Using Galaxy for High-throughput Sequencing (HTS) Analysis and Visualization

Dan Blankenberg
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 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>>>>>>CCCCCCC655
```

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

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

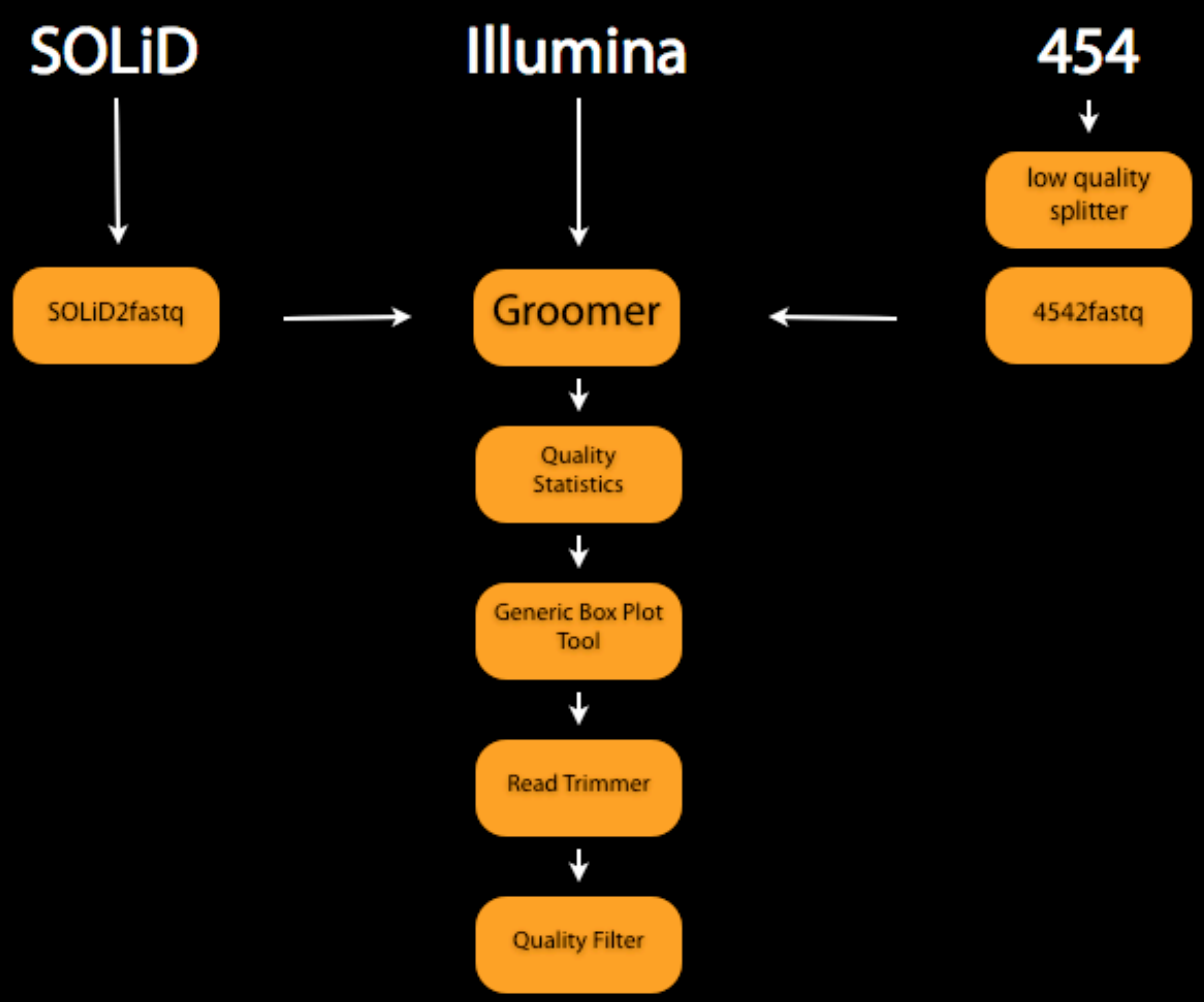
High-throughput Sequencing (HTS) Data

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 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

- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ read score and length
- FASTQ Trimmer
- FASTQ Quality Tr sliding window
- FASTQ Masker b

Combine FASTA and QUAL

FASTA File: 1: 454.fasta

Quality Score File: 2: 454.qual

Force Quality Score encoding: ASCII

Execute

What it does

This tool joins a FASTA file to a Quality Score file, creating a single FASTQ block for each read.

Specifying a set of quality scores is optional; when not provided, the output will be fastqsanger or fastqcssanger (when a csfasta is provided) with each quality score being the maximal allowed value (93).

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

History Options

Combine QUAL and Sequence

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

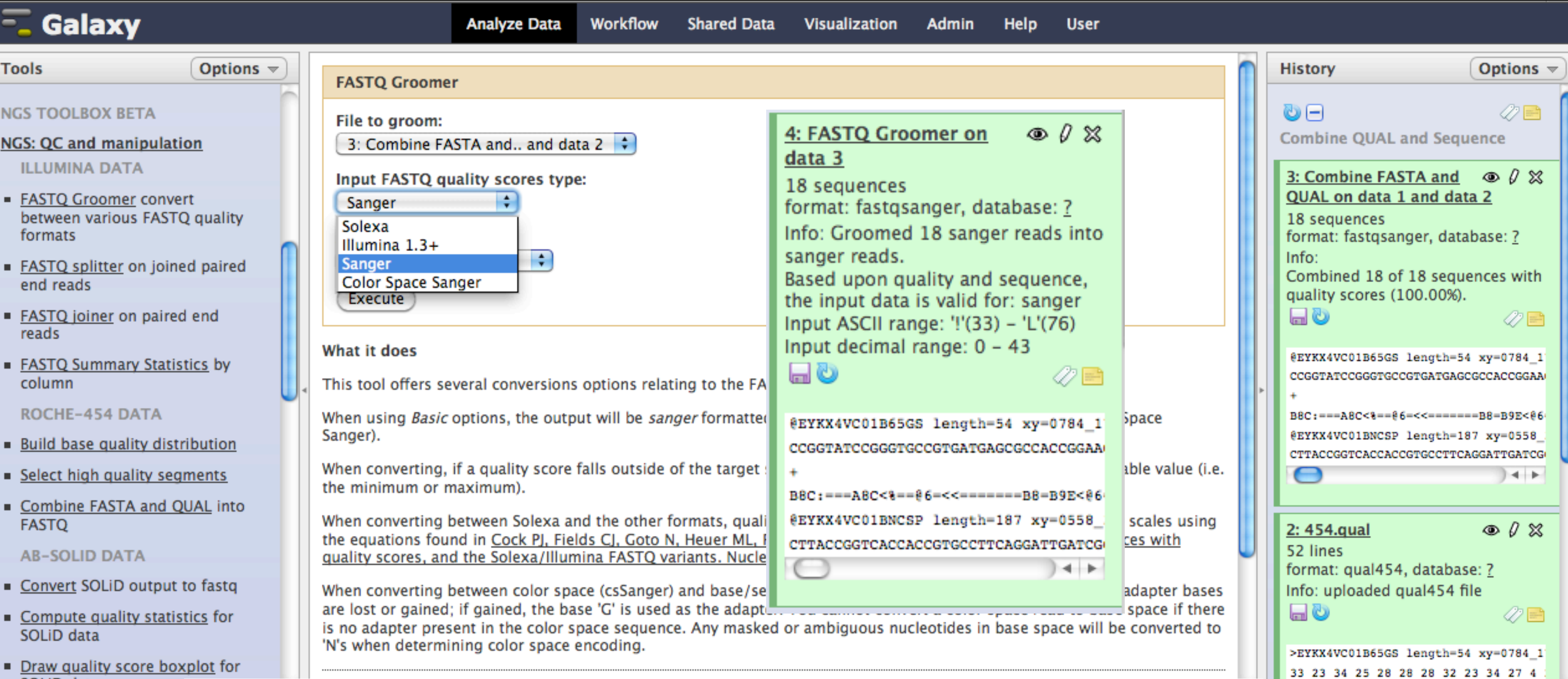
```
>EYKX4VC01B65GS length=54 xy=0784_1
33 23 34 25 28 28 28 32 23 34 27 4
>EYKX4VC01BNCSP length=187 xy=0558_
27 35 26 25 37 28 37 28 25 28 27 36
22 9 23 19 28 28 28 28 26 28 39 32
26 27 37 29 28 26 28 36 28 26 24 38
```

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

```
>EYKX4VC01B65GS length=54 xy=0784_1
CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA
>EYKX4VC01BNCSP length=187 xy=0558_
CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG
GGTGACATCGCCCACCACGGTACTCACTGGCTGGC
CACCACGTTGAGGGTATTCCCCTCGGTTTGTGGCT
```

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

Grooming --> Sanger



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

What it does

This tool offers several conversions options relating to the FA

When using *Basic* options, the output will be *sanger* formatte … Space Sanger).

When converting, if a quality score falls outside of the target … able value (i.e. the minimum or maximum).

When converting between Solexa and the other formats, quali … scales using the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, F … ces with quality scores, and the Solexa/Illumina FASTQ variants. Nucle

When converting between color space (csSanger) and base/se … adapter bases are lost or gained; if gained, the base 'G' is used as the adapt … 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
+
B8C:===A8C<%==@6=<<=======B8=B9E<@6
@EYKX4VC01BNCSP length=187 xy=0558_
CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG
```

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%).

