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

Jeremy Goecks 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 and RNA-seq

## **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 GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

!"\*((((\*\*\*+))%%%++)(%%%%).1\*\*\*-+\*")\*\*55CCF>>>>>CCCCCC65



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**



Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5.

## **Combining Sequences and Qualities**

💳 Galaxy	Analyze Data Workflow Shared Data Visualization Admin Help User	
Tools Options  Com	bine FASTA and QUAL	History Options 🔻
FASTQ splitter on joined paired end reads     FAST	A File:	Combine QUAL and Sequence
FASTQ joiner on paired end     reads	54.fasta 🗘	<u>2: 454.qual</u>
FASTQ Summary Statistics by     Column	ty Score File: 54.qual \$	52 lines format: qual454, database: ?
ROCHE-454 DATA	Quality Score encoding:	
Build base quality distribution		>EYKX4VC01B65GS length=54 xy=0784_1 33 23 34 25 28 28 28 32 23 34 27 4
Select high quality segments     Combine FASTA and QUAL into		>EYXX4VC01BNCSP length=187 xy=0558_ 27 35 26 25 37 28 37 28 25 28 27 36
FASTQ What it	: does ol joins a FASTA file to a Quality Score file, creating a single FASTO block for each read.	22 9 23 19 28 28 28 28 28 26 28 39 32 26 27 37 29 28 26 28 36 28 26 24 38
<u>Convert</u> SOLID output to fastq Specify	ing a set of quality scores is optional; when not provided, the output will be fastqsanger or fastqcssanger (when a	
<u>Compute quality statistics</u> for SOLID data     Use this	is provided) with each quality score being the maximal allowed value (93). s tool, for example, to convert 454-type output to FASTO.	<u>1: 454.fasta</u> ● Ø ⊠
<ul> <li>Draw quality score boxplot for SOLiD data</li> </ul>		format: fasta, database: <u>?</u> Info: uploaded fasta file
GENERIC FASTC GENERIC FASTC	4 xy=0784_1754 region=1 run=R_2007_11_07_16_15_57_ CGCCACCGGAACGAATTCGACTATGCCGAA	a U 🖻 🖉 🔁
BBC:===ABC<%==@6=<<===== Filter FASTQ real @EYKX4VC01BNCSP length=1	==B8=B9E<&6==B;B9<======A8=C: 87 xy=0558 3831 region=1 run=R 2007 11 07 16 15 57	>EYKX4VC01B65GS length=54 xy=0784_1 CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA
score and length CTTACCGGTCACCACCGTGCCTTC	AGGATTGATCGCCAGATCGGTCGGTGCGTCAGGGGGGGGG	>EYKX4VC01BNCSP length=187 xy=0558_ CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG
<ul> <li>FASTQ Quality T</li> <li>FASTQ Quality T</li> <li>GGGGGCTTTGGCCTGTCGTCCGCC</li> </ul>	===Be<<=<<==;=<;e===========;========;===;==	GGTGACATCGCCCACCACGGTACTCACTGGCTGGC CACCACGTTGAGGGTATTCCCCTCGGTTTGTGGCT
FASTO Masker b     REVEX4VC01B8FW0_longtb=9	=E<====E<=============================	
TAAATTTCAAGGAATGCAAATCAG +	GGTCGTGTGTTTAGACTTCGGCTTTAGAGACCTGAATACGTCAAAAAACATAACTTCATGATATCTTGCAGT	
=IC0D=' <b8c9a7===jc2===f @EYKX4VC01BCGYW length=1 GGCCAGCCGGGACAGCGTTGTTGG</b8c9a7===jc2===f 	?*=======P; ===D; =P; *=<==C:==A; ====C:==A; ====C:==A; ====A; ===A; ==A; =A;	
+ =';0<=F=JD2=6=86 <e<9e=ic< td=""><td>/7:=9&lt;=F=;=&lt;&lt;====<le7)=;=<; =:5='C9:IB3"4&lt;IE=E=6&lt;:JC17=F'>;;D&lt;=;JC1==&lt;=F&gt;:LE8-",HA=25==2E&gt;(9)</le7)=;=<;></td><td></td></e<9e=ic<>	/7:=9<=F=;=<<==== <le7)=;=<; =:5='C9:IB3"4&lt;IE=E=6&lt;:JC17=F'>;;D&lt;=;JC1==&lt;=F&gt;:LE8-",HA=25==2E&gt;(9)</le7)=;=<;>	
<pre>%EYKX4VCU1AZXC6 length=1 GGGGGCGTTTGGCCTGTCGTCCGG +</pre>	16 xy=0.292_0.280 region=1 run=R_200/_11_0/_16_15_5/_ CACCTCGCAAGAGCTACAGCAGCGCGGCGGCGGCGATCATCGGCGGGCACGCCGGCCTATATGTCGCCGGAACACACCACCGCCCCCAACGCG	

