

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

GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

! ' ' * ((((* * * +)) % % % + +) (% % % %) . 1 * * * - + * ' ') * * 5 5 C C F > > > > > C C C C C C C 6 5 5

[illegible]

```
S - Sanger      Phred+33, raw reads typically (0, 40)
X - Solexa     Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
```

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

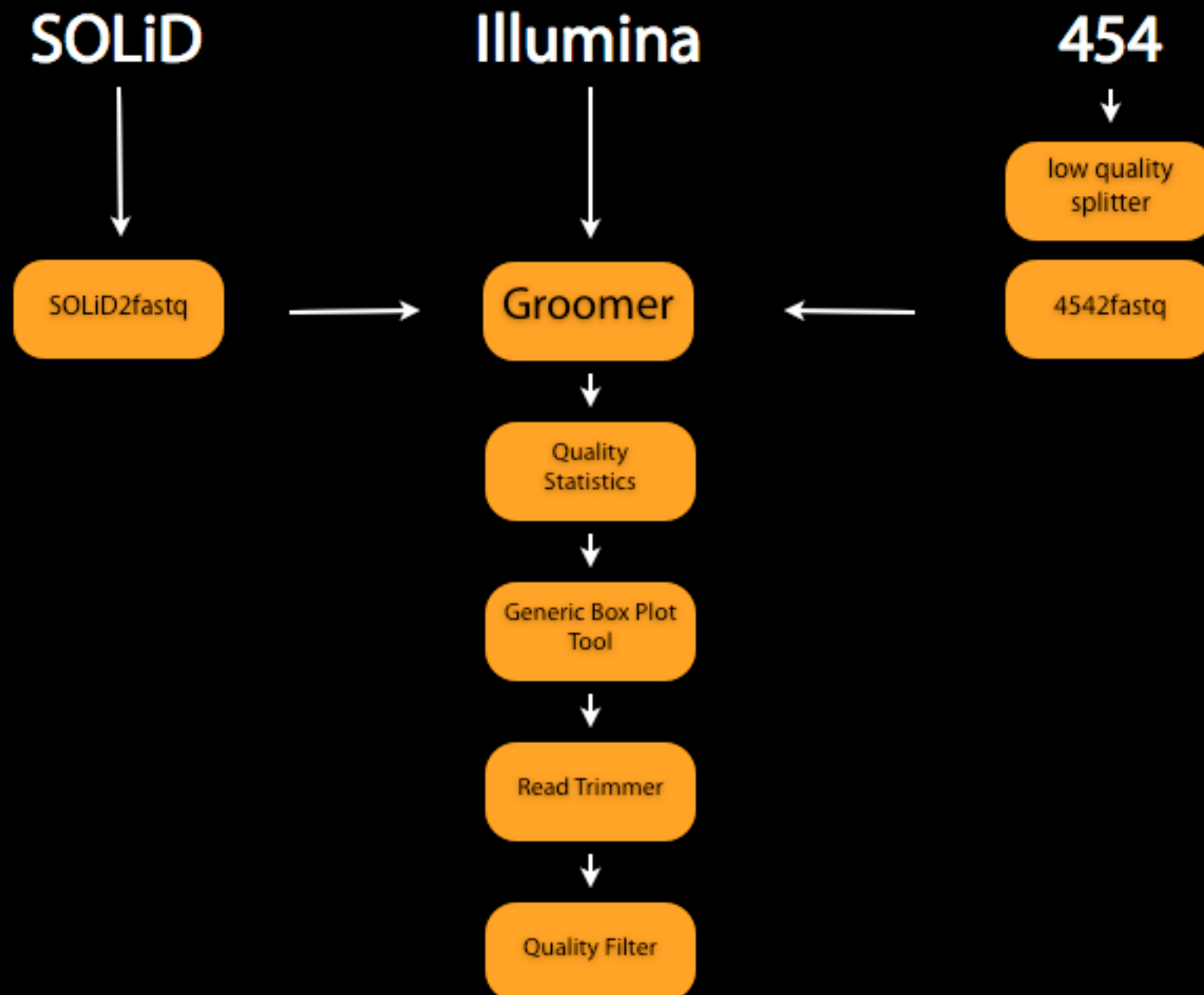
Using Galaxy to Analyze HTS Data

- ✦ 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



Grooming --> Sanger

[Analyze Data](#)
[Workflow](#)
[Shared Data](#)
[Visualization](#)
[Admin](#)
[Help](#)
[User](#)

Tools
Options

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

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 format (Sanger, Solexa, Illumina, etc.).

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

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

When converting between Solexa and the other formats, quality scores are converted using the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, et al. (2009) *Quality scores, and the Solexa/Illumina FASTQ variants*. Nucleic Acids Res 37(21):6157-6161.

When converting between color space (csSanger) and base space (Sanger), 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

CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA

+

BBC:===ABC<@==@6=<<=====BB=B9E<@6

@EYKX4VC01BNCSP length=187 xy=0558_

CTTACCGGTCACCACCGTGCTTCAGGATTGATCG

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

CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA

+

BBC:===ABC<@==@6=<<=====BB=B9E<@6

@EYKX4VC01BNCSP length=187 xy=0558_

CTTACCGGTCACCACCGTGCTTCAGGATTGATCG

2: 454.qual

52 lines

format: qual454, database: ?