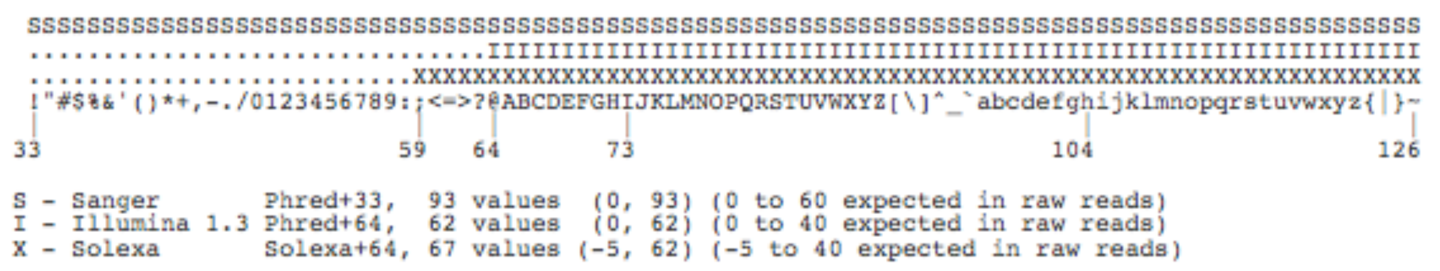
```
@EYKX4VC01B65GS length=54 xy=0784_1
CCGGTATCCGGGTGCCGTGATGAGCGCCACCGGAA
+
B8C:===A8C<%==@6=<<=======B8=B9E<@6
@EYKX4VC01BNCSP length=187 xy=0558_
CTTACCGGTCACCACCGTGCCTTCAGGATTGATCG
```

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



```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`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

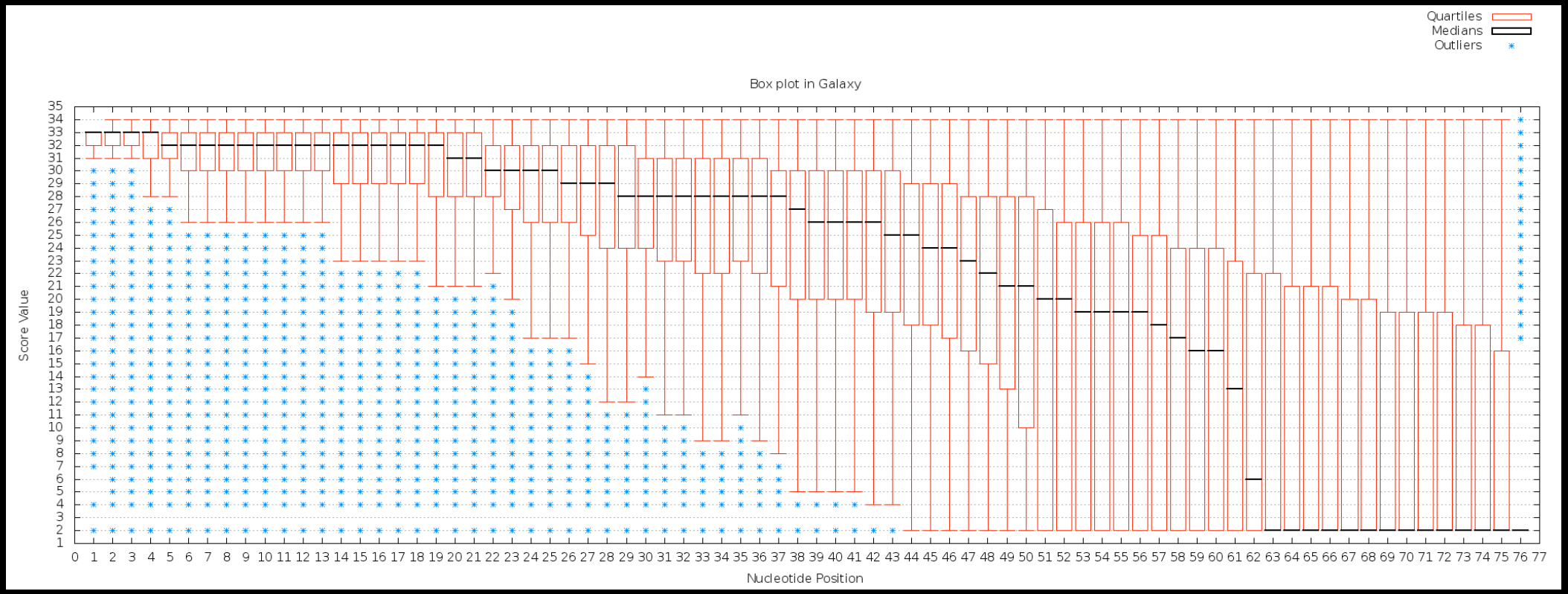
NGS: QC and manipulation

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



FastQC

Galaxy

main.g2.bx.psu.edu

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Tools | Options

fastqc

NGS: QC and manipulation

ROCHE-454 DATA

- Combine FASTA and QUAL into FASTQ

FASTQ QC

- Fastqc: Fastqc QC using FastQC from Babraham

Workflows

FastQC Report

Mon 20 Jun 2011

dataset_1750787.dat

Summary

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

Basic Statistics

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

History | Options

imported: Joe practice 6-14-11 70.5 Mb

120: FastQC.html
11.6 Kb
format: html, database: hg18
HTML file

119: Cuffdiff on data 11, data 13, and data 29: transcript FPKM tracking

118: Cuffdiff on data 11, data 13, and data 29: transcript differential expression testing

117: Cuffdiff on data 11, data 13, and data 29: gene FPKM tracking

116: Cuffdiff on data 11, data 13, and data 29: gene differential expression testing

115: Cuffdiff on data 11, data 13, and data 29: TSS groups FPKM tracking

