Using Galaxy for NGS Analysis

Daniel Blankenberg
Postdoctoral Research Associate
The Galaxy Team
http://UseGalaxy.org

Overview

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

Overview

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

NGS Data

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

A Note on FASTQ

- Contains Sequence data and quality data

```
@UNIQUE_SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- Several Variants exist

Quality Score Comparison

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./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

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009 Dec 16.

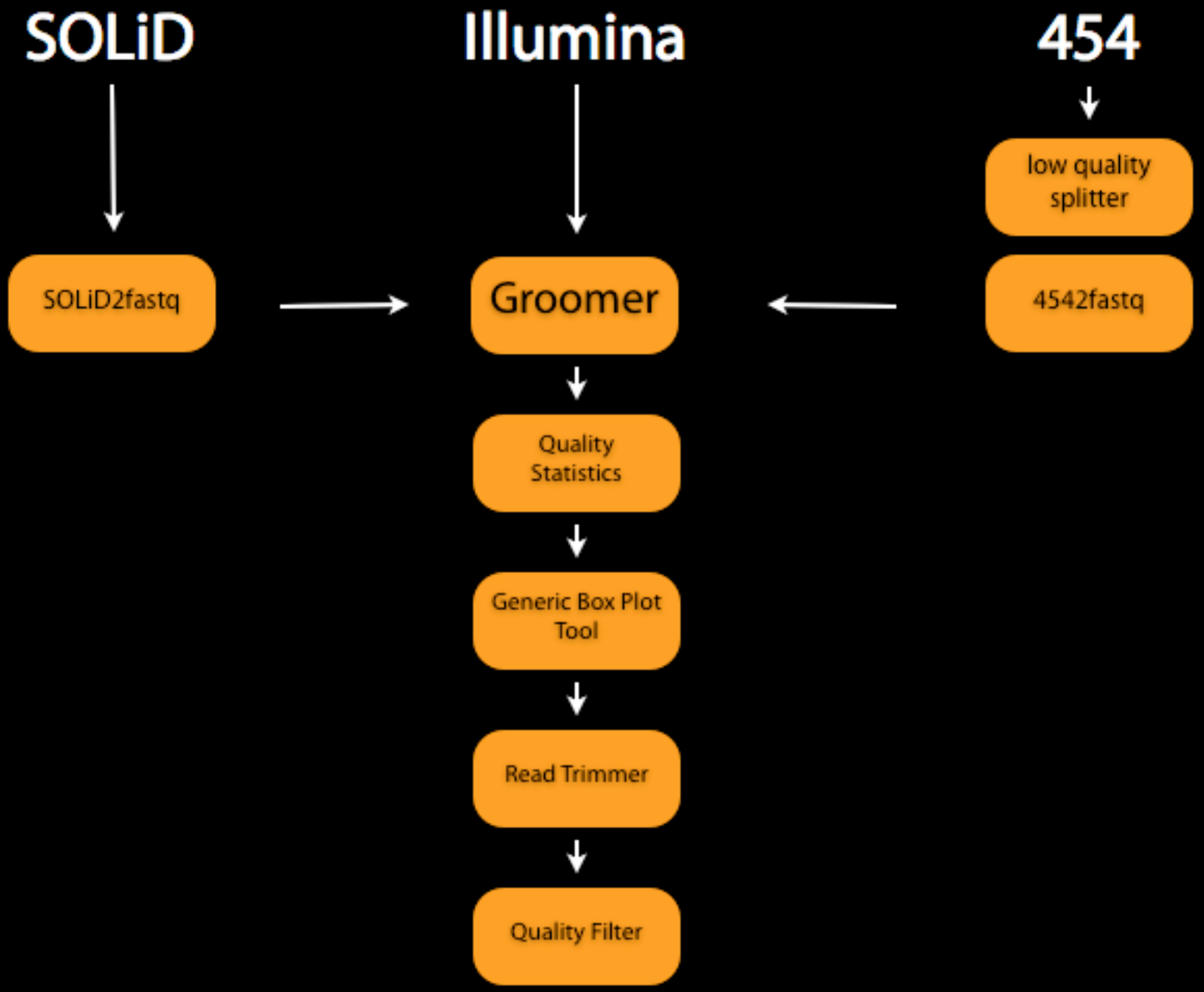
Overview

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

Available NGS Analysis Toolsets

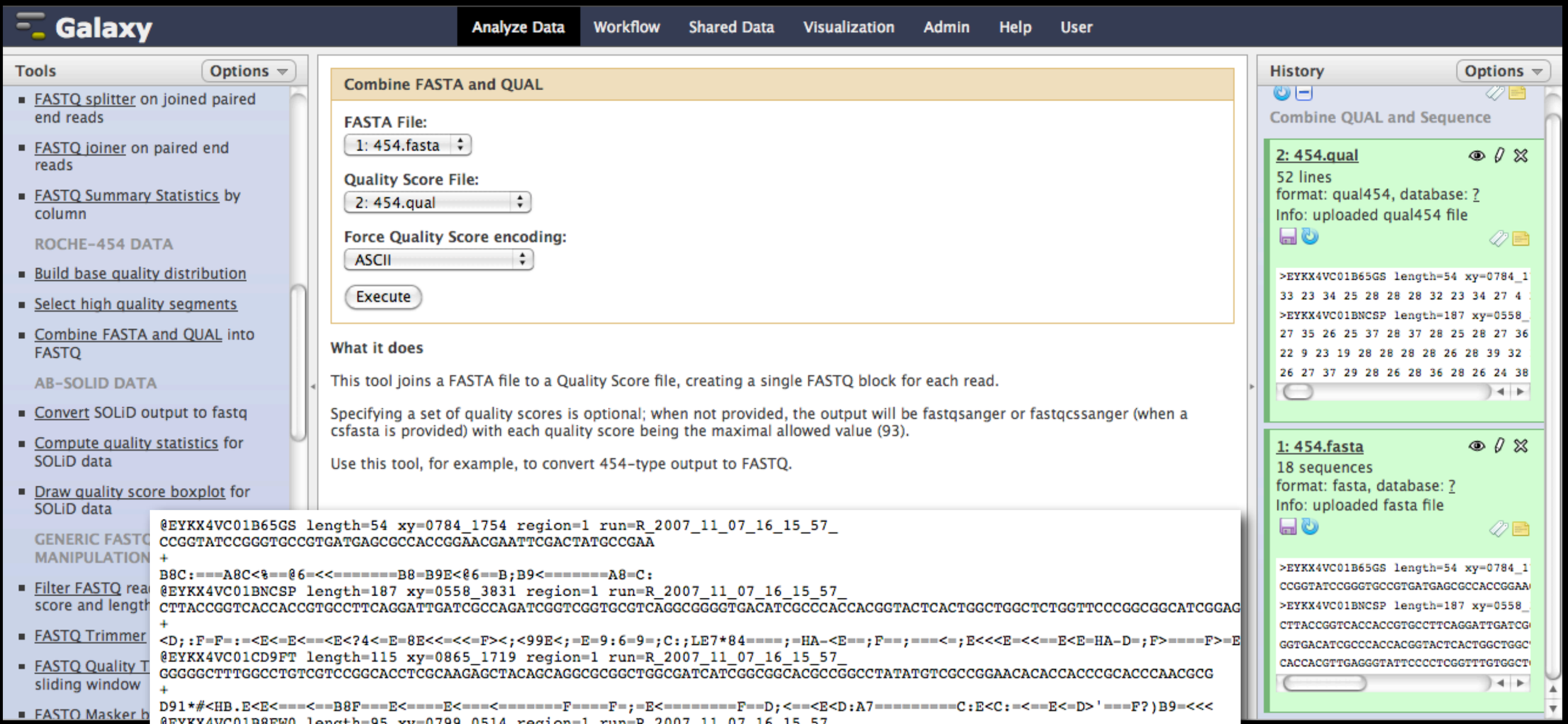
- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMTools
- SNP & INDEL analysis
- Peak Calling / ChIP-seq
- RNA-seq analysis

Prepare and Quality Check



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



Grooming --> Sanger

Galaxy | 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 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, qual scales using the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, 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
```