## Grooming --> Sanger

#### Galaxy

Analyze Data Workflow

Shared Data Visualization Admin Help

User Tools Options v History Options ¬ **FASTO Groomer** NGS TOOLBOX BETA OF File to aroom: 4: FASTO Groomer on • 1 × **Combine QUAL and Sequence** NGS: QC and manipulation 3: Combine FASTA and.. and data 2 📫 data 3 **ILLUMINA DATA** Input FASTQ quality scores type: 18 sequences OUAL on data 1 and data 2 FASTQ Groomer convert Sanger + format: fastgsanger, database: ? between various FASTQ quality 18 sequences Solexa Info: Groomed 18 sanger reads into formats format: fastqsanger, database; ? Illumina 1.3+ sanger reads. + Info: FASTQ splitter on joined paired Sanger Based upon quality and sequence. Combined 18 of 18 sequences with Color Space Sanger end reads quality scores (100.00%). the input data is valid for: sanger Execute FASTQ joiner on paired end 66 Input ASCII range: '!'(33) - 'L'(76) reads Input decimal range: 0 - 43 What it does @EYKX4VC01B65GS length=54 xy=0784 1 FASTQ Summary Statistics by 6 12 🖻 CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA This tool offers several conversions options relating to the FA column ROCHE-454 DATA B8C:===A8C<%==@6=<<====B8=B9E<@6 When using Basic options, the output will be sanger formatte @EYKX4VC01B65GS length=54 xy=0784\_1 Space Sanger). @EYKX4VC01BNCSP length=187 xy=0558 Build base quality distribution CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG When converting, if a quality score falls outside of the target able value (i.e. Select high quality segments 0 4 4 the minimum or maximum). B8C:===A8C<%==@6=<<====B8=B9E<@6 Combine FASTA and OUAL into @EYKX4VC01BNCSP length=187 xy=0558 When converting between Solexa and the other formats, gual scales using FASTO 2: 454.gual O 1 % the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, I CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG ces with 52 lines **AB-SOLID DATA** quality scores, and the Solexa/Illumina FASTQ variants. Nucle ) - i b format: gual454, database: ? Convert SOLiD output to fastq When converting between color space (csSanger) and base/se adapter bases Info: uploaded gual454 file are lost or gained; if gained, the base 'G' is used as the adapter, non cannot convert a color space read to base space if there Compute guality statistics for is no adapter present in the color space sequence. Any masked or ambiguous nucleotides in base space will be converted to SOL iD data 'N's when determining color space encoding. >EYKX4VC01B65GS length=54 xy=0784\_1 Draw guality score boxplot for 33 23 34 25 28 28 28 32 23 34 27 4 SOLID data >EYKX4VC01BNCSP length=187 xy=0558 Quality Score Comparison 27 35 26 25 37 28 37 28 25 28 27 36 **GENERIC FASTO** "#\$%&'()\*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^\_`abcdefghijklmnopgrstuvwxyz{|}~ 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)

Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ\_format

X - Solexa

#### NGS TOOLBOX BETA

#### NGS: QC and manipulation

**ILLUMINA DATA** 

- <u>FASTQ Groomer</u> convert between various FASTQ quality formats
- <u>FASTQ splitter</u> on joined paired end reads
- <u>FASTQ joiner</u> on paired end reads
- <u>FASTQ Summary Statistics</u> by column

#### Box plot in Galaxy \* \* \* ×٠ \* \* Score Value ¥. \* \* \* 10 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 Nucleotide Position

## uality Statistics and Box Plot Tool

#### Graph/Display Data

- <u>Histogram</u> of a numeric column
- <u>Scatterplot</u> of two numeric columns
- <u>Plotting tool</u> for multiple series and graph types
- Boxplot of quality statistics

Quartiles \_\_\_\_\_ Medians \_\_\_\_\_ Outliers \*

## FastQC



## **Read Trimming**

🚾 Galaxy	Analyze Data Workflow Share	ed Data Visualization Admin Help User
Tools Options	FASTQ Trimmer	
MANIPULATION     Filter FASTQ reads by quality	FASTQ File: 2: imported: GM12878ple Dataset	
score and length <ul> <li><u>FASTQ Trimmer</u> by column</li> </ul>	Define Base Offsets as:	
<ul> <li><u>FASTQ Quality Trimmer</u> by sliding window</li> </ul>	Use Absolute for fixed length reads (Illumina, SOLID Use Percentage for variable length reads (Roche/45)	FASTQ Quality Trimmer
FASTQ Masker by quality score	Offset from 5' end:	7: FASTQ Trimmer on data 2
<ul> <li><u>Manipulate FASTQ</u> reads on various attributes</li> </ul>	0 Values start at 0, increasing from the left	Keep reads with zero length:
<ul> <li><u>FASTQ to FASTA</u> converter</li> </ul>	Offset from 3' end:	Trim ends:
FASTQ to Tabular converter     Tabular to FASTQ converter	4 Values start at 0, increasing from the right	5' and 3' 🗘
FASTX-TOOLKIT FOR FASTQ DATA	Keep reads with zero length:	Window size:
<ul> <li><u>Quality format converter</u> (ASCII- Numeric)</li> </ul>	Execute	Step Size:
<u>Compute quality statistics</u>	This tool allows you to trim the ends of reads.	Maximum number of bases to exclude from the window during agg
Draw quality score boxplot	You can specify either absolute or percent-based offs	0
<ul> <li>Draw nucleotides distribution chart</li> </ul>	For example, if you have a read of length 36:	Aggregate action for window:
FASTQ to FASTA converter     Filter by quality	<pre>@Some FASTQ Sanger Read CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA +</pre>	Trim until aggregate score is:
Remove sequencing artifacts	=00.0;B-%78>CBA0>707BBCA4-48%<;;% <b0< td=""><td>&gt;= *</td></b0<>	>= *
	and you cat absolute attrats at 3 and 0:	Quality Score: 0.0 Execute