Info: uploaded qual454 file

>EYKX4VC01B65GS length=54 xy=0784_1

33 23 34 25 28 28 28 32 23 34 27 4

>EYKX4VC01BNCSP length=187 xy=0558_

27 35 26 25 37 28 37 28 25 28 27 36

Quality Score Comparison

Quality Score Comparison

```

SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNPOQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
33          59      64          73          104          126

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)

```

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ_format

Quality Statistics and Box Plot Tool

NGS TOOLBOX BETA

NGS: QC and manipulation

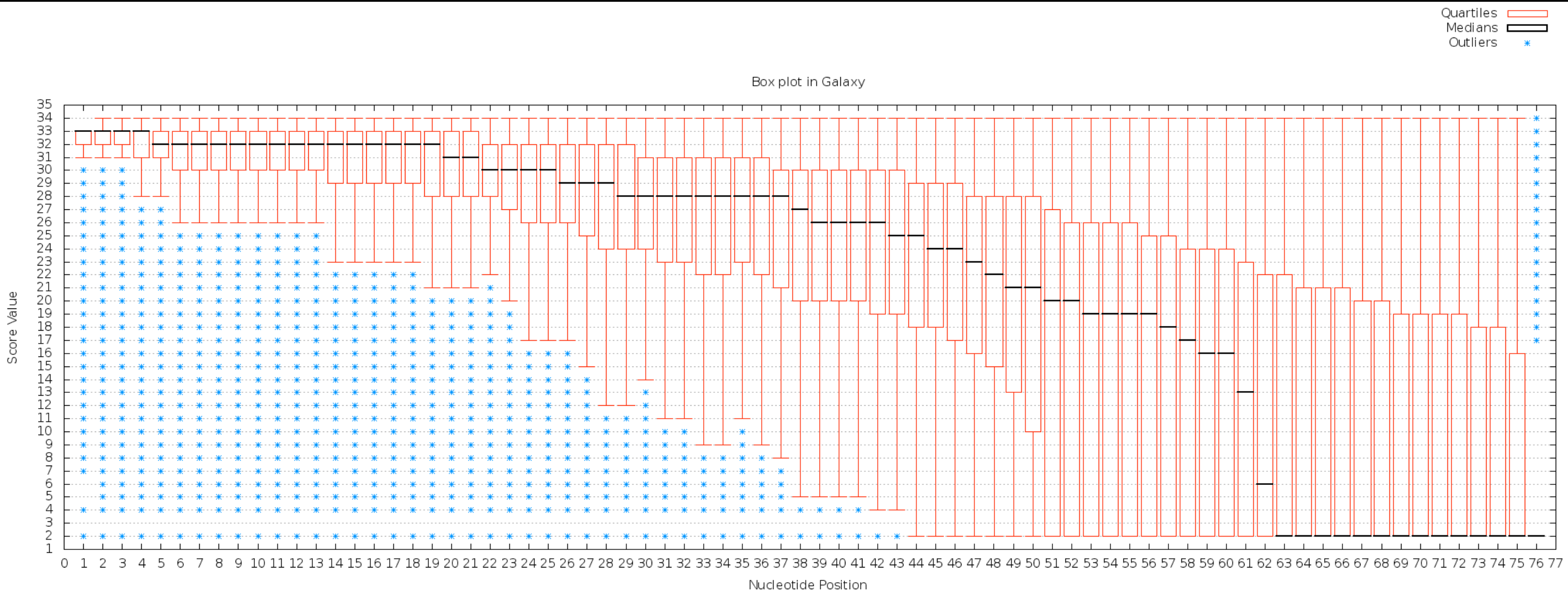
ILLUMINA DATA

- [FASTQ Groomer](#) convert between various FASTQ quality formats
- [FASTQ splitter](#) on joined paired end reads
- [FASTQ joiner](#) on paired end reads
- [FASTQ Summary Statistics](#) by column

Graph/Display Data

- [Histogram](#) of a numeric column
- [Scatterplot](#) of two numeric columns
- [Plotting tool](#) for multiple series and graph types
- [Boxplot](#) of quality statistics

Box plot in Galaxy



FastQC

The screenshot shows the Galaxy web interface with a FastQC report for dataset_1750787.dat. The report is dated Mon 20 Jun 2011. The summary section lists various quality metrics, some of which are highlighted with green checkmarks or yellow exclamation marks. The basic statistics section provides a table of key metrics.

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

Imported: Joe practice 6- 70.5 Mb 14-11

- 120: FastQC.html 11.6 Kb format: html, database: hg18
- 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

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 specified as a percentage of the read length.

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

```
@Some FASTQ Sanger Read
CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA
+
-@@.0;B-?78>CBA@>7@7BBCA4-48%<;%<B@
```

And you set absolute offsets of 2 and 16:

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:

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:

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

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

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

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 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The left sidebar lists various tools under categories like 'Fetch Alignments', 'Statistics', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Metagenomic analyses', 'Human Genome Variation', 'EMBOSS', 'NGS TOOLBOX BETA', '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. The 'What it does' section contains a table of example SAM records and a table of calculated totals.

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	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	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	---	---	---	---

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)

Unified Genotyper

Choose the source for the reference list:

Locally cached

Sample BAM files

Sample BAM file 1

BAM file:

Remove Sample BAM file 1

Add new Sample BAM file

Using reference genome:

Mosquito (Aedes aegypti): AaegL1

dbSNP reference ordered data (ROD):

Selection is Optional

Binding for reference-ordered datas

Add new Binding for reference-ordered data

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called:

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):

30.0

Basic or Advanced GATK options:

Basic

Basic or Advanced Analysis options:

Basic

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

Peak Calling / ChIP-seq analysis

Punctate binding

- ✦ transcription factors

Diffuse binding

- ✦ histone modifications
- ✦ PolII

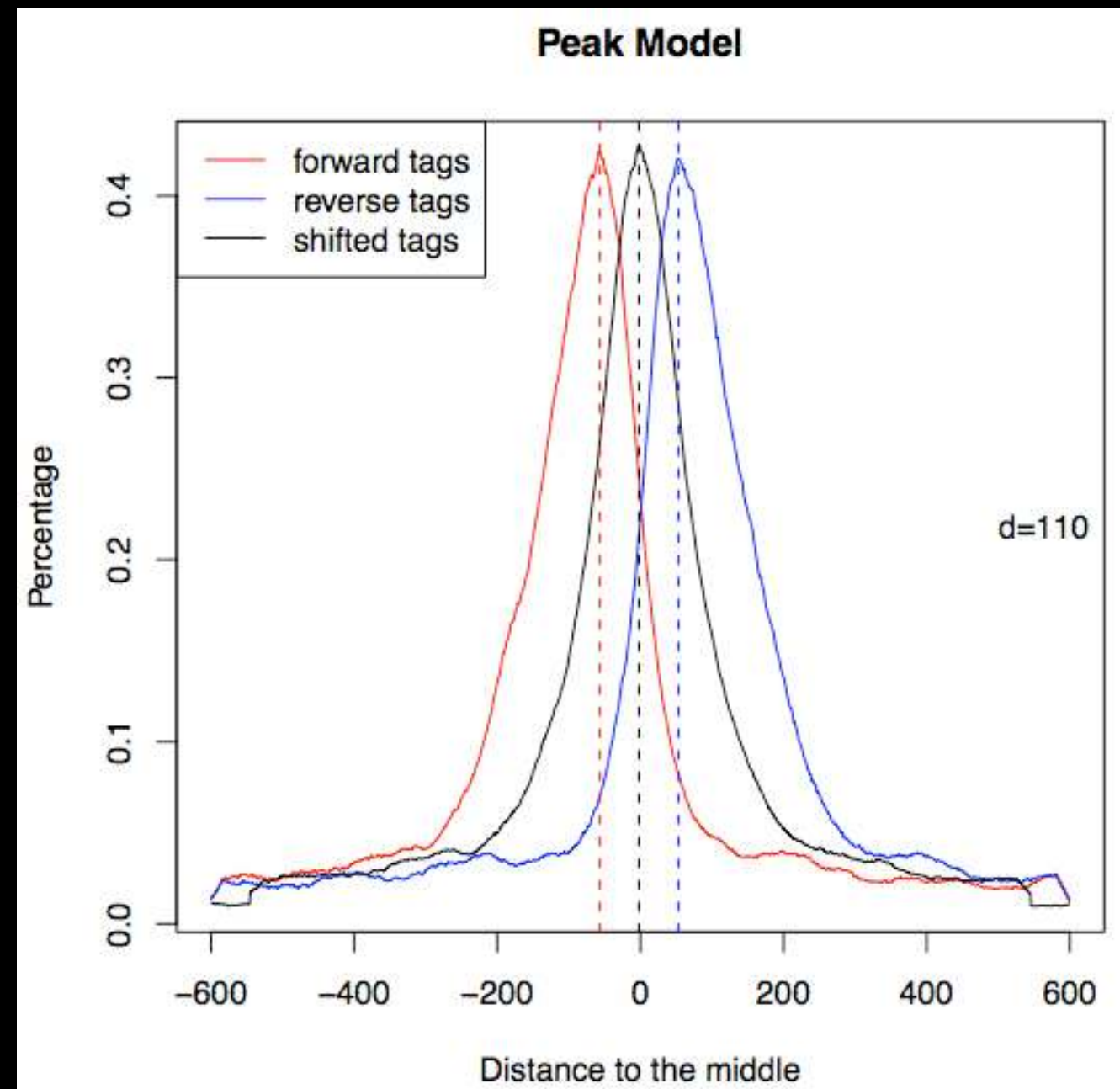
Punctate Binding --> MACS

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)



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)

