGAA+  
B8C:==A8C<@==@6=<<=====B8=B9E<@6=

@EYKX4VC01BNCSP length=187 xy=0558\_1  
CTTACCGGTACCCACCGTGCCCTTCAGGATTGATCG+  
B8C:==A8C<@==@6=<<=====B8=B9E<@6=

**History Options**

Combine QUAL and Sequence

3: Combine FASTA and QUAL on data 1 and data 2

18 sequences  
format: fastqsanger, database: ?  
Info: Combined 18 of 18 sequences with quality scores (100.00%).

@EYKX4VC01B65GS length=54 xy=0784\_1  
CCGGTATCCGGGTGCCGTGATGAGGCCACCGAA+  
B8C:==A8C<@==@6=<<=====B8=B9E<@6=

@EYKX4VC01BNCSP length=187 xy=0558\_1  
CTTACCGGTACCCACCGTGCCCTTCAGGATTGATCG+  
B8C:==A8C<@==@6=<<=====B8=B9E<@6=

2: 454.qual

52 lines  
format: qual454, database: ?  
Info: uploaded qual454 file

@EYKX4VC01B65GS length=54 xy=0784\_1  
33 23 34 25 28 28 32 23 34 27 4+  
@EYKX4VC01BNCSP length=187 xy=0558\_1  
27 35 26 25 37 28 37 28 25 28 27 36+  
S - Sanger Phred+33, 93 values (0, 93) (0 to 60 expected in raw reads)  
I - Illumina 1.3 Phred+64, 62 values (0, 62) (0 to 40 expected in raw reads)  
X - Solexa Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)

Quality Score Comparison

Diagram adapted from [http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format)

## 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

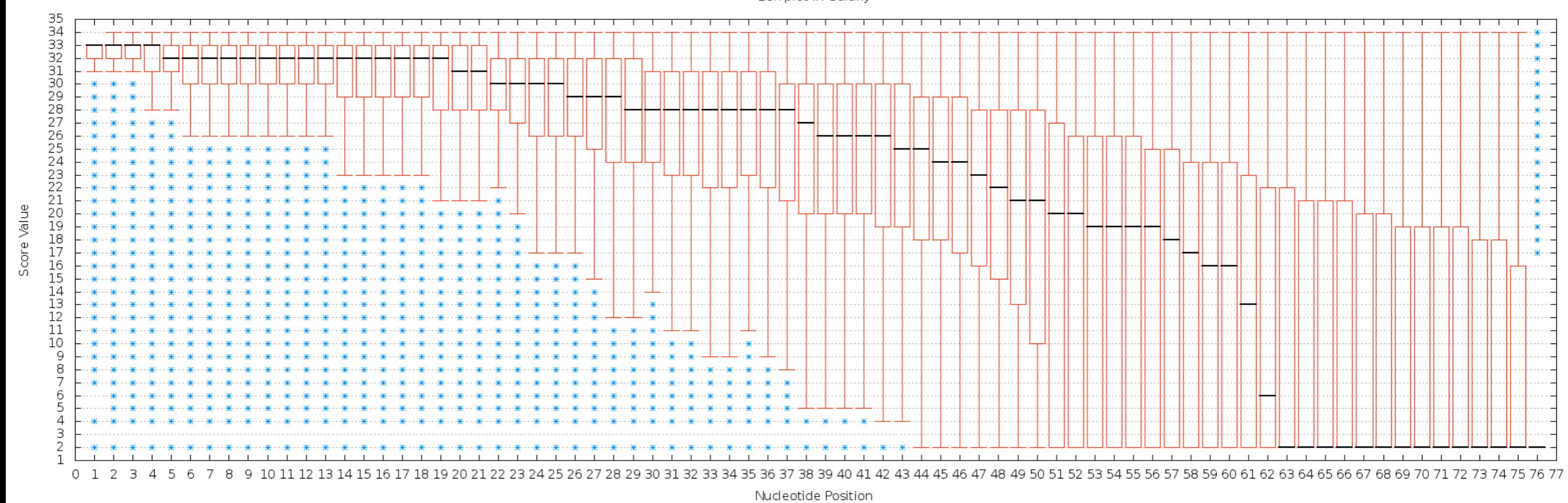
# Quality Statistics and Box Plot Tool

[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

Quartiles  
Medians  
Outliers

Box plot in Galaxy



# FastQC

Galaxy main.g2.bx.psu.edu Analyze Data Workflow Shared Data Visualization Admin Help User

Tools Options fastqc NGS: QC and manipulation ROCHE-454 DATA • Combine FASTA and QUAL into FASTQ FASTQ QC • Fastqc: Fastqc QC using FastQC from Babraham Workflows

## FastQC Report

Mon 20 Jun 2011 dataset\_1750787.dat

### Summary

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

### Basic Statistics

Measure	Value
Filename	dataset_1750787.dat
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9

History Options imported: Joe practice 6- 70.5 Mb 14-11 120: FastQC.html 11.6 Kb format: html, database: hg18 HTML file 119: Cuffdiff on data 11, data 13, and data 29: transcript FPKM tracking 118: Cuffdiff on data 11, data 13, and data 29: transcript differential expression testing 117: Cuffdiff on data 11, data 13, and data 29: gene FPKM tracking 116: Cuffdiff on data 11, data 13, and data 29: gene differential expression testing 115: Cuffdiff on data 11, data 13, and data 29: TSS groups FPKM tracking 114: Cuffdiff on data 11, data 13, and data 29: TSS groups differential expression testing 113: Cuffdiff on data 11, data 13, and data 29: CDS FPKM tracking 112: Cuffdiff on data 11, data 13, and data 29: CDS FPKM differential expression testing 111: Cuffdiff on data 11, data 13, and data 29: CDS overloading differential expression testing

# Read Trimming



Analyze Data   Workflow   Shared Data   Visualization   Admin   Help   User

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 trim the ends of reads. When using the percent-based method, offsets are calculated relative to the total length of the read.