#### Filter FASTQ

#### FASTQ File:

7: FASTQ Trimmer on data 2

A maximum size less than 1 indicates no limit.

A maximum quality less than 1 indicates no limit.

Maximum number of bases allowed outside of quality range:

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

+

#### Minimum Size:

0	
0	
~	

0

0.0

0.0

0

Execute

#### Maximum Size:

Minimum Quality:

Maximum Quality:

This is paired end data:

Quality Filter on a Range of Bases

Add new Quality Filter on a Range of Bases

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:

$\cap$			
υ.			

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

7: FASTQ Trimmer on data 2	\$
equires groomed data: if your da	ta does roome
atch 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	
anipulate Reads	
Add new Manipulate Reads	
Execute	
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 and RNA-seq

## **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
- Ionger reads: LASTZ

## Metagenomics

+ Megablast

## RNA / gapped-reads mapper

Tophat

## **Commonly Used/Default Parameters**

Lastz		
Align sequencing reads in	c	
•		
Against reference sequen	ces that are:	
locally cached 🛟		
Using reference genome:		
Aedes aegypti: AaegL1		
If your genome of interest i	s not listed, contact the Galaxy team	
Output format:		
SAM ‡		
Lastz settings to use:		
Commonly used		
For most mapping needs u	se Commonly used settings. If you want full control use Full List	
Select mapping mode:		
Roche-454 98% identity		
Roche-454 98% identity	aforence name?	
Roche-454 95% identity	ererence namer:	
Roche-454 90% identity		
Roche-454 85% identity	v this identity (%):	
Illumina 95% identity		
Illumina 85% identity		
Do not report matches ab	ove this identity (%):	
100		
Do not report matches th	at cover less than this percentage of each read:	
0		
Convert lowercase bases	to uppercase.	
Voc A	to uppercase.	
162		
Execute		

-	-		_
-	~		-
с.	-	۰.	<i>c</i> .

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	Full Parameter List
Lastz settings to use: Full Parameter List Commonly used Full Parameter List which strang 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 ‡	Do you want to modify the reference name?:
Perform chaining of HSPs?:	No ‡
	Do not report matches below this identity (%):
	0
	Do not report matches above this identity (%):
30	100
V dran threshold	Do not report matches that cover less than this percentage of each read:
	0
V_drop throshold:	Convert lowercase bases to uppercase:
9370	
Sat the threshold for USPs (unganned extensions scoring lower are discarded):	Execute
3000	What it does
Set the threshold for gapped alignments (gapped extensions scoring lower are discarded): 3000	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.
Involve entropy when filtering HSPs?:	
Do you want to modify the reference name?:	input formats
No 🗘	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 and RNA-seq

## **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
- <u>flagstat</u> provides simple stats on BAM files

🗧 Galaxy		Analyze Data	Workf	low Shared Da	ta Visualiza	ation Admin	Help	User
Tools Options 💌	Indel Analysis							ñ
Fetch Alignments Get Genomic Scores Operate on Genomic Intervals Statistics Graph/Display Data Regional Variation Multiple regression	Select sam file to 54: BAM-to-SA Frequency thres 0.015 Cutoff Execute	o analyze: M on datnvert hold:	ed SAM	;				
Evolution       Metagenomic analyses       Human Genome Variation       EMBOSS       NGS TOOLBOX BETA	What it does Given an input sam threshold. The way can indicate an exa "ACTGCTCGAT"):	n file, this tool p this frequency act match or a r	orovides a of occur nismatch	nalysis of the ind ence is calculated For SAM containi	els. It filters o is different fo ng the followi	ut matches that r deletions and i ng bits of inforn	do not mee nsertions. T nation (assu	t the frequency The CIGAR string's "M" iming the reference
NGS: QC and manipulation NGS: Mapping NGS: SAM Tools NGS: Indel Analysis Filter Indels for SAM Extract indels from SAM Indel Analysis	CHROM POS CIGAR ref 3 2W113M ref 1 2W13M ref 4 4W213M ref 2 2W23M ref 6 3W12W ref 6 3W12W ref 6 3W13M ref 5 5M13M ref 3 2W102M The following total	SEQ TACTTC ACGCT GTTCAAGAT CTCCGG AACCTGG TTCAAT CTCTGTT CTAT CGCTA TGCC s would be calc	culated (t	is is an intermedi	ate step and n	ot output):		
NGS: Peak Calling	POS BASE NUMREAD	S DELPROPCALC	DELPROP	INSPROPSTARTCALC	INSSTARTPROP	INSPROPENDCALC	INSENDPROF	2
NGS: RNA Analysis	1 A 2 A C	2 2/2 1 1/3 2 2/3	1.00 0.33 0.67	 				
RGENETICS	3 C T -	1 1/5 3 3/5 1 1/5	0.20 0.60 0.20		===			