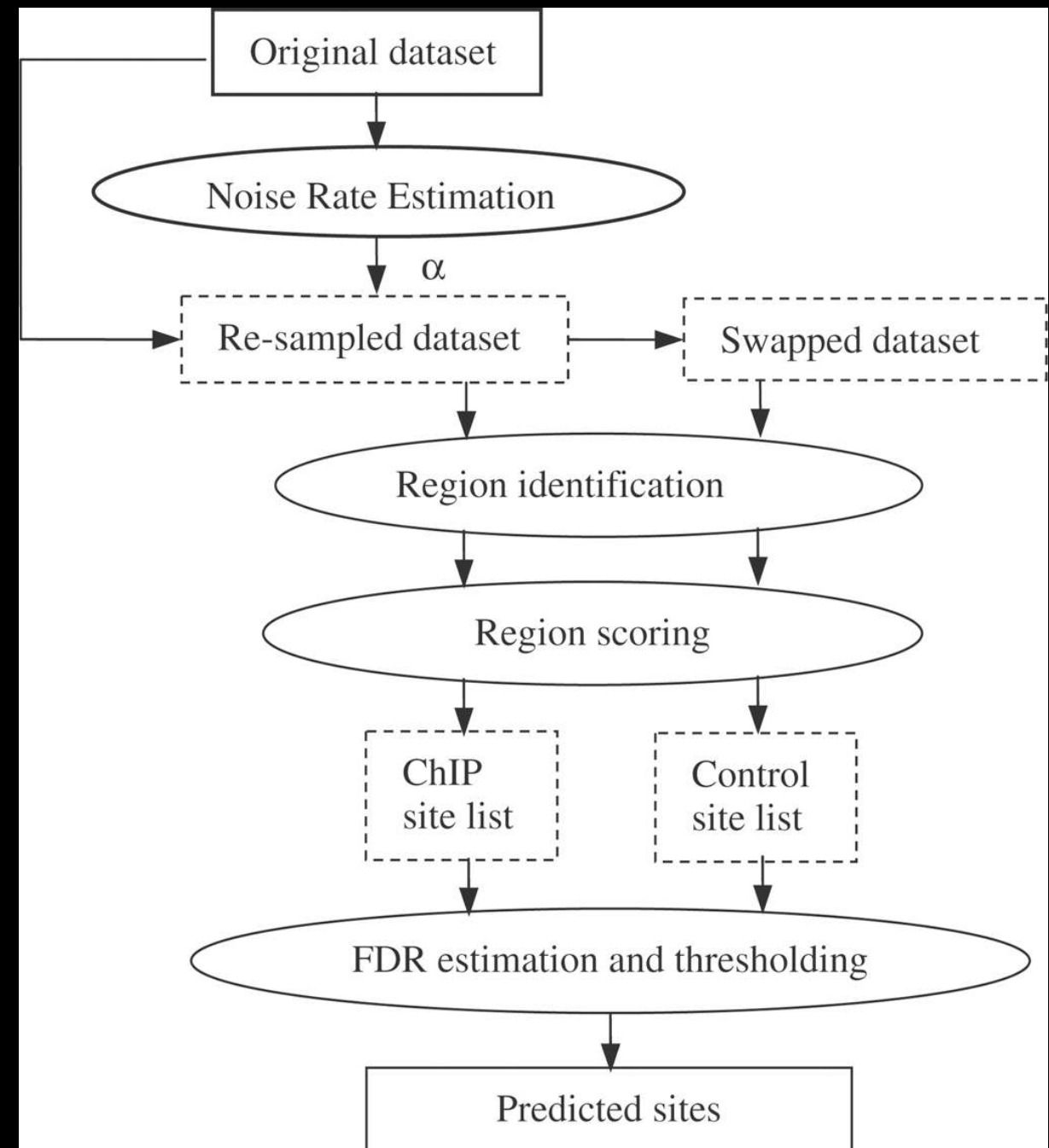
CCAT

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

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

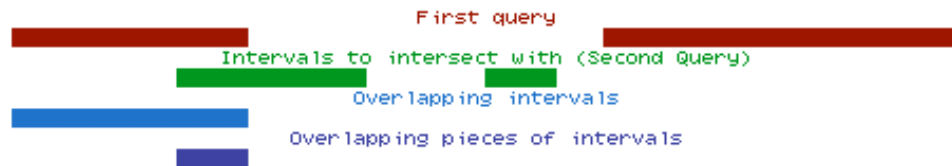
Advanced Options:
Hide Advanced Options

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

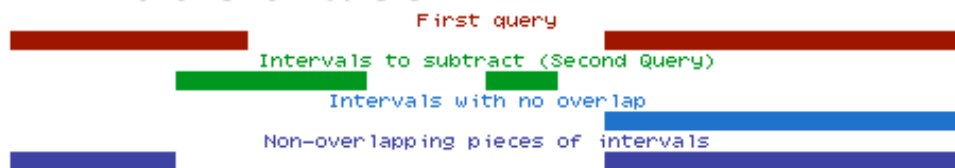


I have Peaks, now what?

A Intersect



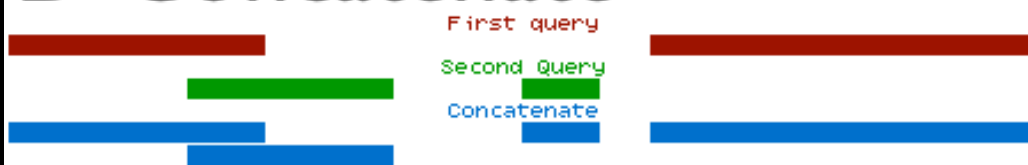
B Subtract



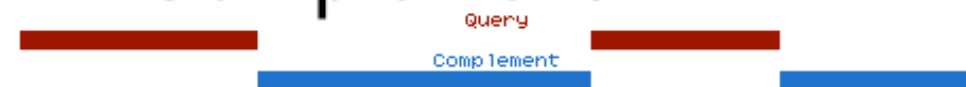
C Merge



D Concatenate



E Complement



F Cluster



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 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 UTR803

5 UTR574

coding exons2743

introns13746

other12499

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

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

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

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?:
 ⬆⬆

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?:
 ⬆⬆

RNA-Seq FASTQ file:
 ⬆⬆

Must have Sanger-scaled quality values with ASCII offset 33

TopHat settings to use:
 ⬆⬆

You can use the default settings or set custom values for any of Tophat's parameters.

Cufflinks

Goal: transcript assembly and quantitation

Input: aligned RNA-Seq reads, usually from TopHat

Outputs

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

The screenshot shows the Cufflinks web interface with the following settings:

- Cufflinks** (Title)
- 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 a
- 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** button