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

```
@Some FASTQ Sanger Read  
CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA  
+  
=@@ .@;B-@?8>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**

## 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

# Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

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

# Mapping HTS Data

Collection of interchangeable mappers

- ♦ accept fastq format, produce SAM/BAM

Mappers for

- ♦ DNA
- ♦ RNA
- ♦ Local realignment

# Mappers

## DNA

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

## Metagenomics

- ♦ Megablast

## RNA / gapped-reads mapper

- ♦ Tophat

# Commonly Used/Default Parameters

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:

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:

## Lastz

Align sequencing reads in:

53: FASTQ to FASTA on data 7

Against reference sequences that are:

locally cached

Using reference genome:

Aedes aegypti: AaegL1

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

Output format:

SAM

Lastz settings to use:

Full Parameter List

Commonly used

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

# Full Parameter List

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.

# Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

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

# SNPs & INDELS

## SNPs from Pileup

- ♦ Generate
- ♦ Filter

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Tools' (selected), 'Options', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. On the left, a sidebar titled 'NGS: SAM Tools' lists various 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', and 'flagstat provides simple stats on BAM files'. The main panel is titled 'Indel Analysis' and contains the following form fields:

- Select sam file to analyze: dropdown menu showing '54: BAM-to-SAM on dat..nverted SAM'
- Frequency threshold: input field with value '0.015'
- Cutoff (button)
- Execute (button)

Below the form is a section titled 'What it does' with the following text:

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	TTCAAT
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	INSPROPCALC	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	---	---	---	---
4	-	1	1/5	0.20	---	---	---	---
	A	1	1/6	0.17	---	---	---	---

# GATK Tools

Local re-alignment

Base re-calibration

Genotyping

Alpha status

- ♦ please try, report bugs
- ♦ available on test server:  
<http://test.g2.bx.psu.edu/>

## NGS: GATK Tools

### REALIGNMENT

- [Realigner Target Creator](#) for use in local realignment
- [Indel Realigner](#) – perform local realignment

### BASE RECALIBRATION

- [Count Covariates](#) on BAM files