114: Cuffdiff on data 11, data 13, and data 29: TSS groups differential expression testing

113: Cuffdiff on data 11, data 13, and data 29: CDS FPKM tracking

112: Cuffdiff on data 11, data 13, and data 29: CDS FPKM differential expression testing

111: Cuffdiff on data 11, data 13, and data 29: CDS overloading diffential expression testing

Read Trimming

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Tools — Options

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts

FASTQ Trimmer

FASTQ File: 2: imported: GM12878..ple Dataset

Define Base Offsets as: Absolute Values

Use Absolute for fixed length reads (Illumina, SOLiD
Use Percentage for variable length reads (Roche/45

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 offs
trimmed. When using the percent-based method, offs

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

```
@Some FASTQ Sanger Read
CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA
+
=@@.@;B-%?8>CBA@>7@7BBCA4-48%<;;%<B@
```

And you set absolute offsets of 2 and 9:

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 aggr

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 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
- longer reads: LASTZ

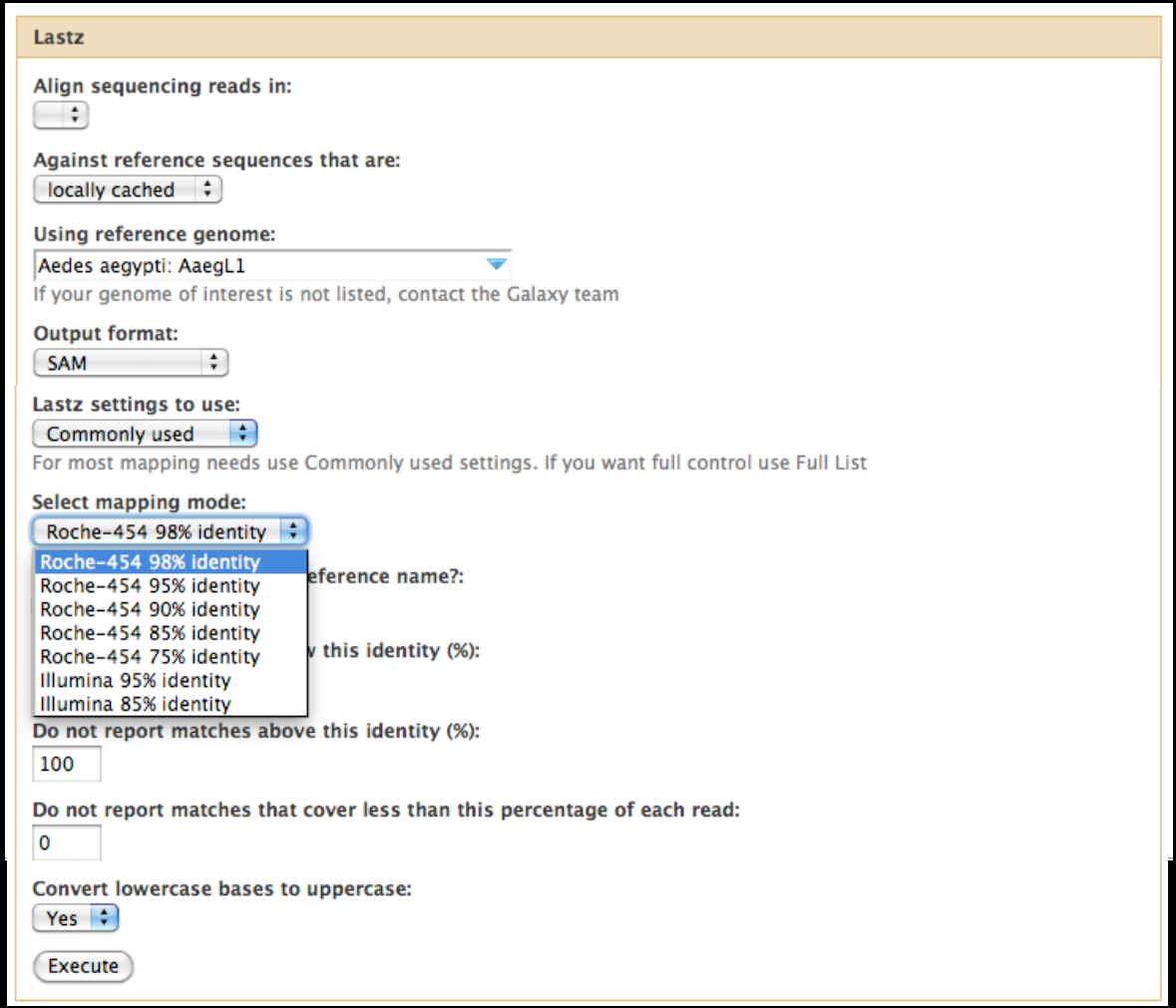
Metagenomics

- Megablast

RNA / gapped-reads mapper

- Tophat

Commonly Used/Default Parameters



Full Parameter List

Lastz

Align sequencing reads in:
53: FASTQ to FASTA on data 7

Against reference sequences that are:
locally cached

Using reference genome:
Aedes aegypti: AaegL1
If your genome of interest is not listed, contact the Galaxy team