Cuffcompare

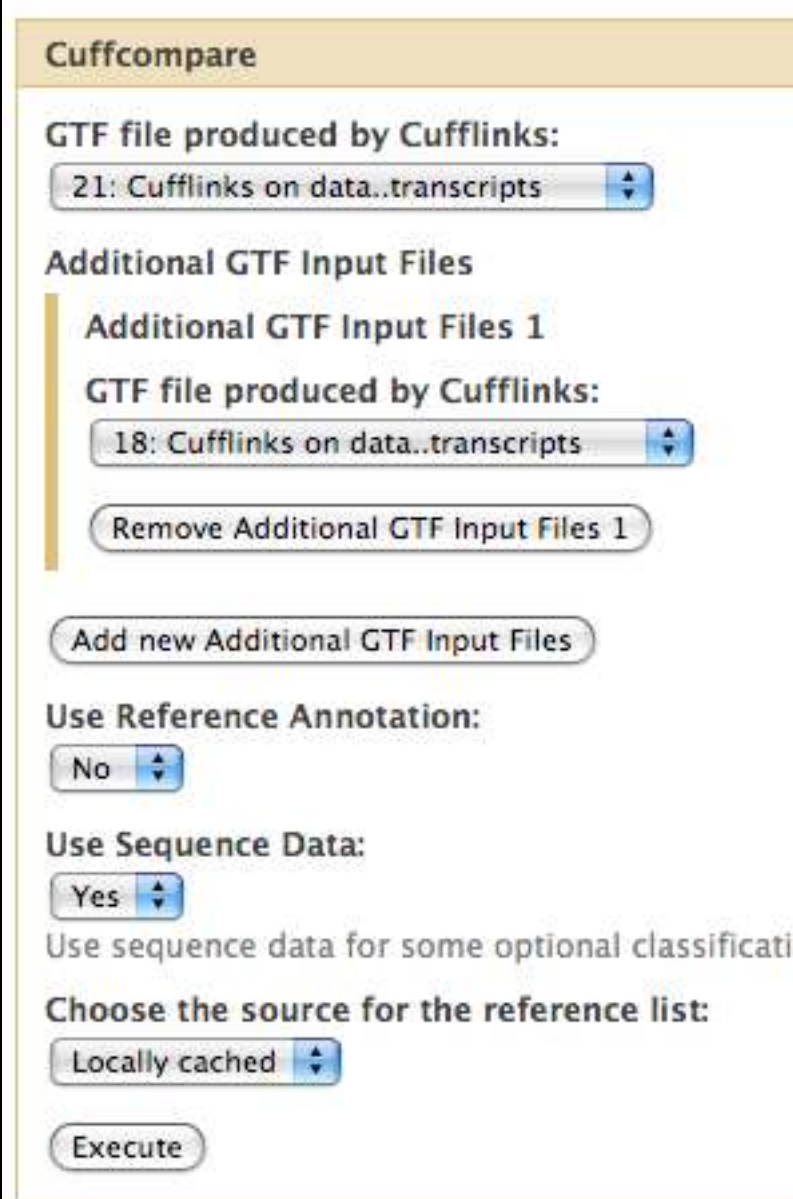
Goals

- ✦ 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



The screenshot displays the Cuffcompare web interface. At the top, the title 'Cuffcompare' is shown in a yellow header. Below this, the 'GTF file produced by Cufflinks:' section features a dropdown menu with the selected option '21: Cufflinks on data..transcripts'. The 'Additional GTF Input Files' section is highlighted with a yellow bar and contains a sub-section 'Additional GTF Input Files 1'. This sub-section includes another 'GTF file produced by Cufflinks:' dropdown menu with the option '18: Cufflinks on data..transcripts', a 'Remove Additional GTF Input Files 1' button, and an 'Add new Additional GTF Input Files' button. The 'Use Reference Annotation:' section has a dropdown menu set to 'No'. The 'Use Sequence Data:' section has a dropdown menu set to 'Yes', with a note below it stating 'Use sequence data for some optional classification'. The 'Choose the source for the reference list:' section has a dropdown menu set to 'Locally cached'. At the bottom, there is an 'Execute' button.

Cuffcompare

GTF file produced by Cufflinks:
21: Cufflinks on data..transcripts

Additional GTF Input Files

Additional GTF Input Files 1

GTF file produced by Cufflinks:
18: Cufflinks on data..transcripts

Remove Additional GTF Input Files 1

Add new Additional GTF Input Files

Use Reference Annotation:
No

Use Sequence Data:
Yes

Use sequence data for some optional classification

Choose the source for the reference list:
Locally cached

Execute

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

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 on

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

Perform Bias Correction:
Yes
Bias detection and correction can significantly improve accuracy of transcript abundance est

Reference sequence data:
Locally cached

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

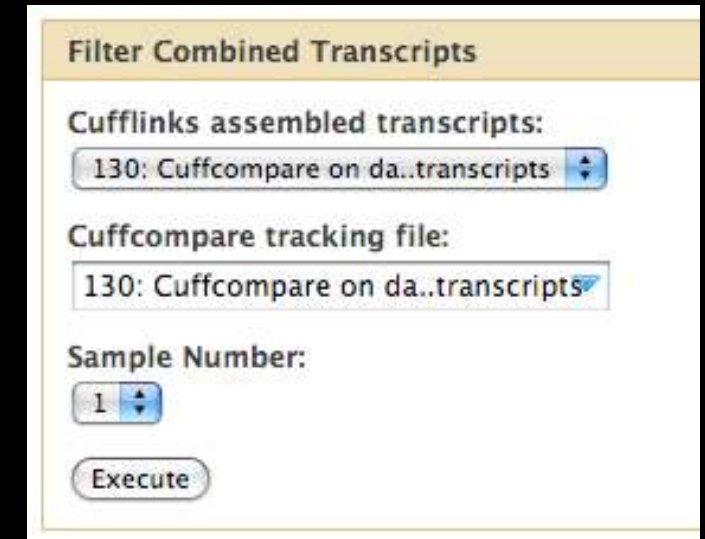
Execute

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



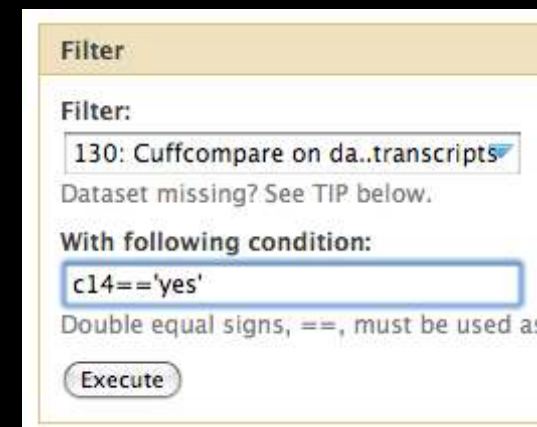
Filter Combined Transcripts

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

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

Sample Number:
1

Execute



Filter

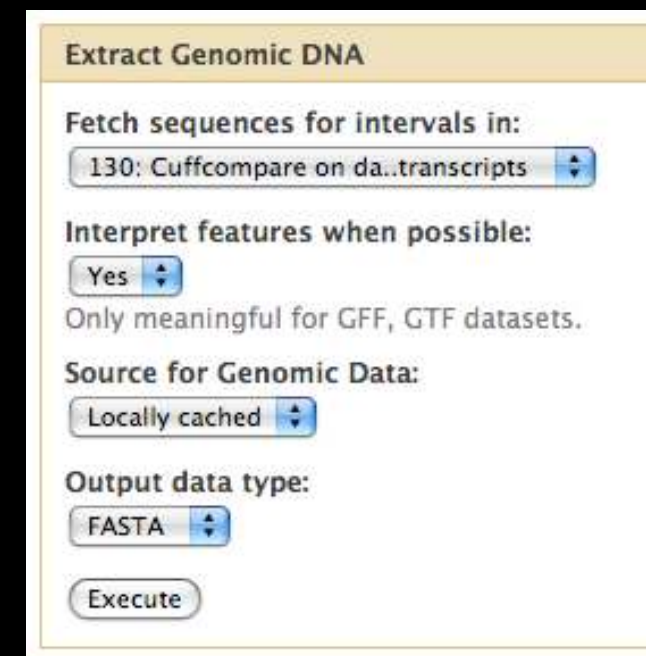
Filter:
130: Cuffcompare on da..transcripts