- [Table Recalibration](#) on BAM files
- [Analyze Covariates](#) – perform local realignment

### GENOTYPING

- [Unified Genotyper](#) SNP and indel caller

# Unified Genotyper

## Inputs

- ♦ BAM files

*Lots of possible parameters*

## Output

- ♦ VCF file(s)

The screenshot shows the command-line interface for the Unified Genotyper tool. It includes sections for sample BAM files, reference genome selection, dbSNP data, binding for reference-ordered data, and GATK analysis options.

- Sample BAM files:** A list containing "Sample BAM file 1". Below it is a "BAM file:" dropdown and a "Remove Sample BAM file 1" button. There is also an "Add new Sample BAM file" button.
- Using reference genome:** A dropdown menu set to "Mosquito (Aedes aegypti): AaegL1".
- dbSNP reference ordered data (ROD):** A dropdown menu set to "Selection is Optional".
- Binding for reference-ordered data:** An "Add new Binding for reference-ordered data" button.
- Calling thresholds:** Two input fields for phred-scaled confidence thresholds:
  - "The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called": Value 30.0.
  - "The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted (and filtered if less than the calling threshold)": Value 30.0.
- GATK options:** A dropdown menu set to "Basic".
- Analysis options:** A dropdown menu set to "Basic".
- Execute:** A "Execute" button at the bottom.

# Overview

High-throughput Sequencing (HTS) Data

## Using Galaxy to Analyze HTS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- **Binding sites analysis and peak calling**
- Transcriptome analysis

Galaxy for Sequencing Facilities

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

# Peak Calling / ChIP-seq analysis

## Punctate binding

- ♦ transcription factors

## Diffuse binding

- ♦ histone modifications
- ♦ PolII

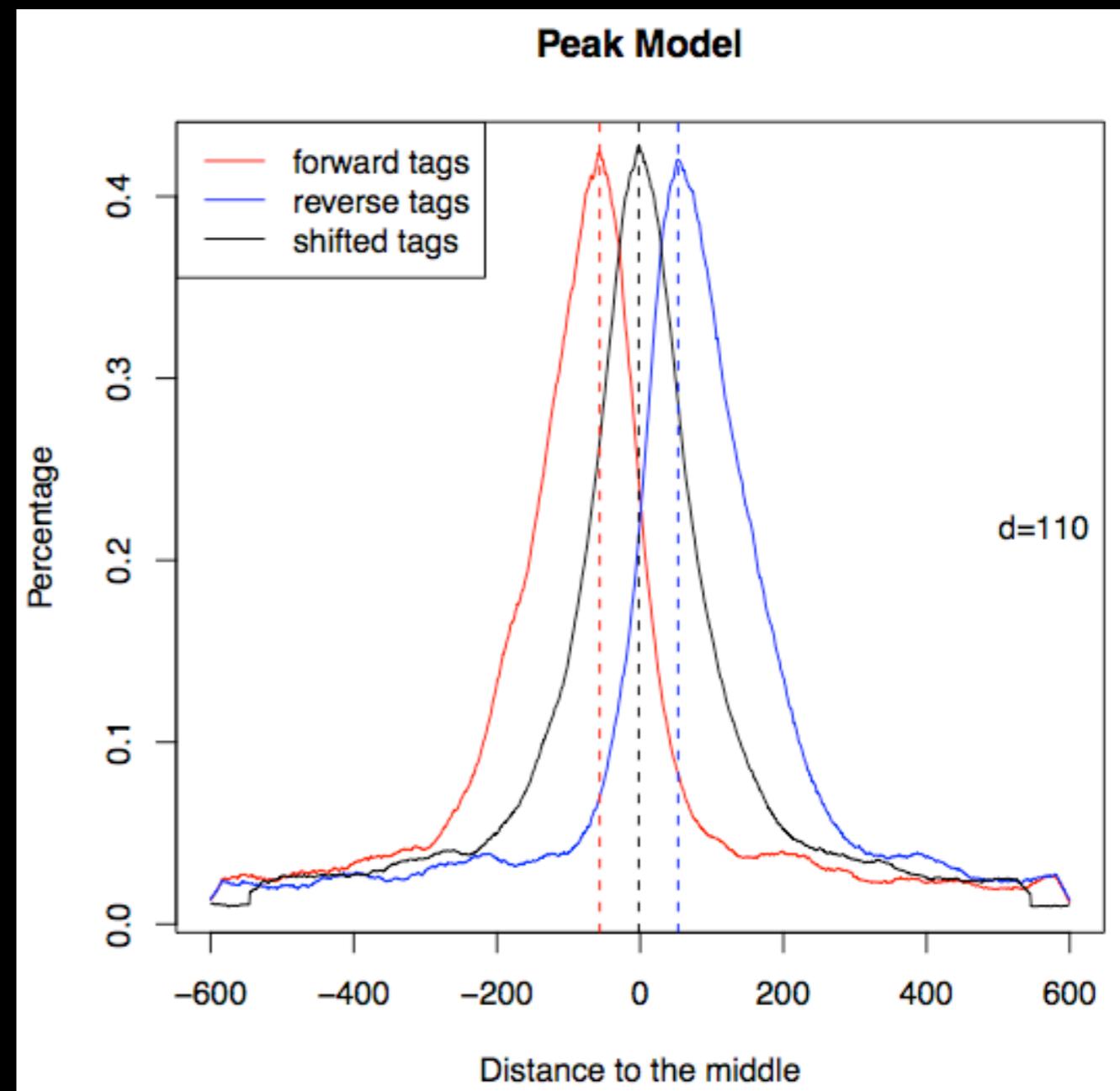
# Punctate Binding --> MACS

## Inputs

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

## Outputs

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



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

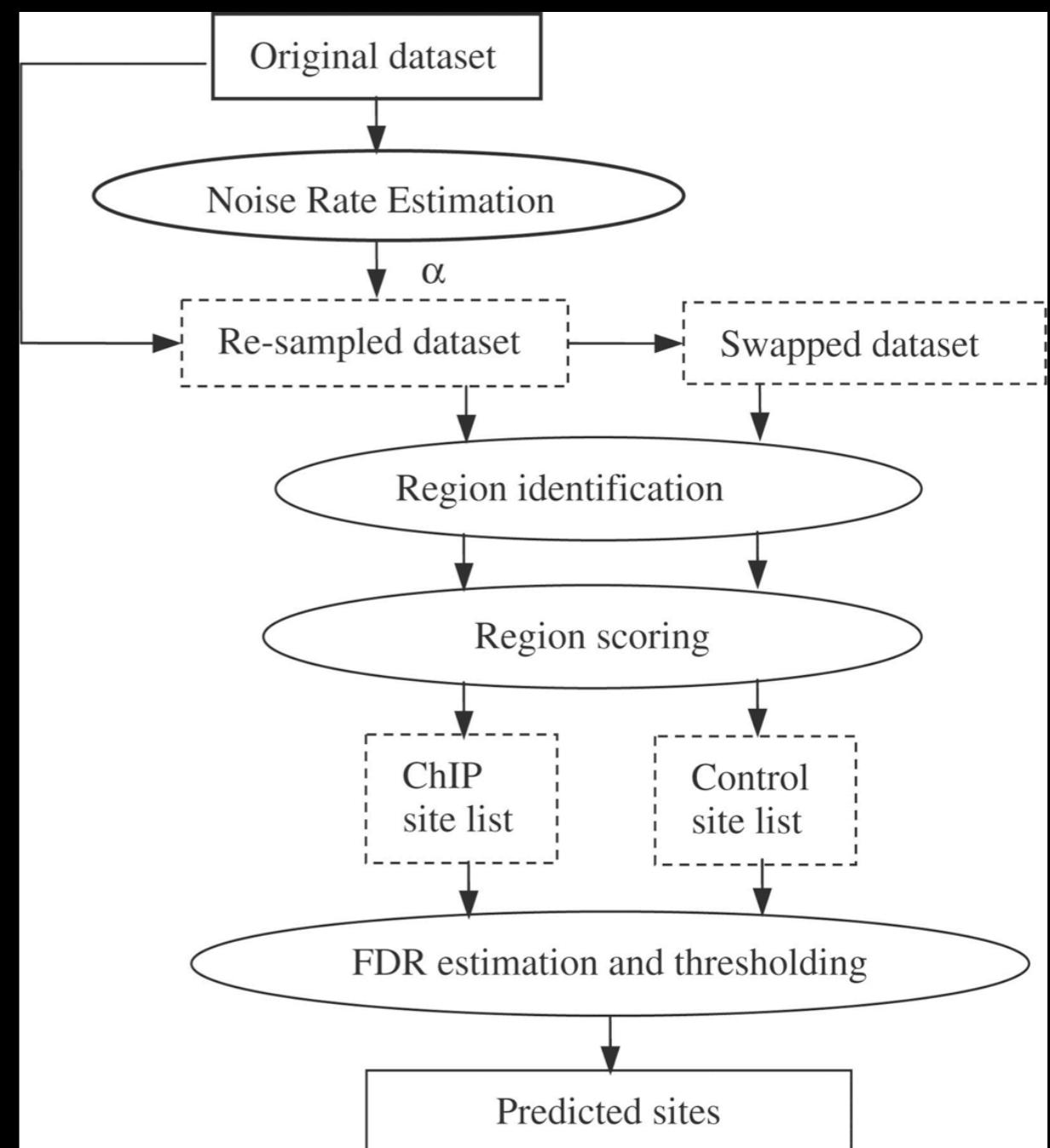
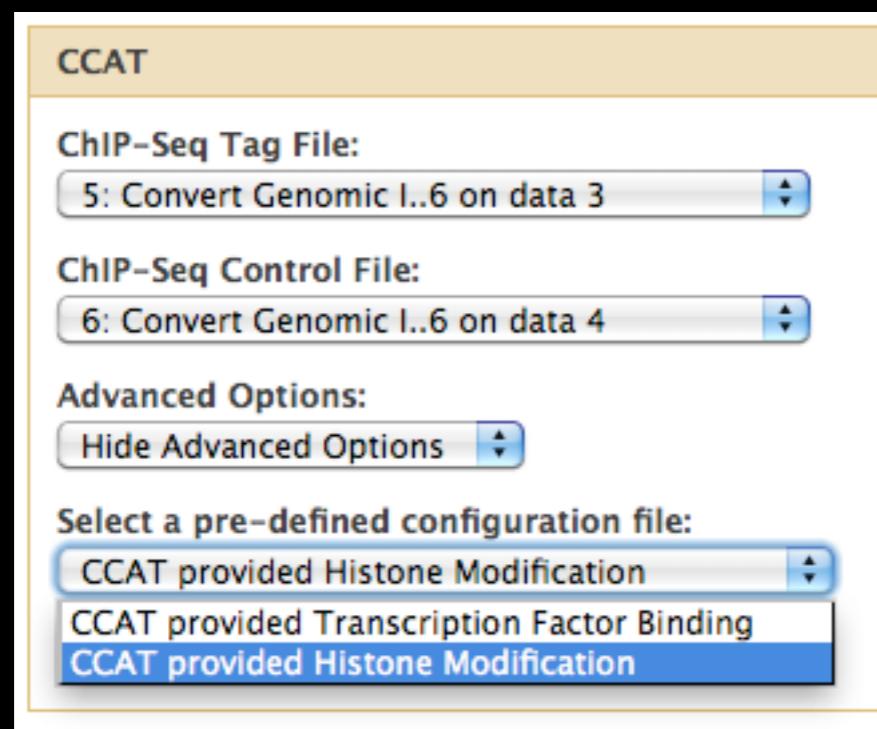
# MACS --> GeneTrack



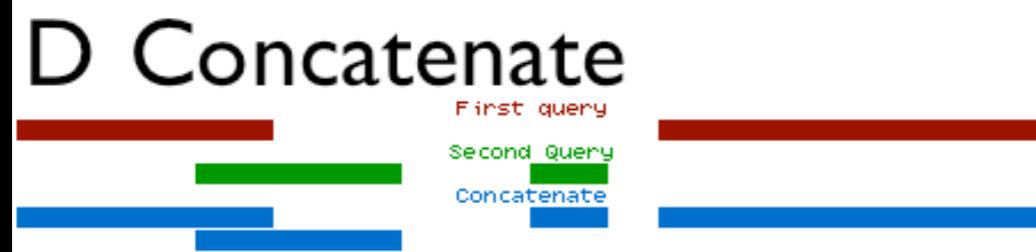
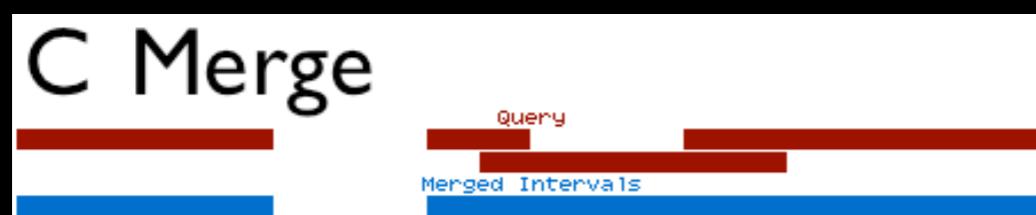
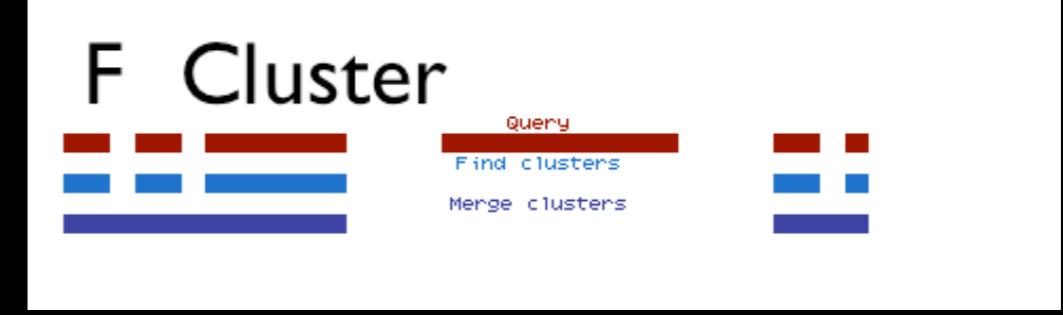
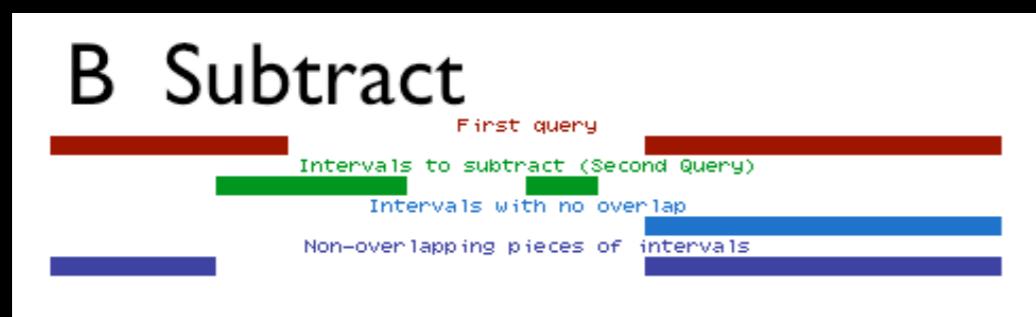
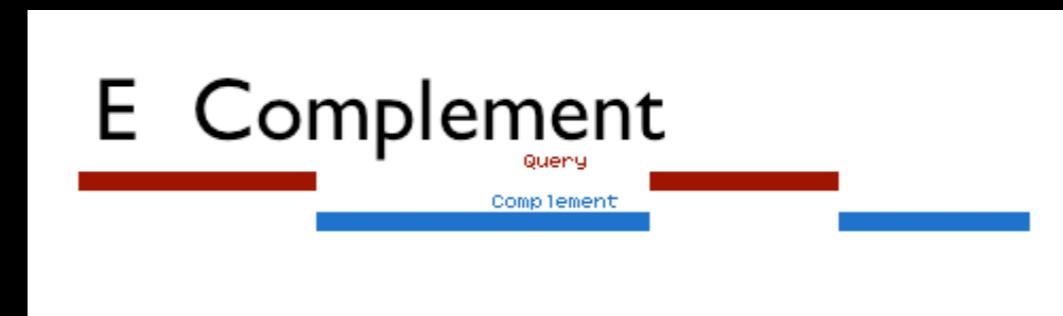
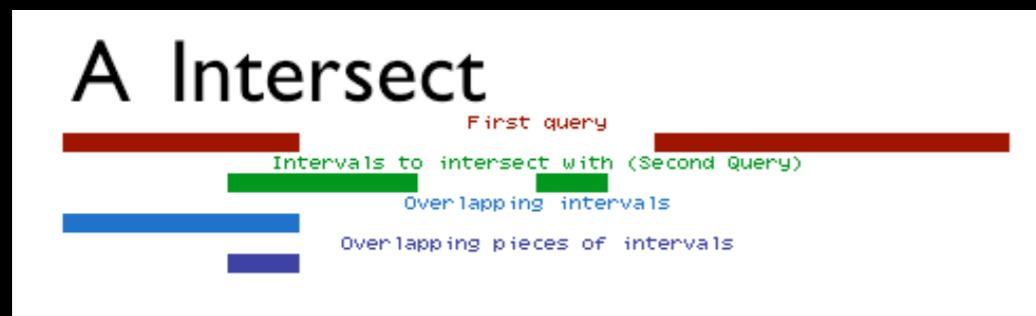
Albert I, Wachi S, Jiang C, Pugh BF. GeneTrack--a genomic data processing and visualization framework. Bioinformatics. 2008 May 15;24(10):1305-6. Epub 2008 Apr 3.

# Diffuse Binding

## CCAT (Control-based ChIP-seq Analysis Tool)



# I have Peaks, now what?



Compare to other annotations using interval operations