## **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

- <u>Realigner Target Creator</u> for use in local realignment
- <u>Indel Realigner</u> perform local realignment

**BASE RECALIBRATION** 

- <u>Count Covariates</u> on BAM files
- <u>Table Recalibration</u> on BAM files
- <u>Analyze Covariates</u> perform local realignment

GENOTYPING

 <u>Unified Genotyper</u> SNP and indel caller

## **Unified Genotyper**

## Inputs

+ BAM files

# *Lots* of possible parameters

## Output

VCF file(s)

Unified Genoty	ber
Choose the sou	rce for the reference list:
Locally cached	•
Sample BAM file	25
Sample BAM	file 1
BAM file:	
Remove Sam	ble BAM file 1
Add new Sample	BAM file
Using reference	genome:
Mosquito (Aede	s aegypti): AaegL1 🔝
dbSNP reference	e ordered data (ROD):
Selection is Opti	onal 🗘
Binding for refe	rence-ordered datas
Add new Binding	a for reference-ordered data
be called:	nred-scaled confidence threshold at which variants not at trigger track sites should
30.0	
The minimum p be emitted (and	hred-scaled confidence threshold at which variants not at 'trigger' track sites should filtered if less than the calling threshold):
30.0	-
Pasis or Advan	ad CATK options:
Basic	eu GATK options.
Basic or Advand	ed 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 and RNA-seq

## Peak Calling / ChIP-seq analysis

## Punctate binding

transcription factors

## Diffuse binding

- histone modifications
- + Polli

## **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**



Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F, Sung WK. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics. 2010 May 1;26(9):1199-204.

## I have Peaks, now what?

A Intersect First query Intervals to intersect with (Second Query) Overlapping intervals Overlapping pieces of intervals	E Complement Guery Complement
B Subtract First query Intervals to subtract (Second Query) Intervals with no overlap Non-overlapping pieces of intervals	F Cluster Query Find clusters Merge clusters
C Merge Rerged Intervals	
D Concatenate First query Second Query Concatenate	

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     3 UTF       Get Data     5 UTF     codir       Send Data     intro     codir       ENCODE Tools     Lift-Over     codir       Text Manipulation     codir     codir	R 803 R 574 ng exons 2743 ons 13746 r 12499				History Options - <u>2: MACS peak calls (broadPeak)</u> 21,728 regions, format: interval, database: mm9 Info: Comparison of the second seco
Filter and Sort Join, Subtract and Group Join two Queries side by side on a specified field Compare two Queries to find					display at UCSC <u>main test</u>   view in <u>GeneTrack</u>   display at Ensembl <u>Current</u> 1.Chrom 2.6tart 3.End 4 5 6 7 8 9 chr1 4132666 4133002 . 0 . 16.04 14.366 0.0 chr1 4322446 4323079 . 0 . 27.07 26.185 0.0
common or distinct rows  Subtract Whole Query from another query  Group data by a column and perform aggregate operation				P-1	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.
on other columns.    Column Join  Convert Formats  Extract Features					1: UCSC Main on Mouse: refGene      ()      (genome)     28,108 regions, format: bed, database: mm9     Info: UCSC Main on Mouse: refGene (genome)
Fetch Sequences Fetch Alignments Get Genomic Scores Operate on Genomic Intervals Statistics					I display at UCSC main test   view in GeneTrack   display at Ensembl Current           1.chrom 2.Start         3.Bod         4.Name         5         6.1           chr1         134212701         134230065         NM_028778         0         +
<u>Wavelet Analysis</u> Graph/Display Data Regional Variation					chr1 134212701 134230065 NM_001195025 0 + chr1 33510655 33726603 NM_008922 0 - chr1 58714963 58752833 NM_175370 0 - chr1 25124320 25886552 NM_175642 0 - 160945,328960,353082,363947,364951,389516,3932
Multivariate Analysis					

## **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:

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 and RNA-seq

## Transcriptome Analysis (with a reference genome)

## TopHat Cufflinks/compare/diff

#### NGS: RNA Analysis

#### RNA-SEQ

- <u>Tophat</u> Find splice junctions using RNA-seq data
- <u>Cufflinks</u> transcript assembly and FPKM (RPKM) estimates for RNA-Seg data
- <u>Cuffcompare</u> compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- <u>Cuffdiff</u> find significant changes in transcript expression, splicing, and promoter use