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

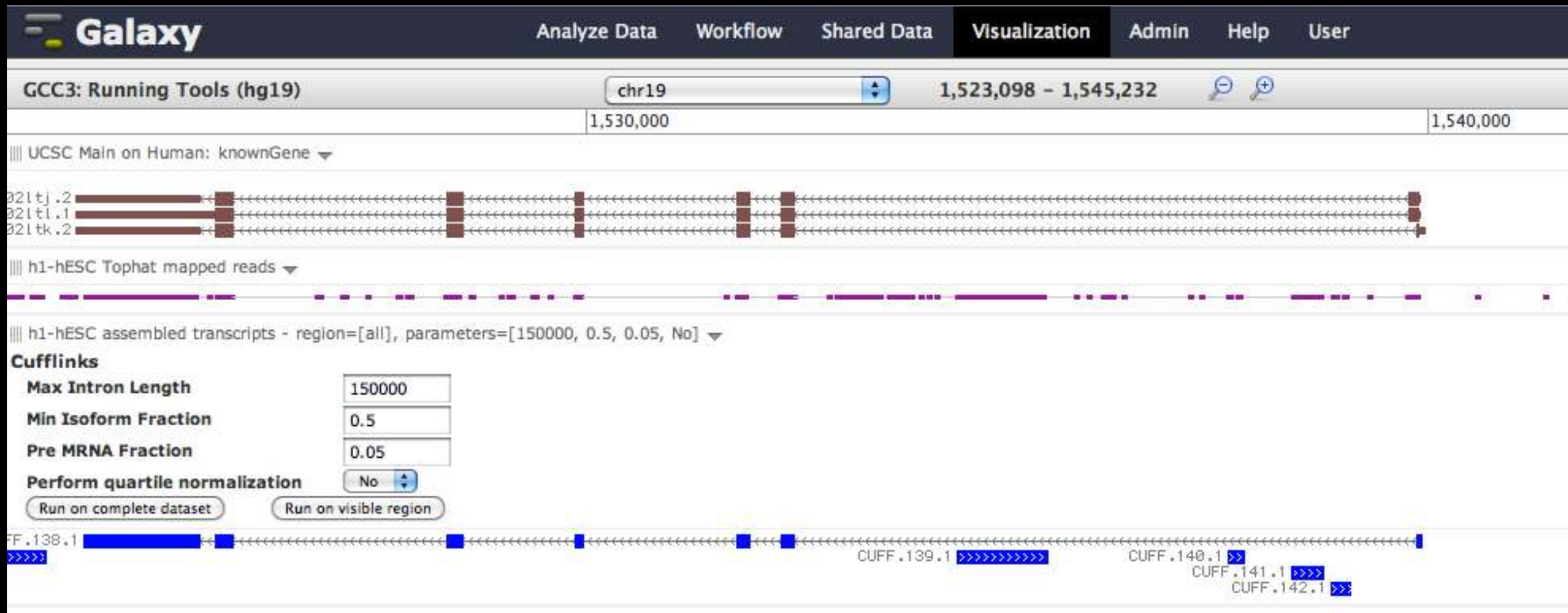
Only meaningful for GFF, GTF datasets.

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

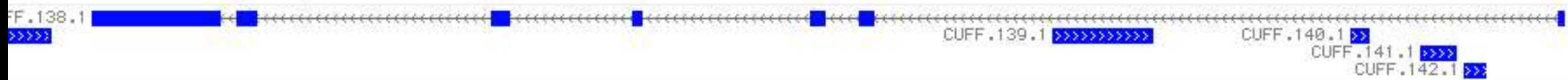
0.05

Perform quartile normalization

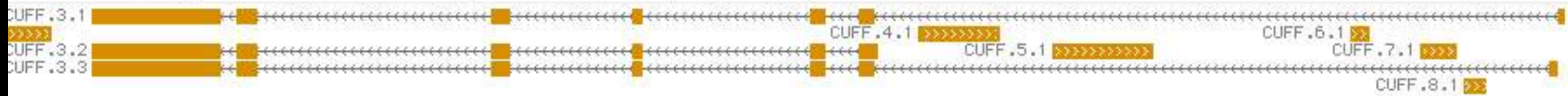
No

Run on complete dataset

Run on visible region



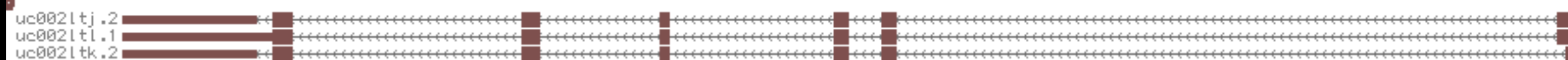
➔ Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No] ▼



1,530,000

1,540

UCSC Main on Human: knownGene ▼



h1-hESC Tophat mapped reads ▼

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

0.001

Perform quartile normalization

No

Run on complete dataset

Run on visible region



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



Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.001, 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