# 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 Data Workflow Shared Data Admin Help User

Tools Options

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

History Options

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	-
160945,328960,353082,363947,364951,389516,393					

# 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**

# Overview

High-throughput Sequencing (HTS) Data

## Using Galaxy to Analyze HTS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

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

# Transcriptome Analysis (with a reference genome)

TopHat

Cufflinks/compare/diff

<u>NGS: RNA Analysis</u>
<u>RNA-SEQ</u>
<ul style="list-style-type: none"><li>■ <a href="#">Tophat</a> Find splice junctions using RNA-seq data</li><li>■ <a href="#">Cufflinks</a> transcript assembly and FPKM (RPKM) estimates for RNA-Seq data</li><li>■ <a href="#">Cuffcompare</a> compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments</li><li>■ <a href="#">Cuffdiff</a> find significant changes in transcript expression, splicing, and promoter use</li></ul>
<u>FILTERING</u>
<ul style="list-style-type: none"><li>■ <a href="#">Filter Combined Transcripts</a> using tracking file</li></ul>

1. Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).
2. Trapnell et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621

# TopHat

## Map RNA (FASTQ) to a reference Genome

- ♦ gapped mapper

## Outputs

- ♦ BAM file of accepted hits
- ♦ BED file of splice junctions

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

Goal: transcript assembly and quantitation

Input: aligned RNA-Seq reads, usually from TopHat

## Outputs

- assembled transcripts (GTF)
- genes' and transcripts' coordinates, expression levels

Cufflinks

SAM or BAM file of aligned RNA-Seq reads:  
13: Tophat on data 1...cepted\_hits

Max Intron Length:  
300000

Min Isoform Fraction:  
0.05

Pre mRNA Fraction:  
0.05

Perform quartile normalization:  
No

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance genes.

Use Reference Annotation:  
No

Perform Bias Correction:  
Yes

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Reference sequence data:  
Locally cached

Set Parameters for Paired-end Reads? (not recommended):  
No

Execute

# Cuffcompare

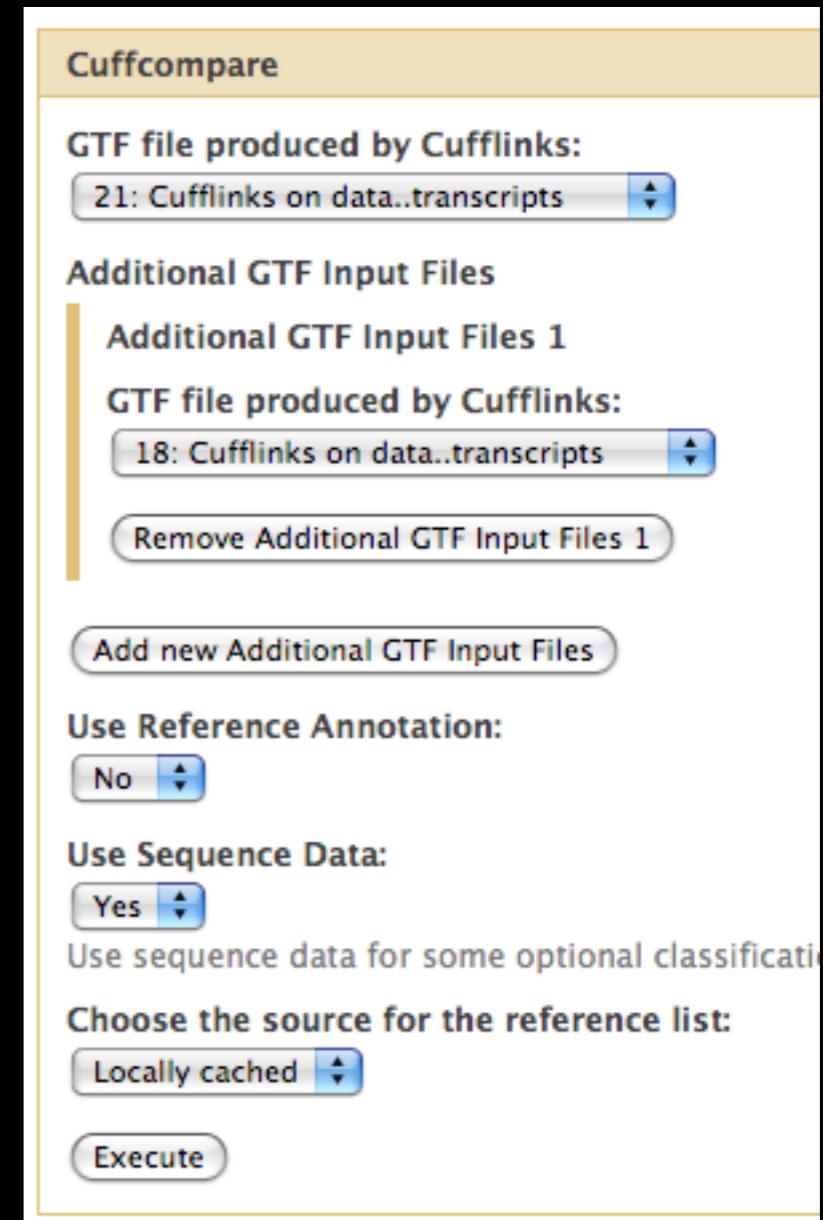
## Goals

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

Inputs: assembled transcripts from Cufflinks

## Outputs:

- Transcripts Combined File
- Transcripts Accuracy File
- Transcripts Tracking Files



# Cuffdiff

## Goals

- differential expression testing
- transcript quantitation

## Inputs

- Combined set of transcripts