FILTERING

 <u>Filter Combined Transcripts</u> using tracking file

Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).
 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:

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

#### Cufflinks

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

Max Intron Length:

Min Isoform Fraction:

Pre MRNA Fraction:

Perform quartile normalization:

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

Use Reference Annotation:

Perform Bias Correction:

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

Reference sequence data:

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

Cuffcompare
GTF file produced by Cufflinks:
21: Cufflinks on datatranscripts
Additional GTF Input Files
Additional GTF Input Files 1
GTF file produced by Cufflinks:
18: Cufflinks on datatranscripts
(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 classificati
Choose the source for the reference list:
Locally cached 文
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:

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 or

Perform quartile normalization:

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):

Execute

## **Next Steps**

Filter

Filter:

c14=='yes'

With following condition:

Double equal signs, ==, must be used as

## Filtering

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

Extract transcript sequences and profile sequences for function





## **Integrating Tools and Visualization**

GCC3: Running Tools (hg19)       chr19       1,523,098 - 1,545,232       P       P         1,530,000       1,540,000       1,540,000         III UCSC Main on Human: knownGene マ       P       P       P         21 ti_2       P       P       P       P         III USSC Main on Human: knownGene マ       P       P       P       P         III USSC Tophat mapped reads マ       P       P       P       P         III h-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] マ       Cuffinks       P         Max Intron Length       150000       0.5       P       P         Min Isoform Fraction       0.5       P       P       P         F. 138.1       Cuff I 41, 1055       Cuff I 41, 1055       Cuff I 41, 1055         T. 138.1       Cuff I 41, 1055       Cuff I 41, 1055       Cuff I 41, 1055	GCC3: Running Tools (hg19)       chr19       1,523,098 - 1,545,232             IUCSC Main on Human: knownGene	Galaxy	Analyze Data	Workflow	Shared Data	Visualization	Admin	Help	User	
1,530,000       1,540,000         III UCSC Main on Human: knownGene ▼         22[t1].2       1         10       1         11       1         11       1         12[t1].2       1         12[t1].2       1         12[t1].2       1         11       1         11       1         11       1         12[t1].2       1         11       1         11       1         11       1         11       1         11       1         11       1         11       1         11       1         11       1         11       1         11       1         11       1         11       1         12	1,530,000       1,540,000         III UCSC Main on Human: knownGene ▼         2215,2       1         2216,2       1         III h1-hESC Tophat mapped reads ▼         III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼         Cufflinks         Max Intron Length       150000         Min Isoform Fraction       0.05         Per MRNA Fraction       0.05         Perform quartile normalization       No €         F. 138.1       CUFF.139.1         CUFF.148.1       000         CUFF.148.1       000	GCC3: Running Tools (hg19)	Chr19		• 1	,523,098 - 1,54	5,232	₽₽		
UCSC Main on Human: knownGen ▼ 221tJ.2 22	UCSC Main on Human: knownGene ▼         221tj.2         221tj.2         21th.1         221tj.2         21th.2         21th.2 </td <td></td> <td>1,530,000</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>1,540,000</td>		1,530,000							1,540,000
221tj.2 2 2 2 2 2 2 2 2 2 2 2 2 2	21 tj.2 21	UCSC Main on Human: knownGene 👻								
III h1-hESC Tophat mapped reads ▼ III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼ Cufflinks Max Intron Length 150000 Min Isoform Fraction 0.5 Per MRNA Fraction 0.05 Perform quartile normalization No ♀ F.138.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1 CUFF.148.1	III h1-hESC Tophat mapped reads ▼ III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼ Cufflinks Max Intron Length 150000 Min Isoform Fraction 0.5 Per MRNA Fraction 0.5 Perform quartile normalization No € Fr. 138.1 CUFF.148.1 CUFF.149.1 CUFF.149.1 CUFF.149.1 CUFF.142.1	221tj.2 221tl.1 221tk.2	•••••••••••••••		***************************************		· · · · · · · · · · · · · · · · · · ·	· · · · · · · · · · · · · · · · · · ·	·····	
III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼         Cufflinks         Max Intron Length         IS0000         Min Isoform Fraction         0.5         Per MRNA Fraction         0.05         Perform quartile normalization         No \$         ** <tr< td=""><td>III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] →          Cufflinks         Max Intron Length         150000         Min Isoform Fraction         0.5         Perform quartile normalization         Run on complete dataset         Run on visible region</td><td>    h1-hESC Tophat mapped reads 👻</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr<>	III h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] →          Cufflinks         Max Intron Length         150000         Min Isoform Fraction         0.5         Perform quartile normalization         Run on complete dataset         Run on visible region	h1-hESC Tophat mapped reads 👻								
CUFF.139.1 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	CUFF.139.1 S>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	IIII n1-nESC assembled transcripts - region=[all], param         Cufflinks         Max Intron Length         Min Isoform Fraction         0.5         Pre MRNA Fraction         Run on complete dataset         FE 138.1	eters=[150000, 0.5, 0.05, M	vol 🗢						
		>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	•••••••••••••••••••••••••••••••••••••••		CUFF.139.1		CUFF.140 C	0.1 >> UFF.141.1 CUFF.1	•••• 42.1 <mark>•••</mark>	•