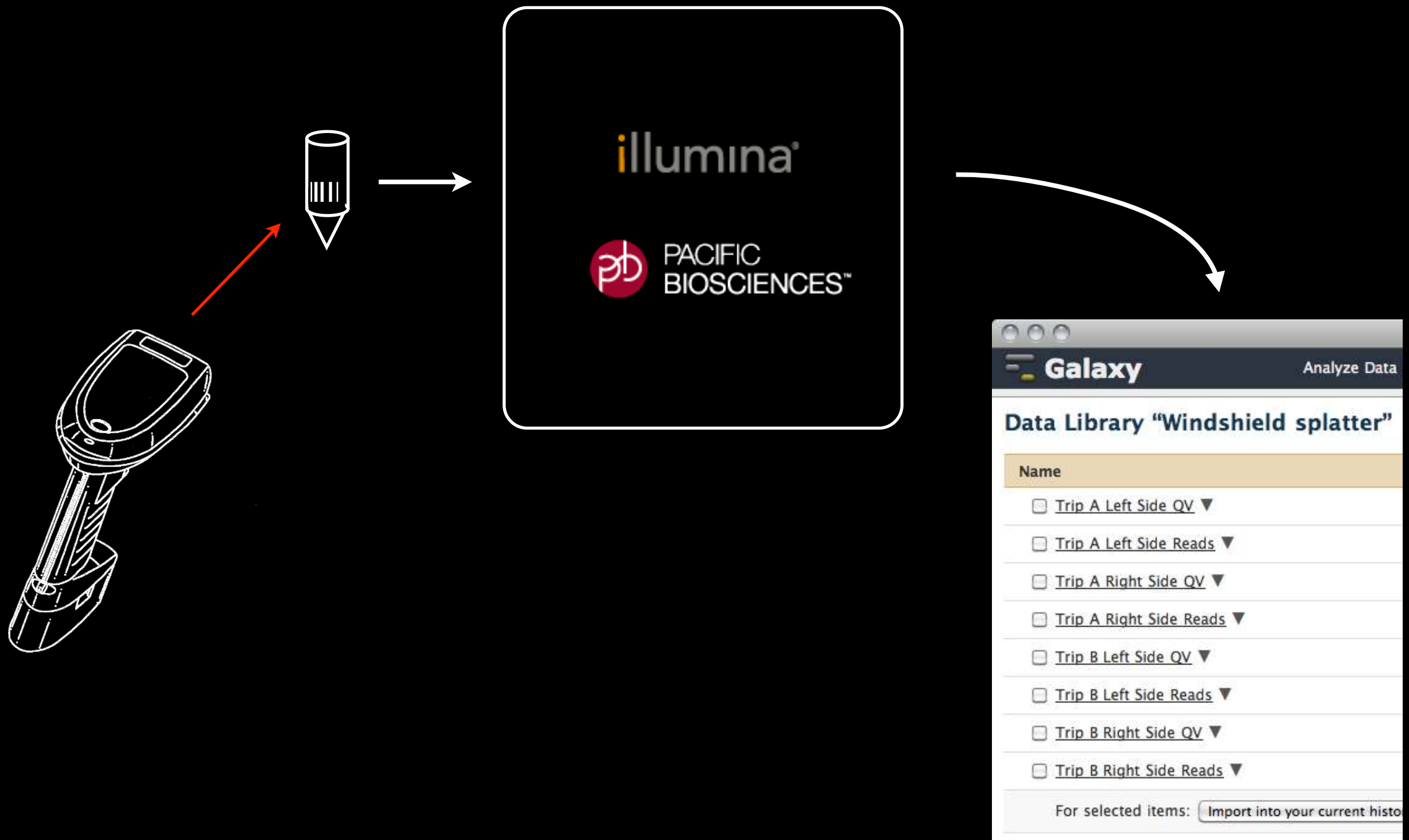
High-throughput Sequencing (HTS) Data

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?

customer



facility manager



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

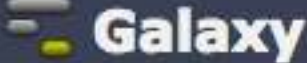
Create and submit a sequencing request

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 DataWorkflowShared DataLabAdminHelpUser

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

Create new form

search  [Advanced Search](#)

☐	Name	Description	Type
☐	Analysis Portal run details ▼		Sample run details template
☐	Atlantic Biosciences Analysis Portal Form ▼		External Service Information Form
☐	Atlantic Biosciences request ▼		Sequencing Request Form
☐	Atlantic Biosciences sample ▼		Sequencing Sample Form

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

[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

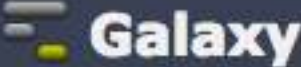
[Edit](#)

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

Layout1

Run type	Read length	Number of Lanes	Alignment target	Processing time	Comments
SelectField: 	SelectField: 	TextField: 	TextField: 	SelectField: 	TextField:
- (optional) Options: SR PE	- (optional) Options: 36 50 75 100	- (optional)	- (optional)	- (optional) Options: Std Rush option3	- (optional)

Configure the Sequencer



Analyze DataWorkflowShared DataLabAdminHelpUser

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 typesCreate new external service

searchAdvanced Search

☐	Name	Description	External Service Type	Last Updated
☐	Analysis Portal service ▼		Atlantic Biosciences Analysis Portal	3 minutes ago

For 0 selected externalservices: DeleteUndo

Edit external service

Name:
Analysis Portal service

Description:

Version:
1

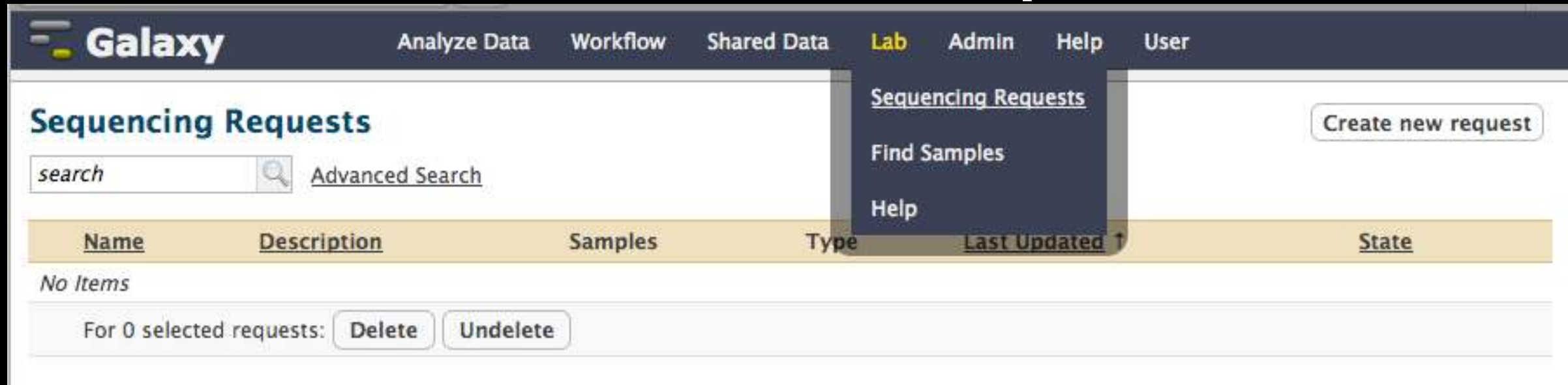
Hostname or IP address:
192.168.56.101
(Required)

User name:
administrator
(Required)

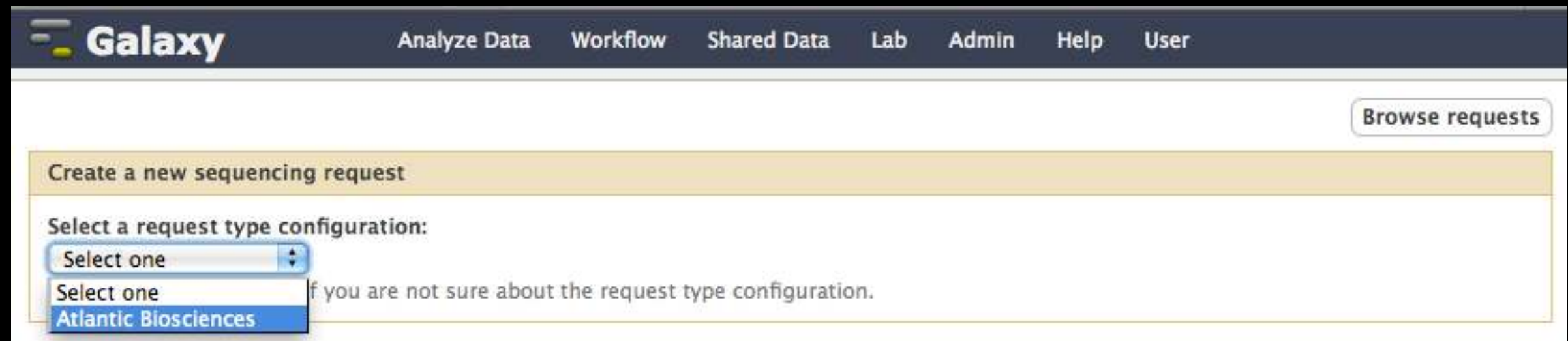
Password:
.....
(Required)