- mapped reads from 2+ samples

## Outputs

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

**Cuffdiff**

**Transcripts:**  
29: Cuffcompare on da..transcripts  
A transcript GTF file produced by cufflinks, cuffcompare, or other source.

**Perform replicate analysis:**  
No  
Perform cuffdiff with replicates in each group.

**SAM or BAM file of aligned RNA-Seq reads:**  
11: Tophat on data 9...cepted\_hits  
13: Tophat on data 1...cepted\_hits

**SAM or BAM file of aligned RNA-Seq reads:**  
13: Tophat on data 1...cepted\_hits

**False Discovery Rate:**  
0.05  
The allowed false discovery rate.

**Min Alignment Count:**  
1000  
The minimum number of alignments in a locus for needed to conduct significance testing or

**Perform quartile normalization:**  
No  
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression.

**Perform Bias Correction:**  
Yes  
Bias detection and correction can significantly improve accuracy of transcript abundance estimation.

**Reference sequence data:**  
Locally cached

**Set Parameters for Paired-end Reads? (not recommended):**  
No

**Execute**

# Next Steps

## Filtering

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

Extract transcript sequences and profile sequences for function

**Filter Combined Transcripts**

Cufflinks assembled transcripts:  
130: Cuffcompare on da..transcripts

Cuffcompare tracking file:  
130: Cuffcompare on da..transcripts

Sample Number:  
1

Execute

**Filter**

Filter:  
130: Cuffcompare on da..transcripts  
Dataset missing? See TIP below.

With following condition:  
c14=='yes'  
Double equal signs, ==, must be used as

Execute

**Extract Genomic DNA**

Fetch sequences for intervals in:  
130: Cuffcompare on da..transcripts

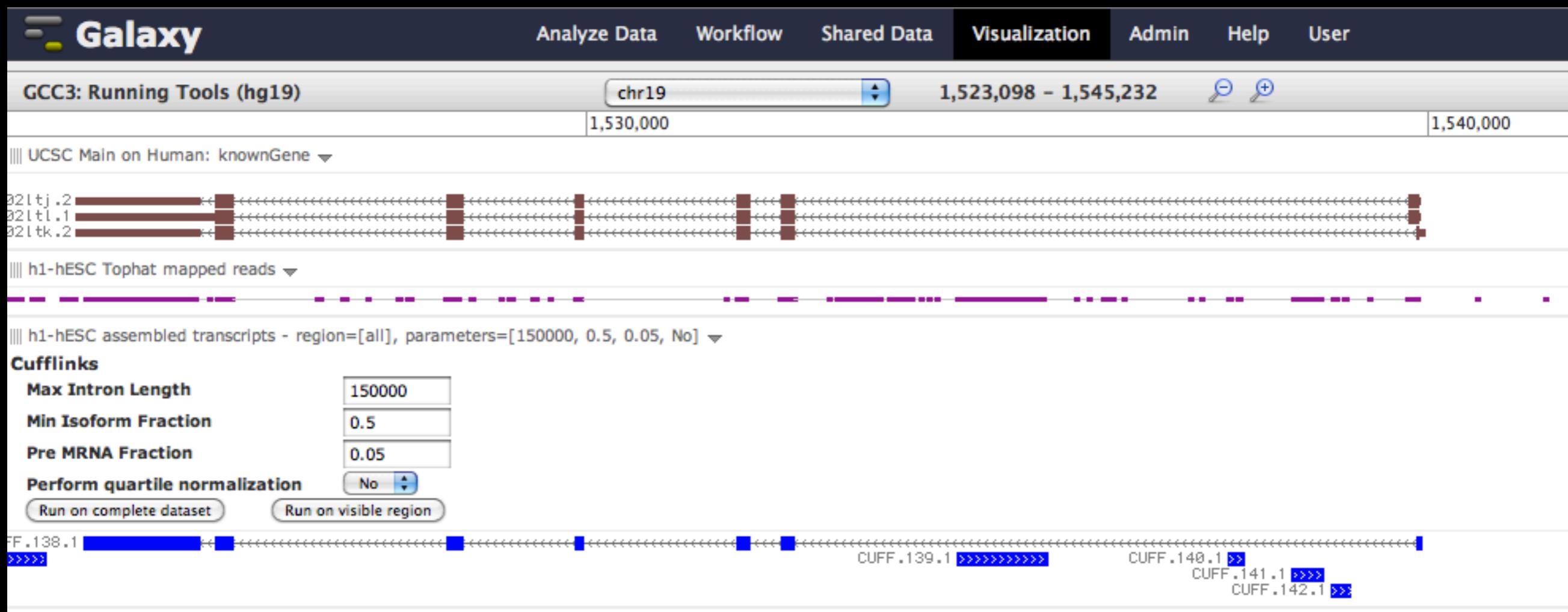
Interpret features when possible:  
Yes  
Only meaningful for GFF, GTF datasets.

Source for Genomic Data:  
Locally cached

Output data type:  
FASTA

Execute

# Integrating Tools and Visualization



|||| h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▾

### Cufflinks

Max Intron Length	150000
Min Isoform Fraction	0.05
Pre mRNA Fraction	0.05
Perform quartile normalization	No <input type="button" value="▼"/>
<input type="button" value="Run on complete dataset"/> <input type="button" value="Run on visible region"/>	



→ Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No] ▾





**Working to add GATK Unified Genotyper (and  
more!) to Trackster as well**

# Working with HTS Tools

Often challenging

- many parameters
- time intensive
- evaluating results difficult

Good options

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

# Overview

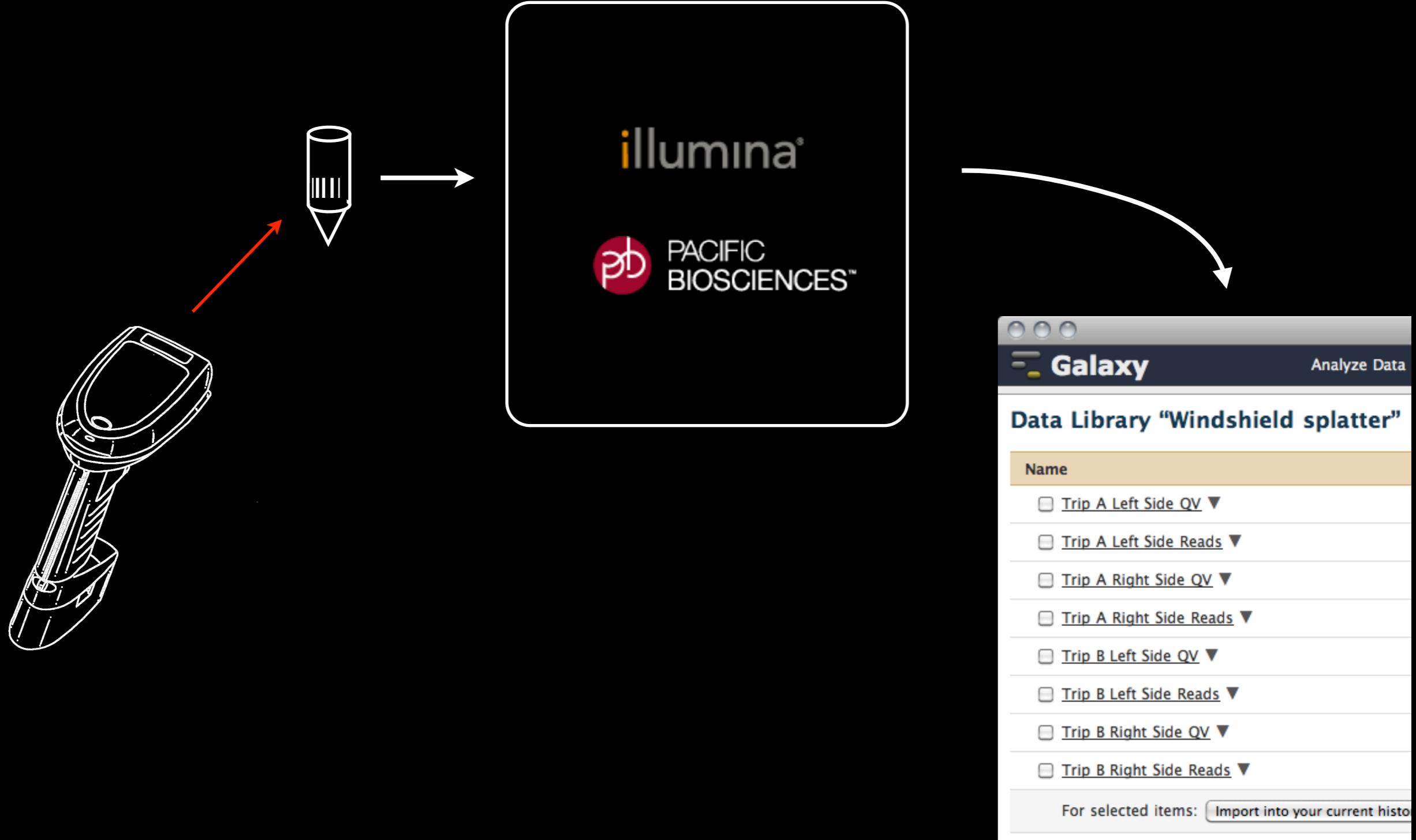