h1-hESC assembled transcripts - regi	on=[all], parameters=[150	0000, 0.5, 0.05, No] 🔻		
h1-hESC assembled transcripts - regi Jfflinks Max Intron Length	on=[all], parameters=[150	0000, 0.5, 0.05, №] 🛩		
h1-hESC assembled transcripts - regi ufflinks Max Intron Length Min Isoform Fraction	on=[all], parameters=[150	0000, 0.5, 0.05, No] 👻		
h1-hESC assembled transcripts - regi ufflinks Max Intron Length Min Isoform Fraction Pre MRNA Fraction	on=[all], parameters=[150	0000, 0.5, 0.05, No] 🔻		
h1-hESC assembled transcripts - regi ufflinks Max Intron Length Min Isoform Fraction Pre MRNA Fraction Perform quartile normalization Run on complete dataset Run o	on=[all], parameters=[150 150000 0.05 0.05 No 1 visible region	0000, 0.5, 0.05, No] 👻		

CUFF.3.1	<del></del>				****	***********************************	***************************************
>>>>>	203	and and the constructions that is an a	e voneaue von	the concerne concerne la	CUFF	.4.1 >>>>>>:	CUFF.6.1 >>
CUFF.3.2						CUFF.5.1 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	CUFF.7.1 >>>>
CUFF.3.3				*******************		************************************	***************************************
	Contraction of the second	5 (L)		54 D.S.			CUFF.8.1 >>>



## Working to add GATK Unified Genotyper 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 and RNA-seq



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 🛛 🗛	nalyze Data Workflow Shared Data Lab <mark>Admin</mark> Help User
Administration Security • Manage users • Manage groups • Manage roles Data • Manage data libraries	Forms       Create new form         search       Advanced Search         Name       Description         Analysis Portal run details       Sample run details template         Atlantic Biosciences Analysis Portal Form       External Service Information Form
Server  Reload a tool's configuration  Profile memory usage Manage jobs Form Definitions Manage form definitions	Atlantic Biosciences request       Sequencing Request Form         Atlantic Biosciences sample       Sequencing Sample Form         For 0 selected forms:       Delete         Undelete       Velete
Sample Tracking Manage sequencers and external services Manage request types Sequencing requests Find samples	

	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 held	
Save	

## Configurations can be

- + custom-built
- Ioaded from provided configuration files

Form defin	ition "Atlantic	Biosciences sampl	e" (Sequencing Sam	ple Form)	
Layout1					
Run type	Read length	Number of Lanes	Alignment target	Processing time	Comments
SelectField:	SelectField:	TextField:	TextField:	SelectField:	TextField:
- (optional)	- (optional)	(	(	- (optional)	(and an all
Options: SR PE	50 75 100	- (optional)	- (optional)	Rush option3	- (optional)

## **Configure the Sequencer**

- Galaxy Ana	lyze Data Workflow Shared Data Lab <mark>Admin</mark> Help User
Administration	External Services Reload external service types Create new external service
Security	External Service types Create new external service
Manage users	search Advanced
Manage groups	Search
Manage roles	Name         Description         External Service Type         Last Updated
Data	Analysis Portal service  Atlantic Biosciences Analysis Portal 3 minutes ago
Manage data libraries	For 0 selected external convices: Delete Undek Edit external convice
Server	For 0 selected external services: Delete Onder Edit external service
Reload a tool's configuration	Name:
Profile memory usage	Anaiysis Portai service
Manage jobs	Description:
Form Definitions	
<ul> <li>Manage form definitions</li> </ul>	Version:
Sample Tracking	1
<ul> <li>Manage sequencers and external</li> </ul>	Hostname or IP address:
services	192.168.56.101
<ul> <li>Manage request types</li> </ul>	(Required)
Sequencing requests	User name:
Find samples	administrator
	(Required)
	Password:
	•••••
	(Required)
	Data directory:

## **User Creates a Request**

💳 Galaxy	Analyze Data	Workflow	Shared Data	Lab Admin	Help	User
Sequencing search	Advanced Search			<u>Sequencing Req</u> Find Samples Help	<u>uests</u>	Create new request
<u>Name</u>	Description	Samples	Туре	Last U	odated 1	State
No Items						
For 0 selected	requests: Delete Undelete	2				

🗧 Galaxy	Analyze Data	Workflow	Shared Data	Lab	Admin	Help	User	
								Browse requests
Create a new sequencing red	quest							
Select a request type configu Select one Select one Atlantic Biosciences	<b>iration:</b> u are not sure about	t the request	type configuratio	on.				