Data directory:

User Creates a Request



The screenshot shows the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Lab, Admin, Help, and User. The 'Lab' menu is currently open, displaying 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 table with columns for Name, Description, Samples, Type, Last Updated, and State is present, but it is empty, displaying 'No Items'. To the right of the table is a button labeled 'Create new request'. Below the table, there is a message 'For 0 selected requests:' followed by 'Delete' and 'Undelete' buttons.



This screenshot displays the 'Create a new sequencing request' form in the Galaxy interface. The form is titled 'Create a new sequencing request' and includes a section for 'Select a request type configuration:'. A dropdown menu is open, showing 'Select one' as the current selection, with 'Atlantic Biosciences' listed as an available option. To the right of the dropdown, there is a note: 'If you are not sure about the request type configuration.' In the top right corner of the form area, there is a button labeled 'Browse requests'.

User Creates a Request

Galaxy

Analyze DataWorkflowShared DataLabAdminHelpUser

Sequencing Requests

search

Advanced Search

NameDescriptionSamples

No Items

For 0 selected requests:

Delete

Undelete

Sequencing Requests

Find Samples

Create new 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

Galaxy

Analyze DataWorkflowSha

Create a new sequencing request

Select a request type configuration:

Select one

Select one

Atlantic Biosciences

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

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 (required)		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

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**

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	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

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](#)

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](#)

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

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

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	

For selected samples:

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

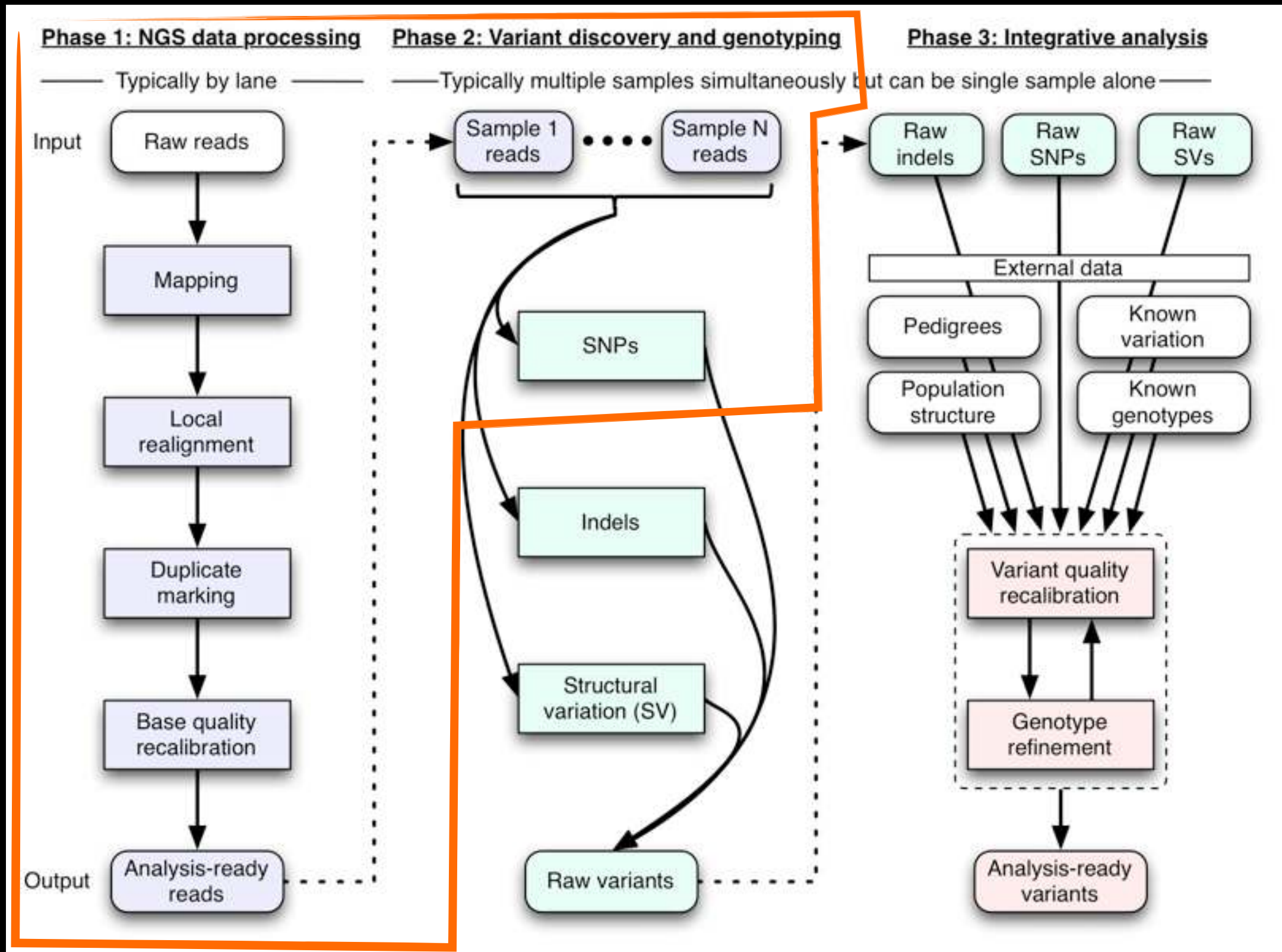
- ✦ <http://usegalaxy.org/u/james/p/exercise-chip-seq>

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, Kernytsky 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.