Output format:
SAM

Lastz settings to use:
Full Parameter List
Commonly used
Full Parameter List
use Commonly used settings. If you want full control use Full List

Which strand to search?:
Both

Select seeding settings:
Seed hits require a 19 bp word with matches i
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

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 exercises: ChIP-seq and RNA-seq

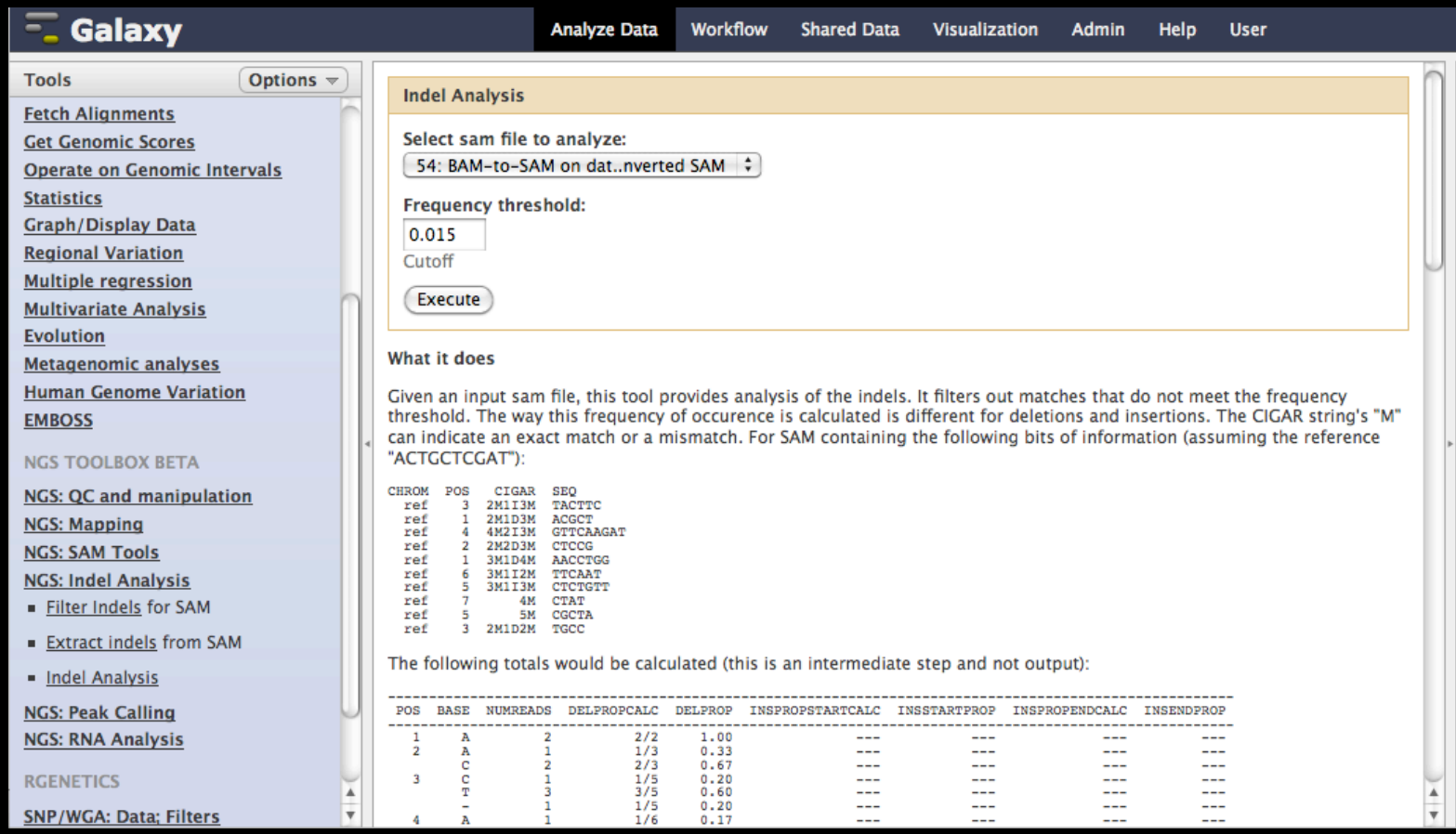
SNPs & INDELs

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
- flagstat provides simple stats on BAM files



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

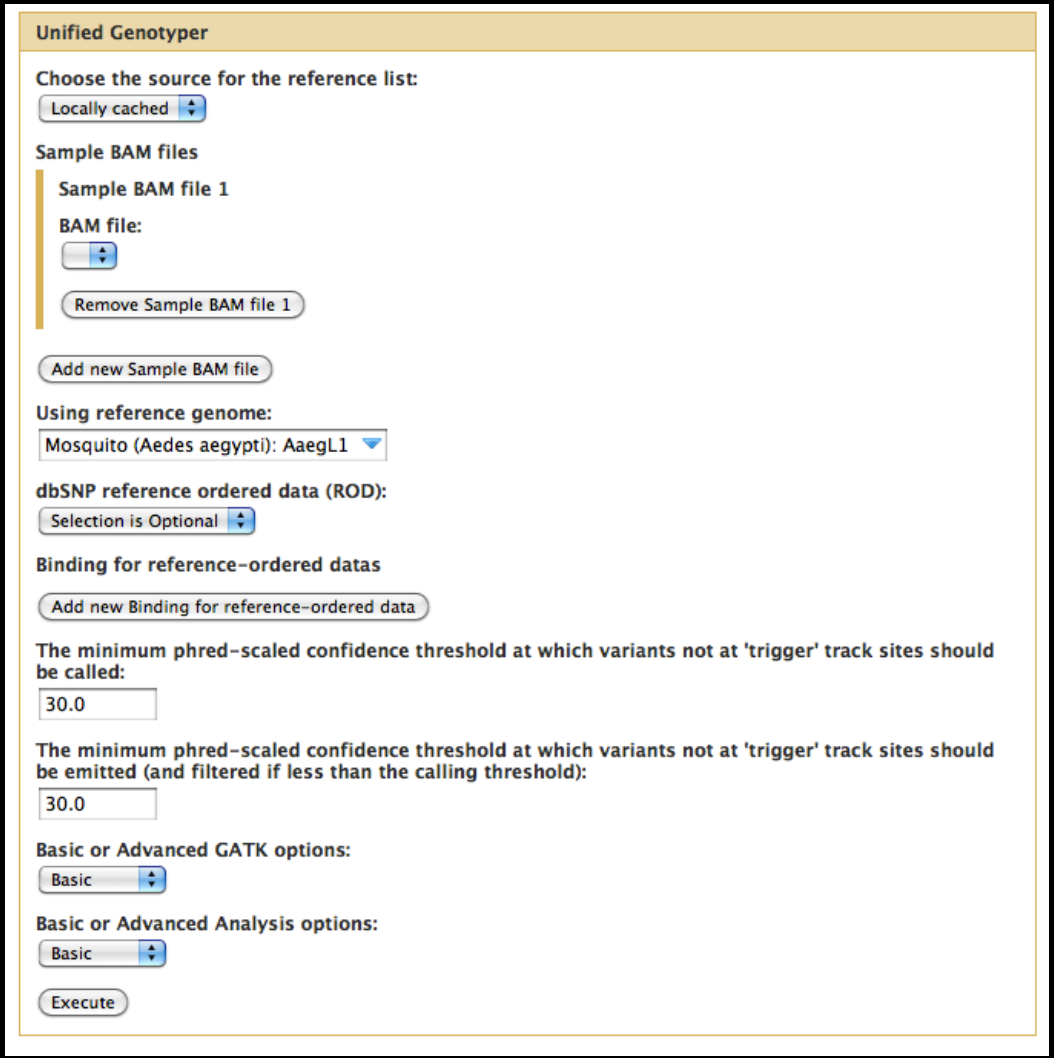
Inputs

- BAM files

Lots of possible parameters

Output

- VCF file(s)



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 exercises: ChIP-seq and RNA-seq

Peak Calling / ChIP-seq analysis

Punctate binding

- transcription factors

Diffuse binding

- histone modifications
- PolII

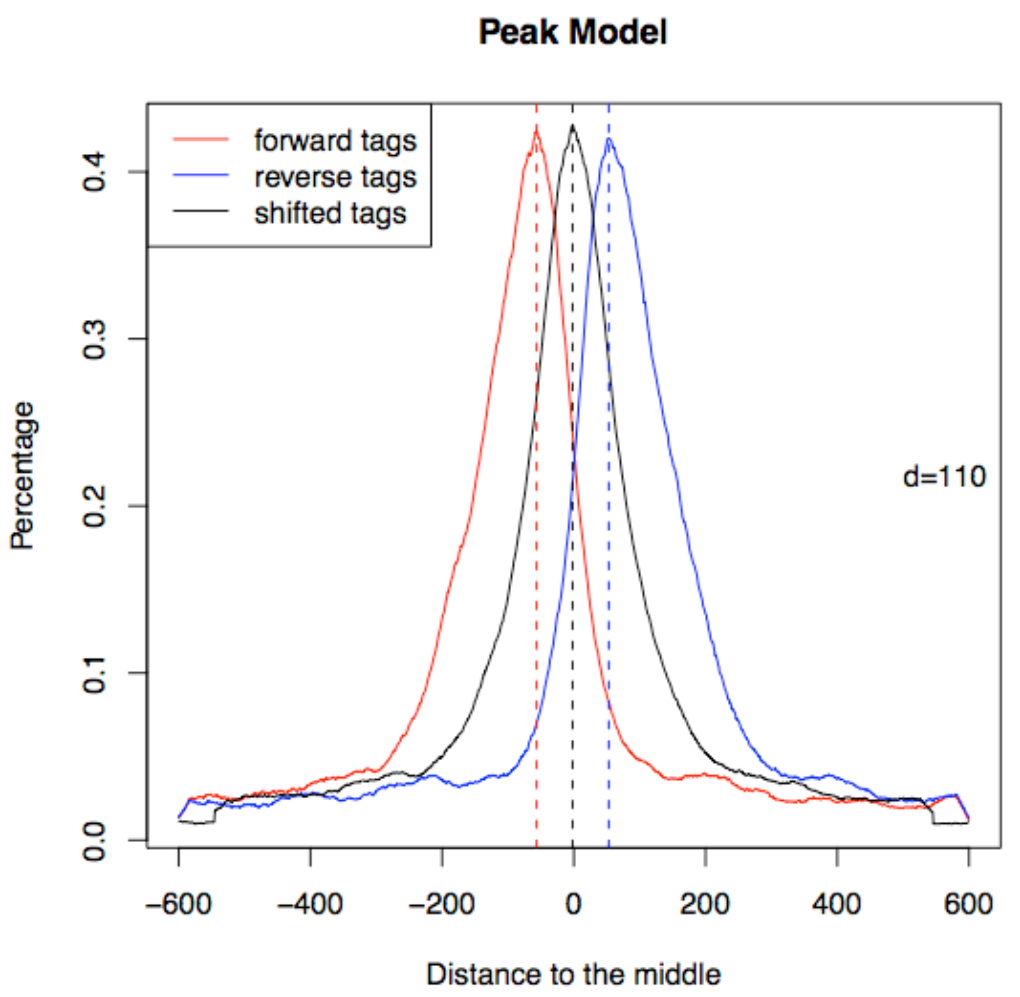
Punctate Binding --> MACS

Inputs

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

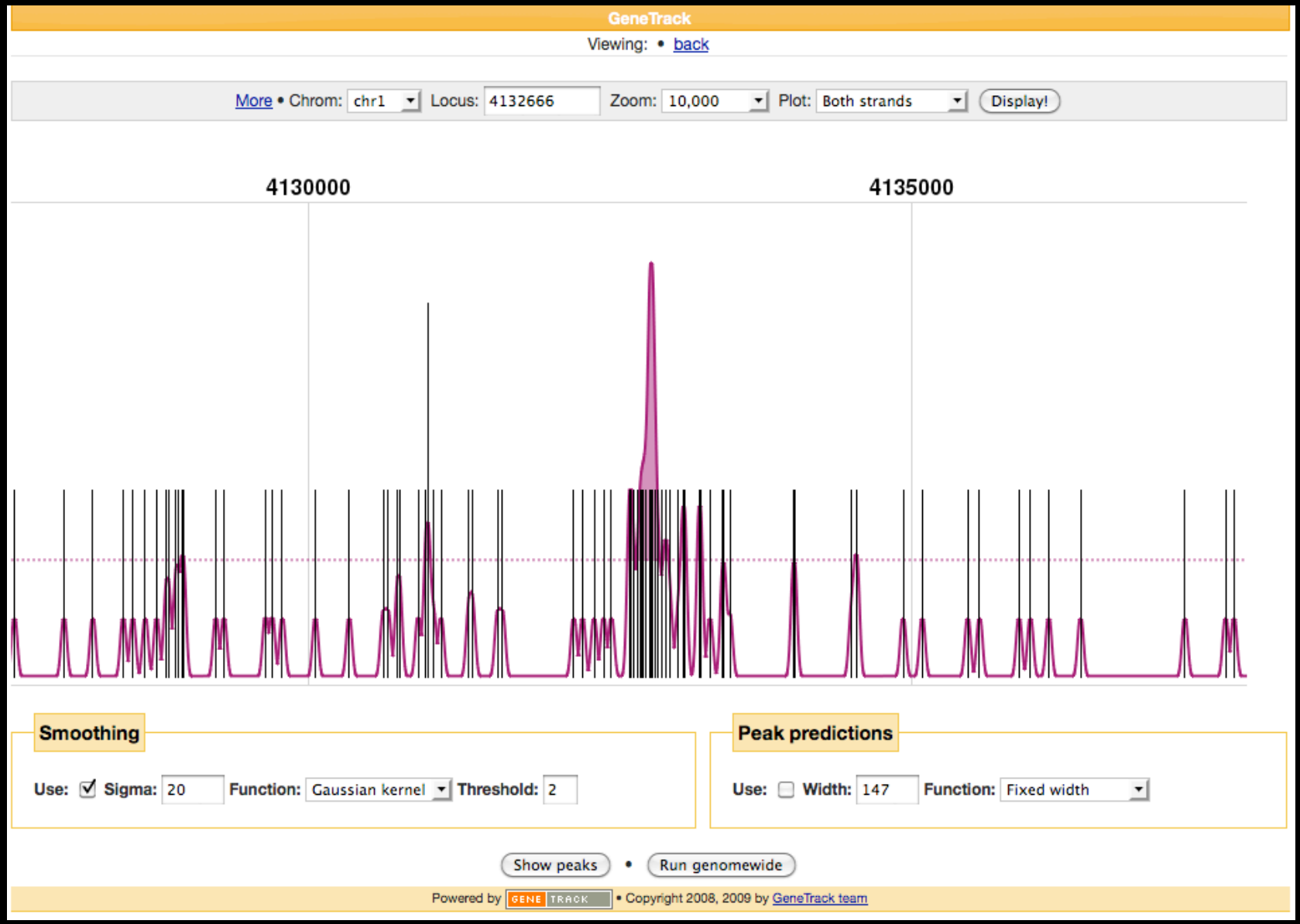
Outputs

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



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

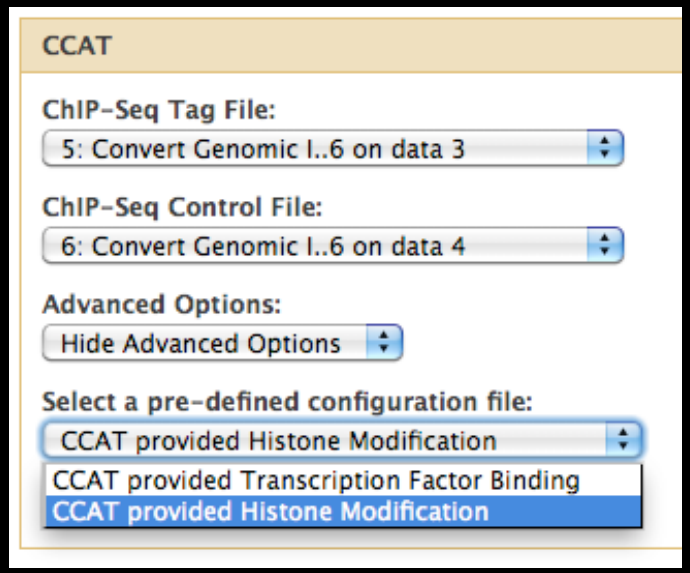
MACS --> GeneTrack

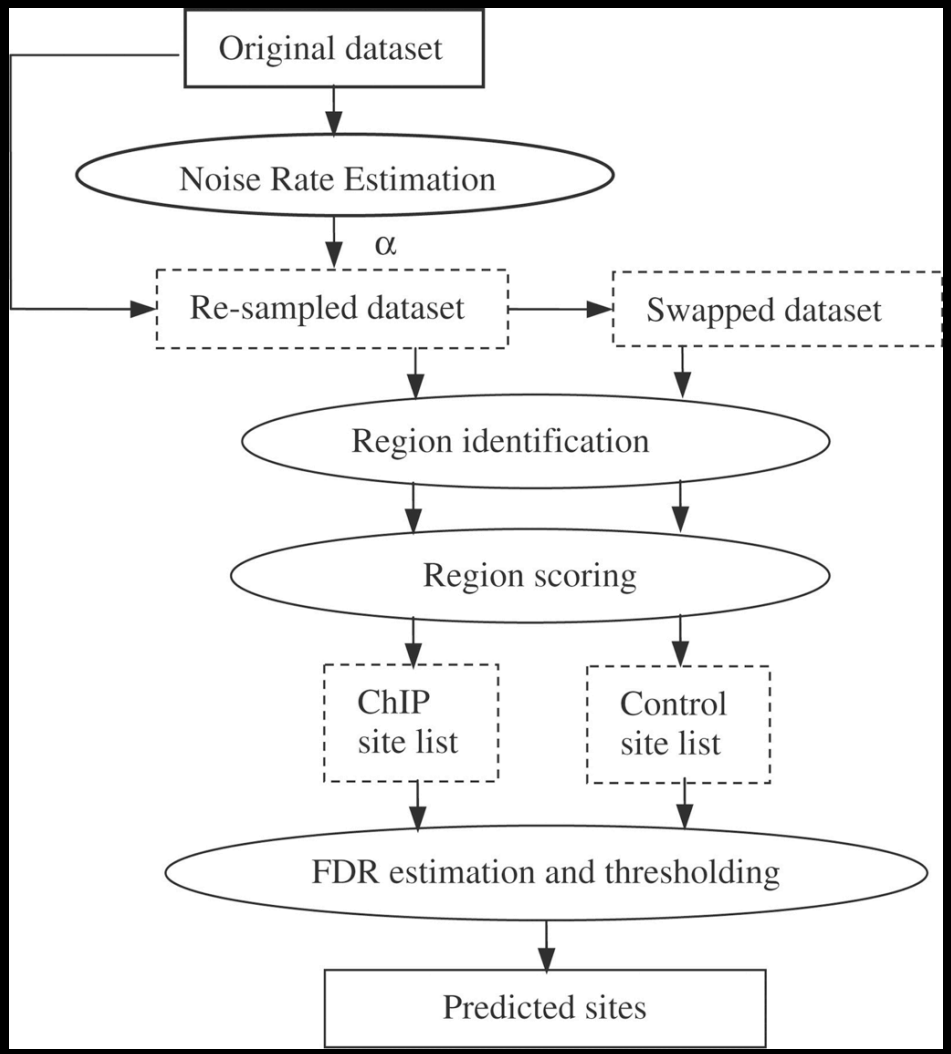


Albert I, Wachi S, Jiang C, Pugh BF. GeneTrack--a genomic data processing and visualization framework. Bioinformatics. 2008 May 15;24(10):1305-6. Epub 2008 Apr 3.