## **User Creates a Request**

💳 Galaxy	Analyze Data	Workflow	Shared Data	Lab	Admin	Help	User	
Sequencing Request	<b>S</b> anced Search			Seque Find S	encing Req Samples	<u>uests</u>		Create new request
Name Descripti	ion	Samples	Create	a new s	equencin	g reques	t	
No Items For 0 selected requests:	Delete Undelete Analyze Data	Workflow	Select a Atlant Contact Name o My firs (Requir Shar Descrip	a reque tic Biosc t the lab of the E st ChIP-s red) ption	st type col iences manager xperiment req Experir	nfigurati ) if you are ment erformer	on: e not sure abou	t the request type configuratio
Create a new sequencing req	juest		(Option Name	nal)		crionnee	a shing the pro-	
Select a request type configu Select one	ration:	the request t	(Option Scienti dan@ office Penn Wartil Unive Unive Option Save	fic Cont bx.psu.e State Ur k Lab rrsity Par d States e: 867-5 hal) Add	act edu office a liversity k PA 1680 309 samples	address 3	•	

## User Adds a Sample

🗧 Galaxy	Analyze Data	Workflow	Shared Data	Lab	Admin	Help	User			
									Request Act	ions 👻
Add Samples to Seque	encing Request "My first Chl	P-seq Expe	eriment"							
Name State	Data Library	Folder			History			Workflow		
Sample_1	Dan's Sequencing Requests 🛟	ChIP-seq		\$	My own C	hIP-seq	Experiment! 🛟	Dan's Ch	IP-seq Workflov	v 🛟
(required)										
For each sample, select the first and then the desired v	e data library and folder in which yo vorkflow.	ou would like	the run dataset	s depos	ited. To au	tomatica	lly run a workfl	ow on run dat	astets, select a	history
Layout1										
Copy 1 samples f	rom sample None 🛟									
Select the sample from w	hich the new sample should be cop	oied or leave s	selection as No	ne to ad	d a new "g	eneric" sa	ample.			
	(Transl)									
Click the Add sample bu	tton for each new sample and click	the Save but	ton when you h	ave fini	shed addin	g sample	es.			
Import samples from	csv file									
							F	listo	ſV	
								0		
							Wor <u>k</u>	flow	to rur	

## Samples Added, Submit Request

🗧 Gala	axy	Analyze Data	Workflow	Shared Data	Lab	Admin	Help	User		
							Edit	samples	Submit request	Request Actions 🔻
Add Sampl	es to Sequencir	ng Request "My first Chl	P-seq Exp	periment"						
Name	State	Data Library		Folder	History				Workflow	
Sample_1	Unsubmitted	Dan's Sequencing Reques	ts	ChIP-seq	My own	ChIP-seq E	xperime	nt!	Dan's ChIP-see	g Workflow
For each sam first and ther	ple, select the data the desired workfl	library and folder in which yo ow.	ou would like	e the run datase	ets deposi	ted. To aut	omatica	lly run a w	vorkflow on run data	astets, select a history
Layout1										
Copy 1	samples from s	ample None 🗘								
Select the s	ample from which f	the new sample should be cop	or leave	selection as N	one to ad	d a new "ge	eneric" s	ample.		
Click the A	dd sample button f	or each new sample.								
Import sa	amples from csv fi	le								