High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

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



Sample information tracked in Galaxy, state changes through laboratory workflow are captured, data is linked back to sample in user's workspace

# Sample Tracking System

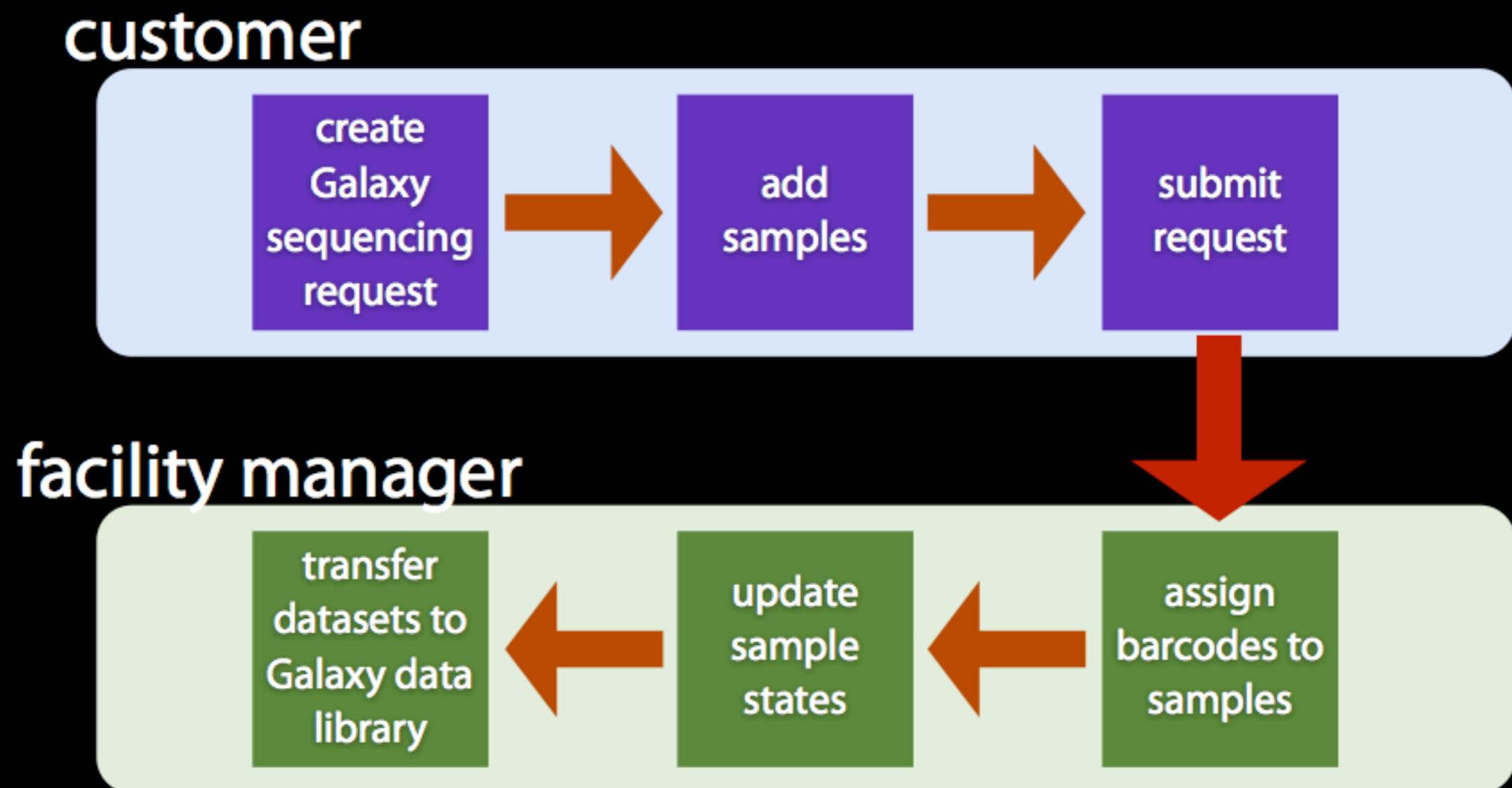
Built-in system for tracking sequencing requests

Customizable interfaces

- Sequencing Facility Managers/Administrators
- Users/Biologists

Streamlines data delivery: sequencing runs to users

# How does it all work?



# Sequencing Facility Managers

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

Create and submit a sequencing request on behalf of another user

Reject an incomplete or erroneous sequencing request

Receive samples and assign them tracking barcodes.

Setup data transfer from the sequencer

# Sequencing Facility Users

Create and submit a sequencing request

Edit and resubmit a rejected sequencing request

Obtain datasets at the end of a sequencing run

Select Libraries and Histories, and Workflows to  
populate and run on sequenced samples.

# Configure Available Request / Sample Options

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](#)

**Forms**

[Advanced Search](#)

[Create new form](#)

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

For 0 selected forms: [Delete](#) [Undelete](#)

View

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

## Name

Atlantic Biosciences request

## 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:  - (optional) Options: SR PE	SelectField:  - (optional) Options: 36 50 75 100	TextField:  - (optional)	TextField:  - (optional)	SelectField:  - (optional) Options: Std Rush option3	TextField:  - (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**

search [Advanced Search](#)

[Reload external service types](#) [Create new external service](#)

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

For 0 selected externalservices: [Delete](#) [Undelete](#)

[Edit external service](#)

Name:

Description:

Version:

Hostname or IP address:   
(Required)

User name:   
(Required)

Password:    
(Required)

Data directory:

# User Creates a Request

The screenshot shows the Galaxy web interface with the 'Sequencing Requests' page. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Lab, Admin, Help, and User. A dropdown menu from the 'Lab' link contains options for Sequencing Requests, Find Samples, and Help. On the left, there's a search bar with a magnifying glass icon and a link to 'Advanced Search'. The main content area displays a table header with columns: Name, Description, Samples, Type, Last Updated ↑, and State. Below the header, a message says 'No Items'. At the bottom, there are buttons for 'Delete' and 'Undelete'.