Diffuse Binding

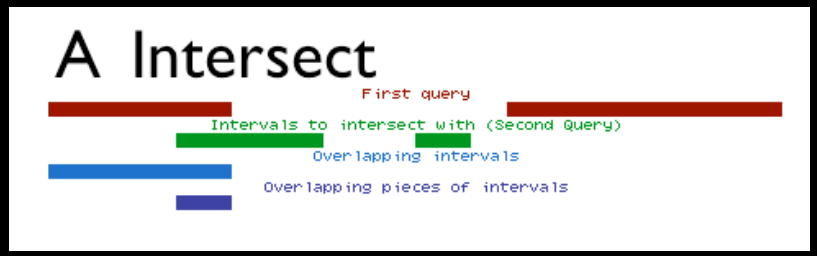
CCAT (Control-based ChIP-seq Analysis Tool)

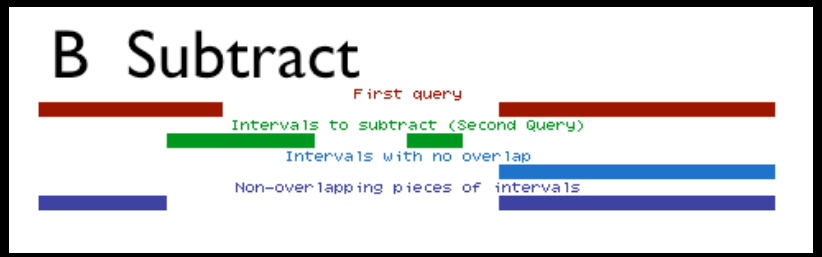


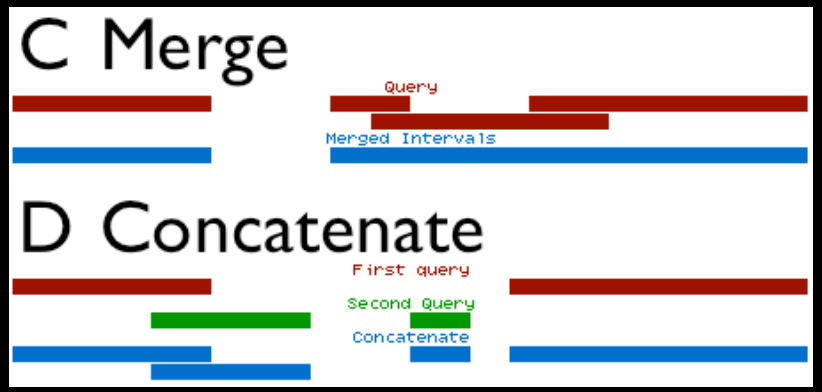


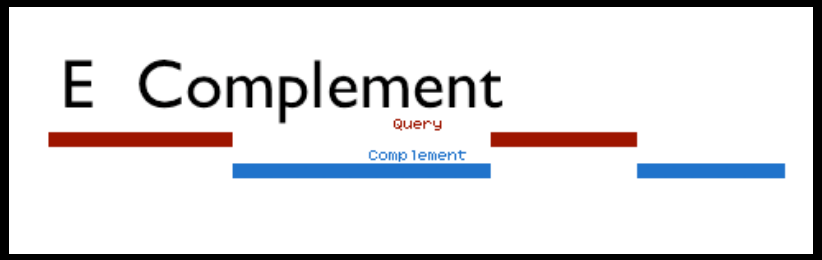
Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F, Sung WK. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics. 2010 May 1;26(9):1199-204.

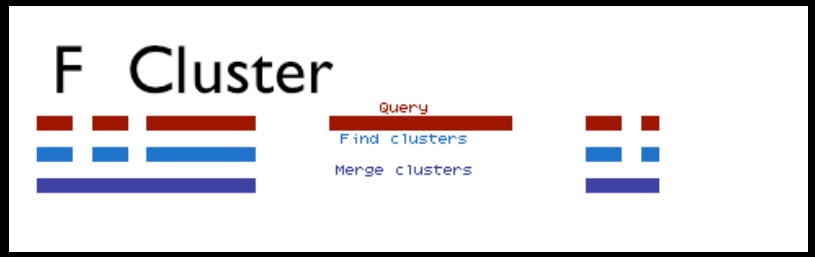
I have Peaks, now what?











Compare to other annotations using interval operations

Secondary Analysis

A simple goal: determine number of peaks that overlap a) coding exons, b) 5-UTRs, c) 3-UTRs, d) introns and d) other regions

Get Data

Import Peak Call data

Retrieve Gene location data from external data resource

Extract exon and intron data from Gene Data (Gene BED To Exon/Intron/Codon BED expander x4)

Create an Identifier column for each exon type (Add column x4)

Create a single file containing the 4 types (Concatenate)

Complement the exon/intron intervals

Force complemented file to match format of Gene BED expander output (convert to BED6)

Create an Identifier column for the 'other' type (Add column)

Concatenate the exons/introns and other files

Determine which Peaks overlap the region types (Join)

Calculate counts for each region type (Group)

Secondary Analysis

Galaxy

Analyze Data | Workflow | Shared Data | Admin | Help | User

Tools — Options

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
  - Join two Queries side by side on a specified field
  - Compare two Queries to find common or distinct rows
  - Subtract Whole Query from another query
  - Group data by a column and perform aggregate operation on other columns.
  - Column Join
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution

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

History — Options

2: MACS peak calls (broadPeak)

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

Info:

| display at UCSC main test | view in GeneTrack | display at Ensembl Current

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

1: UCSC Main on Mouse: refGene (genome)

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

Info: UCSC Main on Mouse: refGene (genome)

| display at UCSC main test | view in GeneTrack | display at Ensembl Current

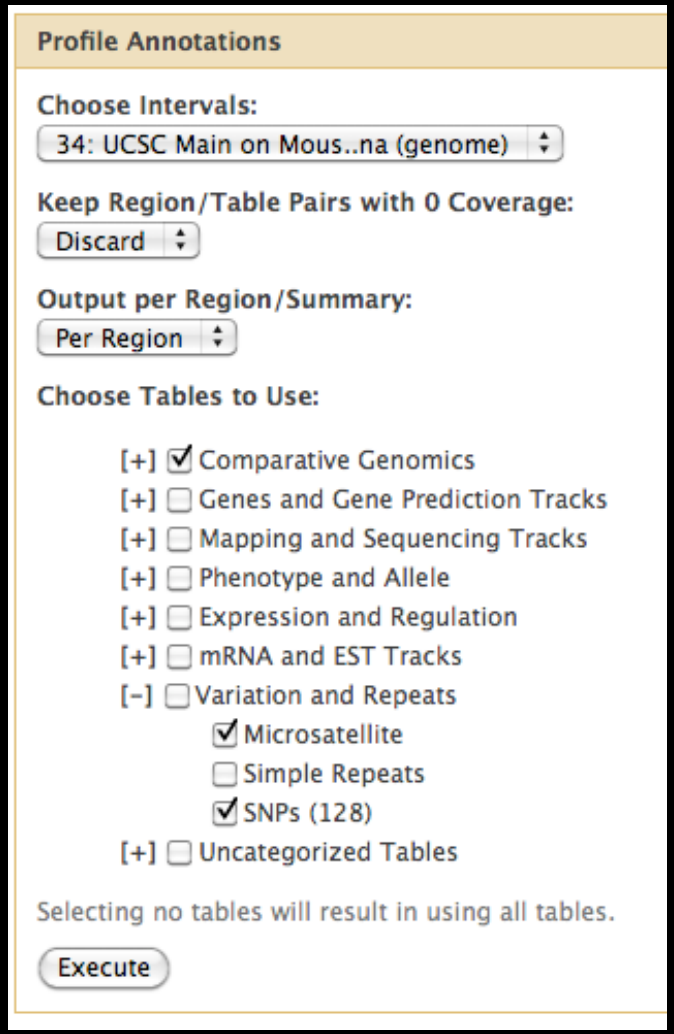
1.Chrom	2.Start	3.End	4.Name	5	6.