## Samples enter "New" state

Request A	🗧 Gala	axy		Analyze Data	Workflow S	Shared Data	Lab	Admin	Help	User	
											Request Actions 🗢
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 1         Name       Barcode       State       Data Library       Folder       History       Workflow       Run Datas         Sample_1       New       Dan's Sequencing Requests       ChIP-seq       My own ChIP-seq Experiment!       Dan's ChIP-seq Workflow       0	The sequence of the sequenc	uencing req	uest has l	been submitted.							
Current state: In Progress Description: This is Experiment was performed using the protocol       User: dan@bx.psu.edu Request type: Atlantic Biosciences         More       More         Samples       Edit ::         Name       Barcode       State       Data Library       Folder       History       Workflow       Run Datase         Sample_1       New       Dan's Sequencing Requests       ChIP-seq       My own ChIP-seq Experiment!       Dan's ChIP-seq Workflow       0	Sequencin	g request "	My first	ChIP-seq Experiment"							
User: dan@bx.psu.edu Request type: Atlantic Biosciences ▶ More Samples Request Samples Edit = Name Barcode State Data Library Folder History Workflow Run Datase Sample_1 New Dan's Sequencing Requests ChIP-seq My own ChIP-seq Experiment! Dan's ChIP-seq Workflow 0 ► Lavout1	Current st In Progress Descriptio	ate: i n: eriment was	sperform	ed using the protocol							
Request type:       Atlantic Biosciences         More       More         Samples       Edit         Name       Barcode       State       Data Library       Folder       History       Workflow       Run Datas         Sample_1       New       Dan's Sequencing Requests       ChIP-seq       My own ChIP-seq Experiment!       Dan's ChIP-seq Workflow       0	User: dan@bx.ps	su.edu									
More         Samples       Edit         Name       Barcode       State       Data Library       Folder       History       Workflow       Run Dataset         Sample_1       New       Dan's Sequencing Requests       ChIP-seq       My own ChIP-seq Experiment!       Dan's ChIP-seq Workflow       0	Request ty Atlantic Bio	/ <b>pe:</b> osciences									
Samples       Edit         Name       Barcode       State       Data Library       Folder       History       Workflow       Run Datas         Sample_1       New       Dan's Sequencing Requests       ChIP-seq       My own ChIP-seq Experiment!       Dan's ChIP-seq Workflow       0	More More										
Name         Barcode         State         Data Library         Folder         History         Workflow         Run Data           Sample_1         New         Dan's Sequencing Requests         ChIP-seq         My own ChIP-seq Experiment!         Dan's ChIP-seq Workflow         0	Samples										Edit samples
Sample_1 <u>New Dan's Sequencing Requests</u> ChIP-seq <u>My own ChIP-seq Experiment!</u> <u>Dan's ChIP-seq Workflow</u> 0	Name	Barcode	State	Data Library	Folder	History	é.			Workflow	Run Datasets
Lavout1	Sample_1		New	Dan's Sequencing Request	s ChIP-see	My own	ChIP-seq	Experime	nt!	Dan's ChIP-seq Workflow	0
Lujoura	Layout1										

## Sequencing Facility is informed of Request

Galaxy	Analyze Data	Workflow	Shared Data	Lab	Admin H	elp User		
Administration Security Manage users	Sequencin search	Create new request						
Manage groups     Manage roles	Name Name	<u>Des</u>	cription S	amples	Туре	Last Updated †	<u>State</u>	User
Data <u>Manage data libraries</u> Server	My first ChIP-sec Experime	✓ Expension of this expension of the	eriment was ormed <u>1</u> g the ocol	L	<u>Atlantic</u> <u>Biosciences</u>	26 minutes ago	In Progress	dan@bx.psu.edu
<ul> <li><u>Reload a tool's configuration</u></li> <li><u>Profile memory usage</u></li> </ul>	new requ	iest 🔻	1	1	Atlantic Biosciences	3 days ago	Complete	customer@corp.com
<ul> <li>Manage jobs</li> <li>Form Definitions</li> </ul>	some experime test	ent desc	st <u>1</u> cription <u>1</u>	<u>1</u>	Atlantic Biosciences	3 days ago	Complete	customer@corp.com
Manage form definitions     Sample Tracking     Manage sequencers and external     condees	For 0 sel	ected requests	: Delete U	Indelete	)			
Manage request types     Sequencing requests								
<ul> <li>Find samples</li> </ul>								

## **Sequencing Facility Receives Samples**

Edit Current Samples of Sequencing Request "My first ChIP-seq Experiment"											
	Name	Barcode	State	Data Library	Folder	History	Workflow	Run Data	asets Delete		
	Sample_1 (required)		<u>New</u>	Dan's Sequencing Requests 🛟	ChIP-seq 🛟	My own ChIP-seq Experim	ent! 🗘 Dan's ChIP-seq	Workflow 🛟 0	*		
For	selected samples:	Select one		\$							
For ea	ch sample, select	the data library and	folder	in which you would like the run dat	tasets deposited. To automatically r	run a workflow on run datast	ets, select a history first and	d then the desired workfl	.wc		
Layout1											
Save Cancel Click the Save button when you have finished editing the samples											
Create new requ											
searcl		Advanced Sear	<u>ch</u>								
	<u>Name</u>			Description		Samples	Туре	Last Updated 1	<u>State</u>		
	My first ChIP-seq	Experiment 🔻		This is Experiment was perfor	med using the protocol	1	Atlantic Biosciences	35 minutes ago	Complete		
	For 0 selected req	uests: Delete	Unde	lete							

## 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/nextgeneration-sequencing-information-managementand-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

💳 Galaxy	Analyze Data	Workflow	Data Libraries	Lab	Visualization	Admin	Help	User
Samples • Define samples and services • Submit samples as a project • View projects Sequencing • Queues • Runs	Samples TJa3 Sample Copy	4		Lane 1 TJa1 S Lane 2 PhiX c Lane 3 BCa1 Lane 4 Lane 5 Lane 6 Lane 7 Lane 8	Sample 2			

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 and RNA-seq











Dannon Baker



**Dave Clements** 



Jeremy Goecks



Dan Blankenberg



Jennifer Jackson



Nate Coraor



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

http://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysisexercise

- Shared Data --> Published Histories --> Import
- start Tophat mapping first (second section), then look at QC (first section)

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

http://usegalaxy.org/u/jeremy/p/usc-exercise-clusters