The screenshot shows the Galaxy web interface with the 'Create a new sequencing request' form. The top navigation bar is identical to the previous screenshot. The main form has a title 'Create a new sequencing request'. It asks 'Select a request type configuration:' and provides a dropdown menu with 'Select one' at the top and 'Atlantic Biosciences' highlighted in blue. A note below the dropdown says 'if you are not sure about the request type configuration.'

# User Creates a Request

Galaxy Analyze Data Workflow Shared Data Lab Admin Help User

## Sequencing Requests

search Advanced Search

Name	Description	Samples
No Items		

For 0 selected requests: [Delete](#) [Undelete](#)

[Create new request](#)

Sequencing Requests

Find Samples

Galaxy Analyze Data Workflow Shared Data

## Create a new sequencing request

Select a request type configuration:

Select one  Atlantic Biosciences

If you are not sure about the request type configuration, contact the lab manager.

Name of the Experiment  
My first ChIP-seq Experiment  
(Required)

Description  
This is Experiment was performed using the protocol  
(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](#)

# User 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
Sample_1 <small>(required)</small>		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 1 samples from sample None

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

Add sample   Save   Cancel

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

▶ Import samples from csv file

History  
Workflow to run

# Samples Added, Submit Request

Galaxy

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	<a href="#">Dan's Sequencing Requests</a>	ChIP-seq	<a href="#">My own ChIP-seq Experiment!</a>	<a href="#">Dan's ChIP-seq Workflow</a>

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.

[Add sample](#)

Click the Add sample button for each new sample.

▶ Import samples from csv file

# Samples enter “New” state



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		<u>New</u>	<u>Dan's Sequencing Requests</u>	ChIP-seq	<u>My own ChIP-seq Experiment!</u>	<u>Dan's ChIP-seq Workflow</u>	0

▶ Layout1

# 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

search

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

For 0 selected requests:

# Sequencing Facility Receives Samples

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

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

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

Click the Save button when you have finished editing the samples

Sequencing Requests

Name	Description	Samples	Type	Last Updated ↑	State