chr1	134212701	134230065	NM_028778	0	+
chr1	134212701	134230065	NM_001195025	0	+
chr1	33510655	33726603	NM_008922	0	-
chr1	58714963	58752833	NM_175370	0	-
chr1	25124320	25886552	NM_175642	0	-

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

Annotation Profiler

One click to determine base coverage of the interval (or set of intervals) by a set of features (tables) available from UCSC

galGal3, mm8, panTro2, rn4, canFam2, hg18, hg19, mm9, rheMac2



Overview

High-throughput Sequencing (HTS) Data

Using Galaxy to Analyze HTS Data

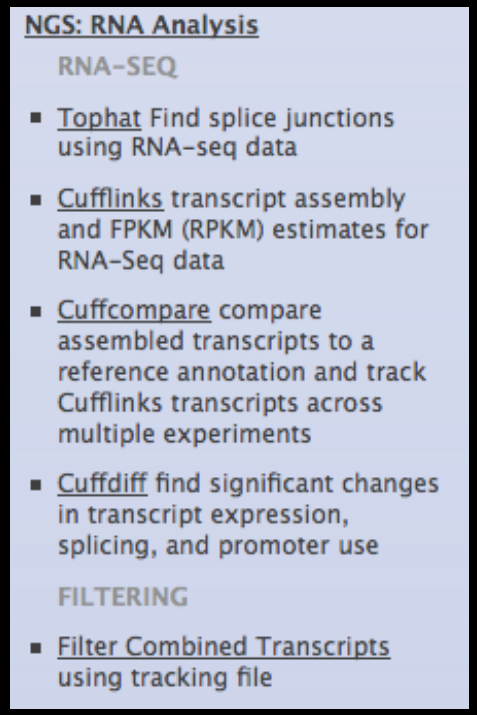
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

Galaxy exercises: ChIP-seq and RNA-seq

Transcriptome Analysis (with a reference genome)

TopHat

Cufflinks/compare/diff



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

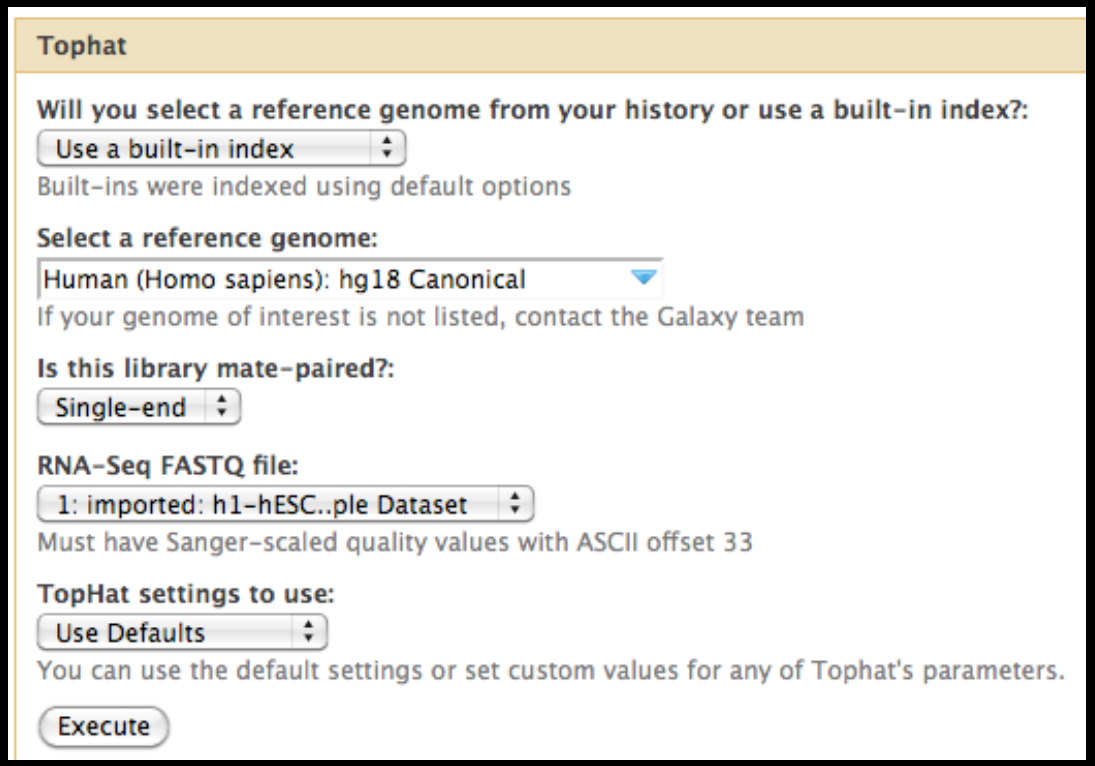
TopHat

Map RNA (FASTQ) to a reference Genome

- gapped mapper

Outputs

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



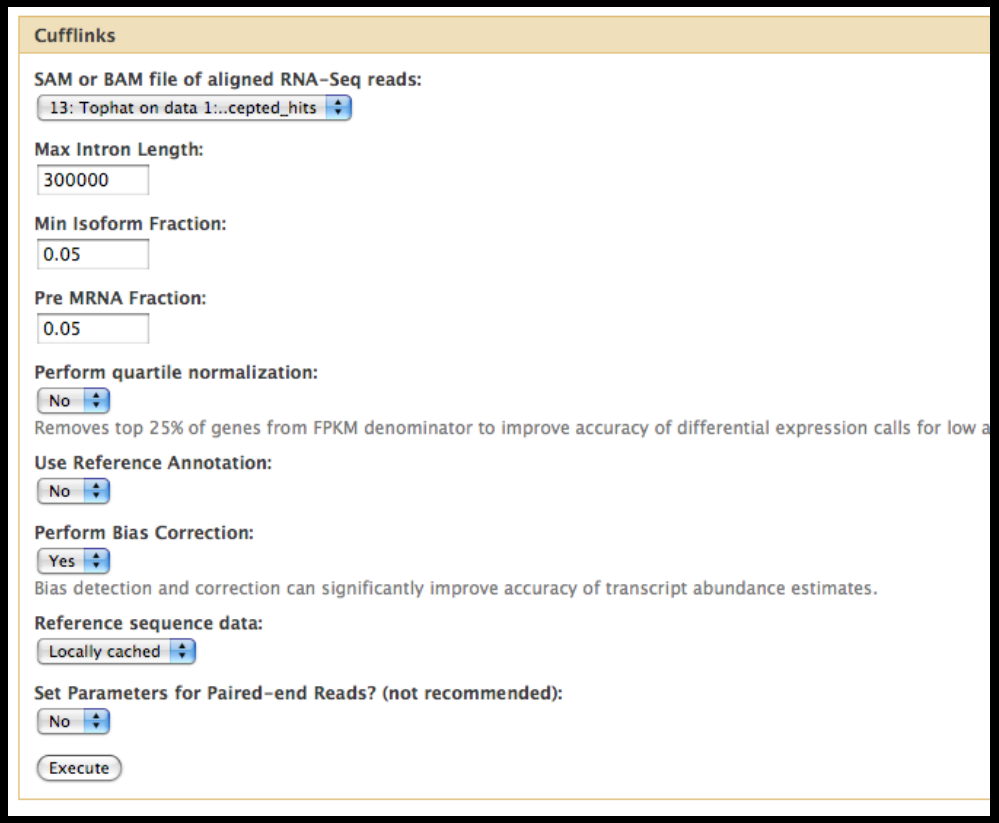
Cufflinks

Goal: transcript assembly and quantitation

Input: aligned RNA-Seq reads, usually from TopHat

Outputs

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



Cuffcompare

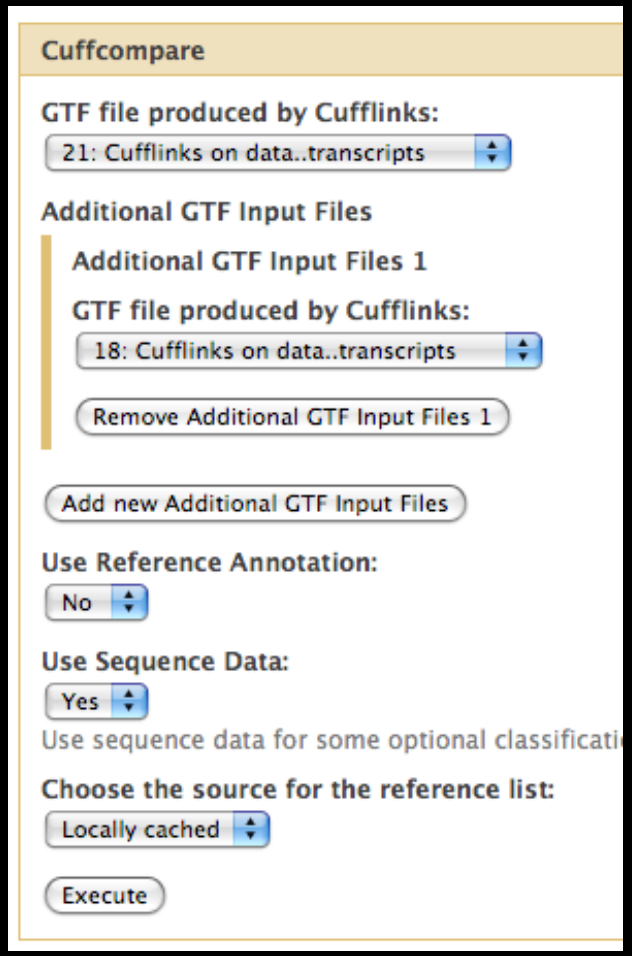
Goals

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

Inputs: assembled transcripts from Cufflinks

Outputs:

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



Cuffdiff

Goals

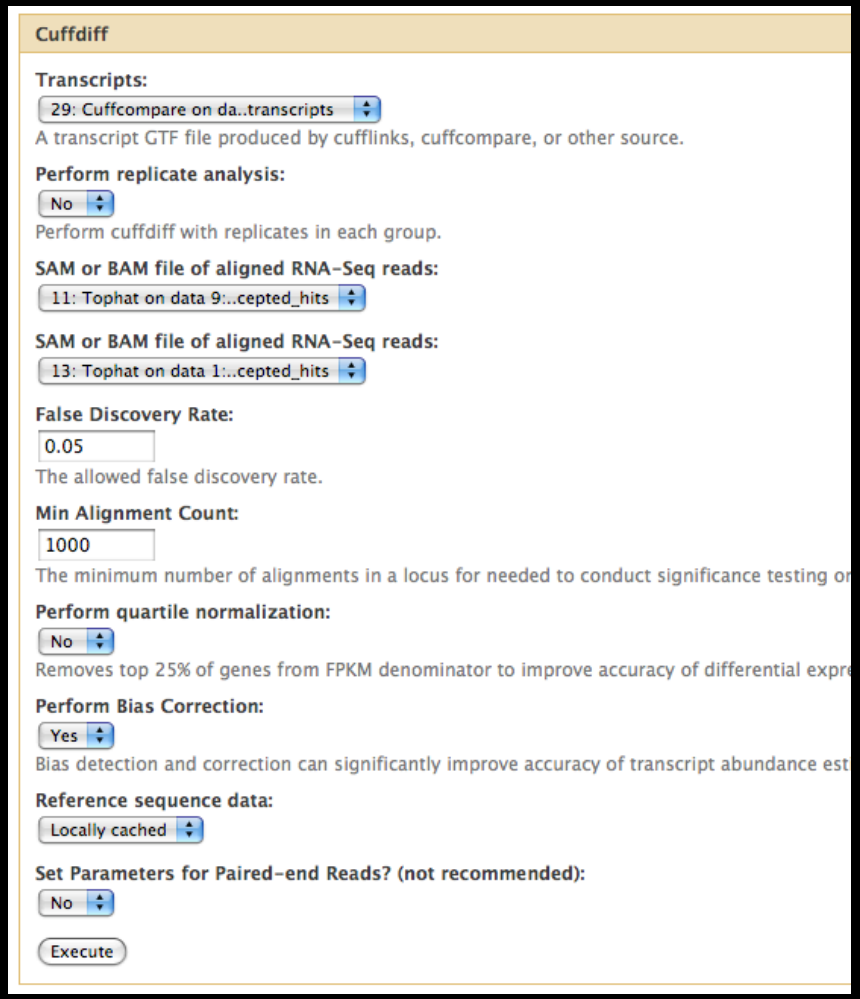
- differential expression testing
- transcript quantitation

Inputs

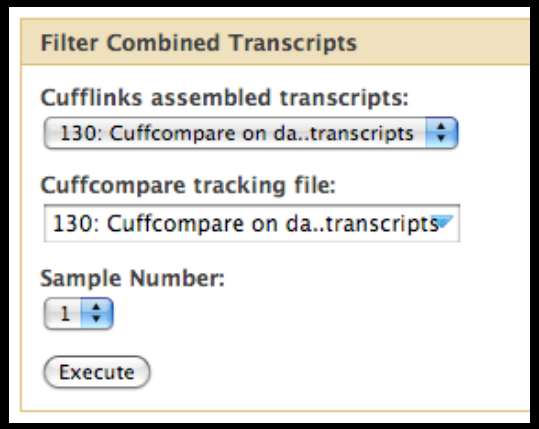
- Combined set of transcripts
- mapped reads from 2+ samples

Outputs

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

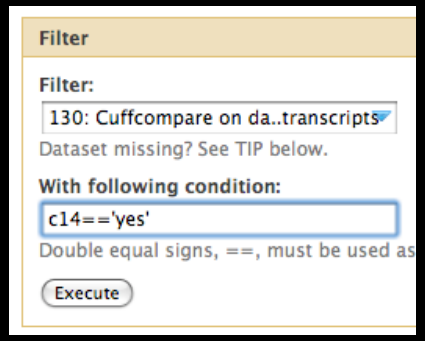


Next Steps

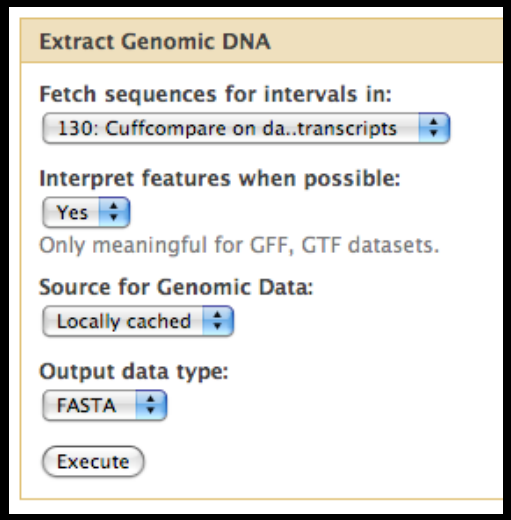


Filtering

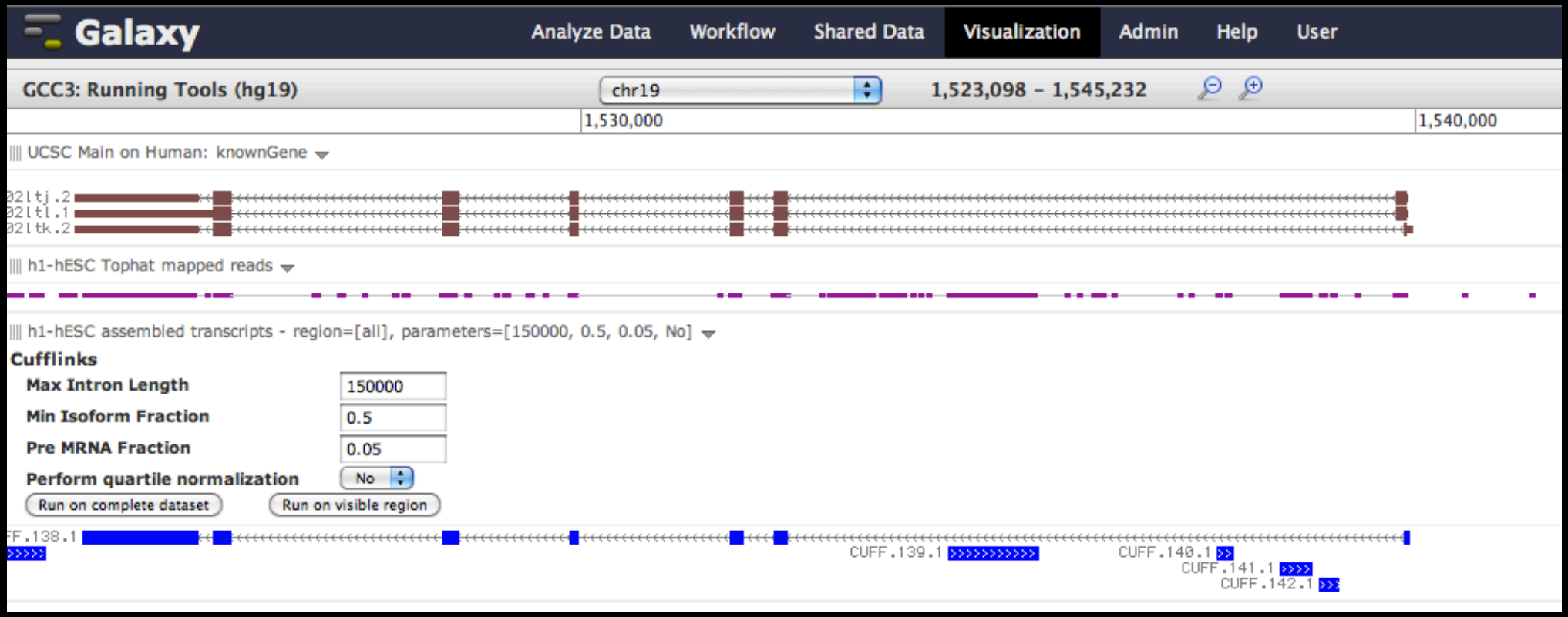
- for differentially expressed elements
- combined transcripts (e.g. for those differentially expressed between samples)

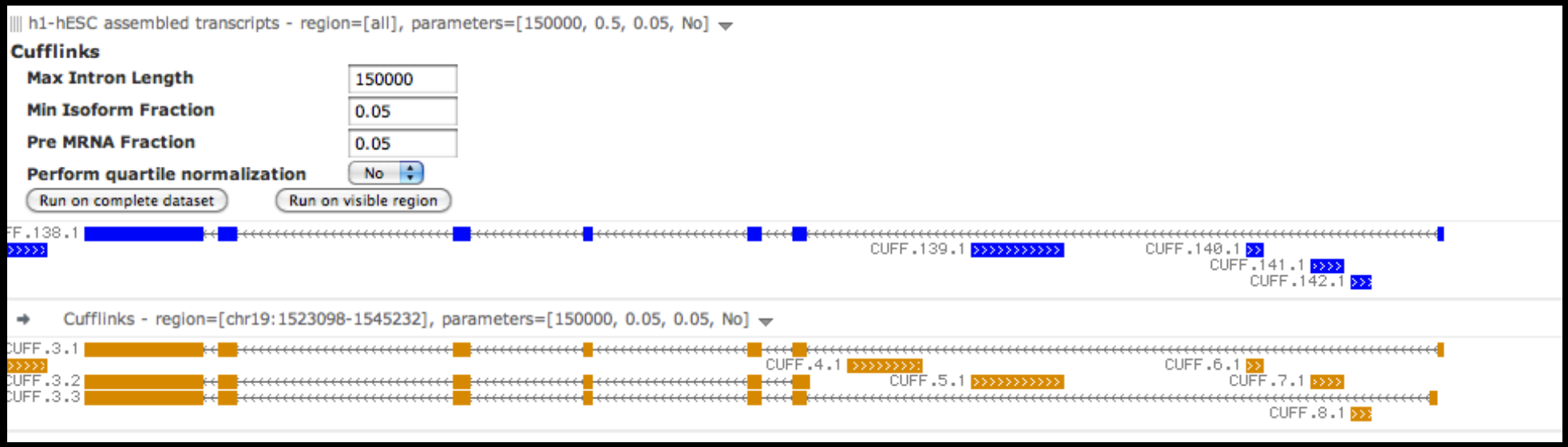


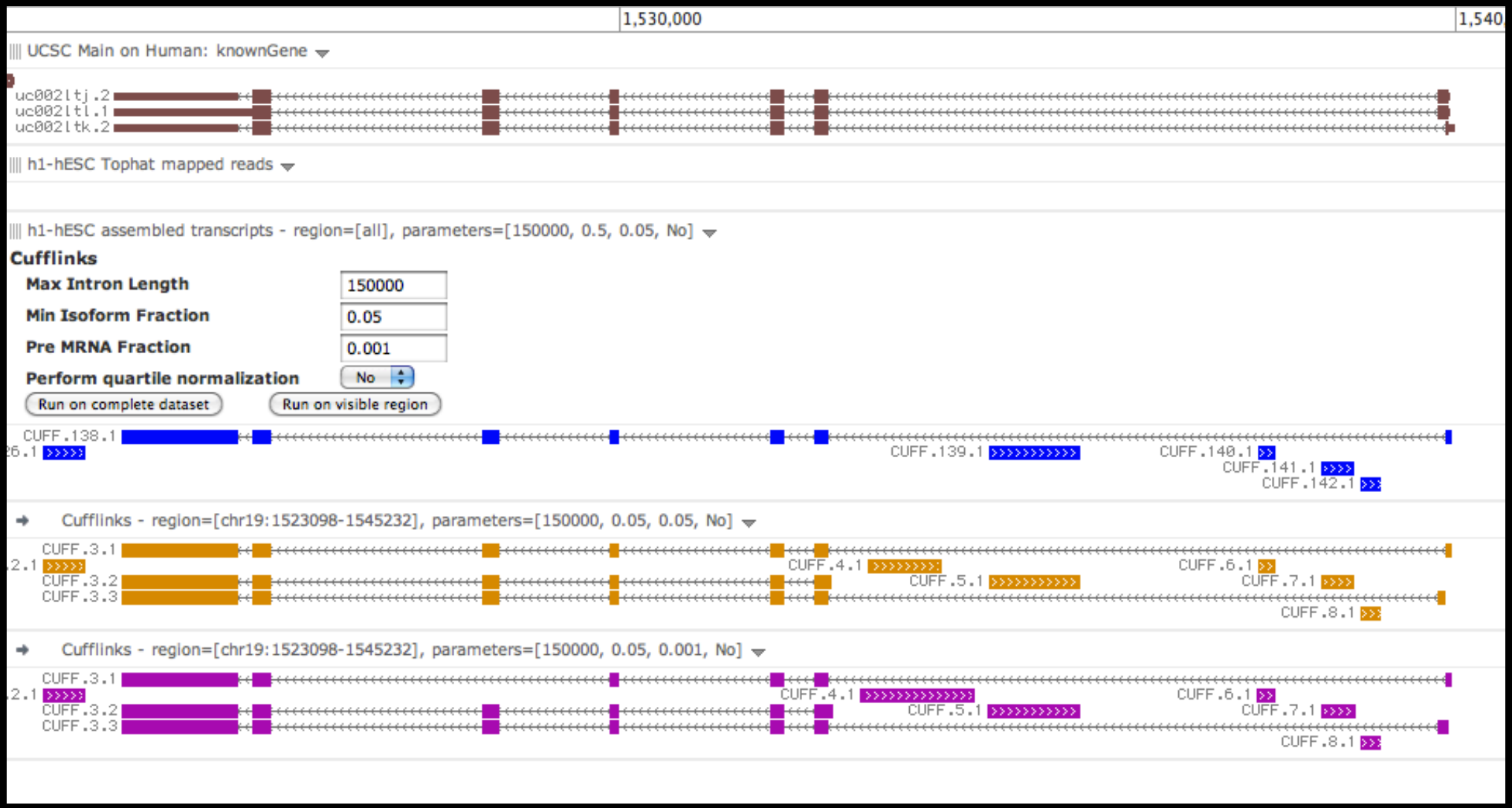
Extract transcript sequences and profile sequences for function



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
FF.138.1
CUFF.139.1
CUFF.140.1
CUFF.141.1
CUFF.142.1
Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No]
CUFF.3.1
CUFF.4.1
CUFF.6.1
CUFF.3.2
CUFF.5.1
CUFF.7.1
CUFF.3.3
CUFF.8.1


1,530,000
1,540
UCSC Main on Human: knownGene
uc002ltj.2
uc002ltl.1
uc002ltk.2
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