My first ChIP-seq Experiment	This is Experiment was performed using the protocol ...	1	Atlantic Biosciences	35 minutes ago	Complete

For 0 selected requests:

## Facility

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

User can watch progress of sequencing request

# Sequencing Finished

Datasets are transferred from sequencer into Galaxy

- ♦ library
- ♦ user's history

Galaxy Workflow is executed on Dataset

User is automatically emailed

# Extending Sample Tracking with ngLims

An add-on written by community contributor Brad Chapman

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

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

<http://bcbio.wordpress.com/2011/01/11/next-generation-sequencing-information-management-and-analysis-system-for-galaxy/>

# Sample tracking is completely extensible

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

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

Interaction with sequence instruments and secondary analysis is completely pluggable

- For services that provide a web / REST API even easier

**Samples**

- [Define samples and services](#)
- [Submit samples as a project](#)
- [View projects](#)

**Sequencing**

- [Queues](#)
- [Runs](#)

**Samples**

TJa3 Sample 4

**Copy****Lane 1**

TJa1 Sample 2

**Lane 2**

PhiX control

**Lane 3**

BCa1 Test sample 1

**Lane 4****Lane 5****Lane 6****Lane 7****Lane 8**

Example: extensions from Brad Chapman for flowcell layout, multiplexing, ...

# Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data

- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy for Sequencing Facilities

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



EMORY

PENNSTATE



Enis Afgan



Dannon Baker



Dan Blankenberg



Nate Coraor



Dave Clements



Jeremy Goecks



Jennifer Jackson



Greg von Kuster



Kanwei Li



James Taylor



Kelly Vincent



Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

# Using Galaxy

Use public Galaxy server: UseGalaxy.org

Download Galaxy source: GetGalaxy.org

Galaxy Wiki: GalaxyProject.org

Screencasts: GalaxyCast.org

Public Mailing Lists

- galaxy-bugs@bx.psu.edu
- galaxy-user@bx.psu.edu
- galaxy-dev@bx.psu.edu

# ChIP-seq and RNA-seq exercises

## RNA-seq

- <http://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise>
  - start Tophat mapping first (second section), then look at QC (first section)

## ChIP-seq

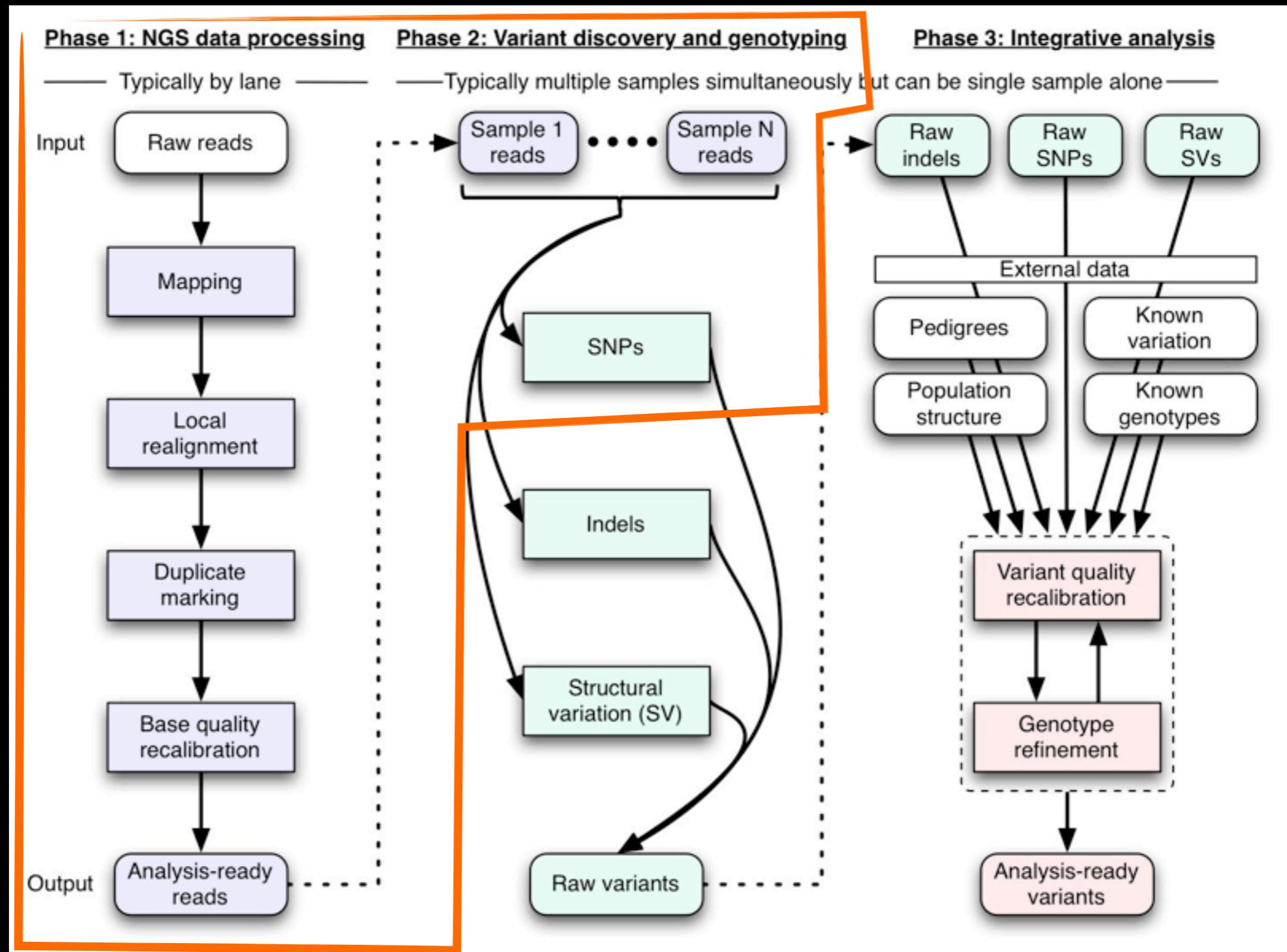
- <http://usegalaxy.org/u/james/p/exercise-chip-seq>

## Variant Detection

- Using GATK, Picard Tools

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

# Variant Detection



Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kurnytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8.