CUFF.138.1
CUFF.139.1
CUFF.140.1
CUFF.141.1
CUFF.142.1
Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No]
CUFF.3.1
CUFF.4.1
CUFF.6.1
CUFF.3.2
CUFF.5.1
CUFF.7.1
CUFF.3.3
CUFF.8.1
Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.001, No]
CUFF.3.1
CUFF.4.1
CUFF.6.1
CUFF.3.2
CUFF.5.1
CUFF.7.1
CUFF.3.3
CUFF.8.1

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

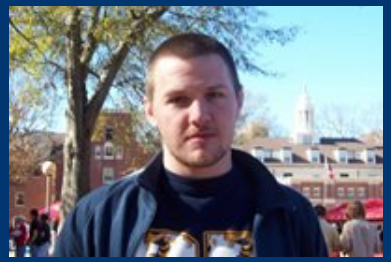
Using Galaxy to Analyze HTS Data

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

Galaxy exercises: ChIP-seq and RNA-seq

EMORY

Enis Afgan



Dannon Baker

Dave Clements

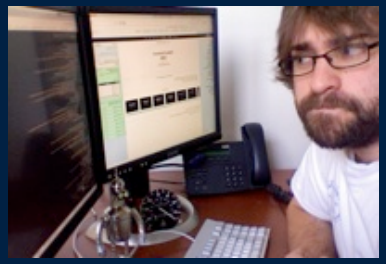
Jeremy Goecks

Kanwei Li

James Taylor

PENNSTATE 1855

Dan Blankenberg



Nate Coraor

Jennifer Jackson

Greg von Kuster

Guru Ananda

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

Chip-seq

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

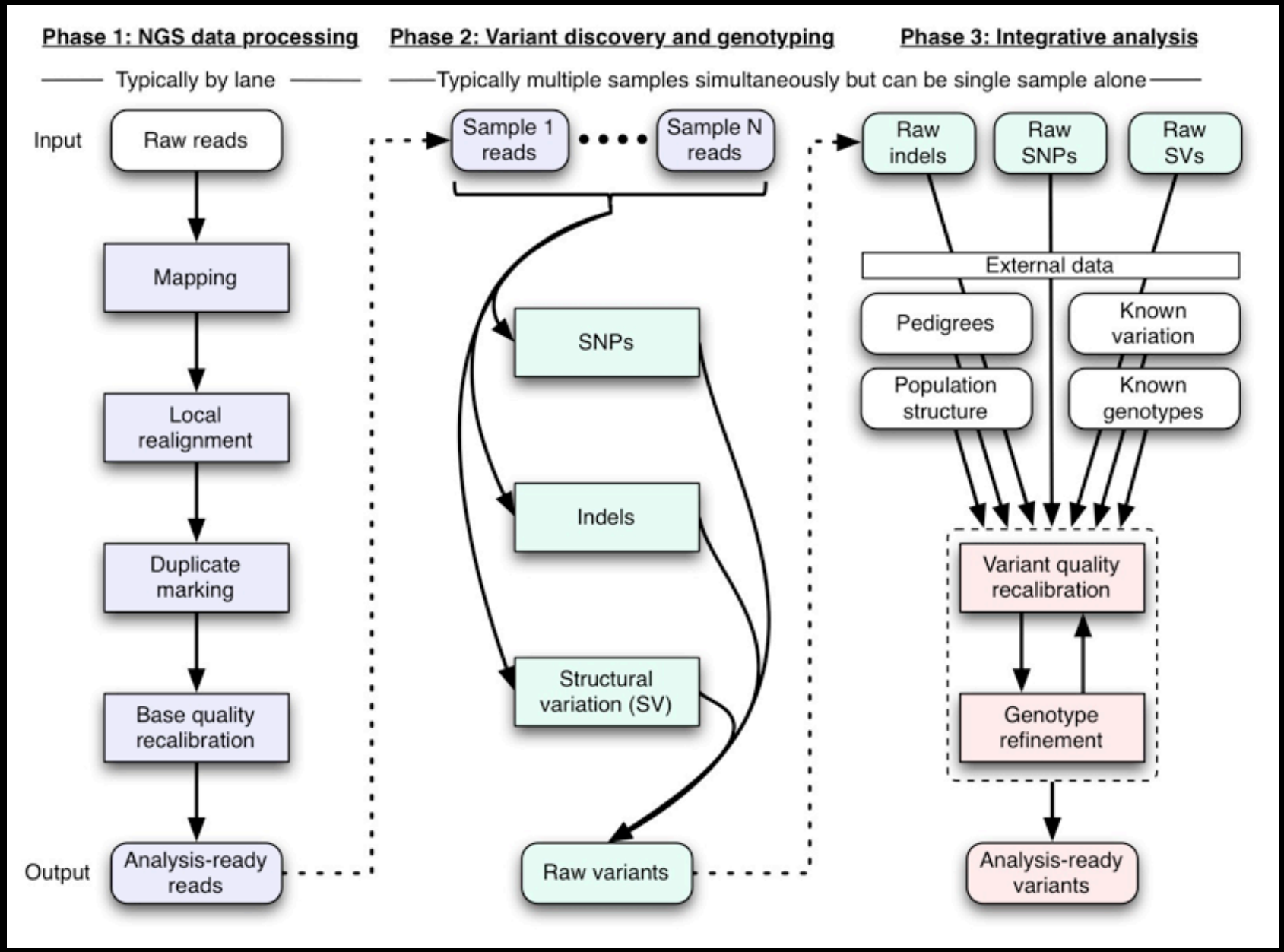
RNA-seq

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

  start Tophat mapping first (second section), then look at QC (first section)

- Add various outputs to a Trackster visualization and play with filtering and reruning tools

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.