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
+
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

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./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

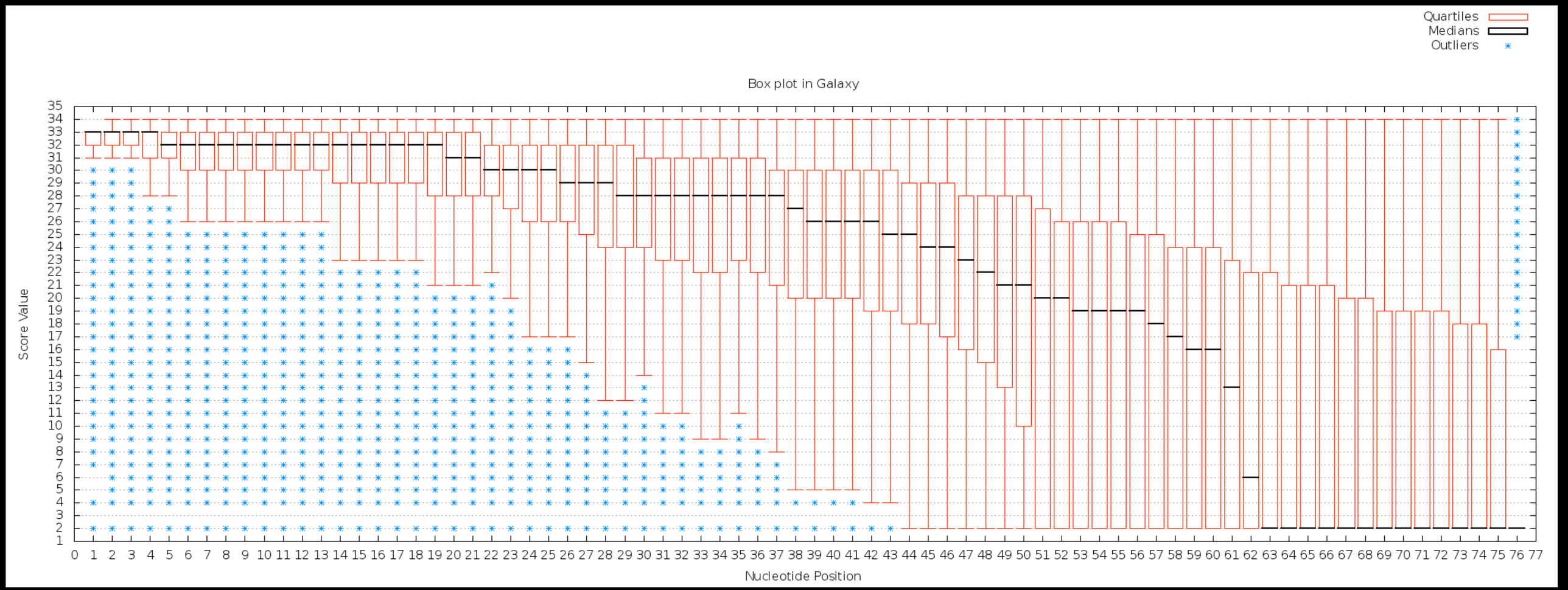
NGS: QC and manipulation

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



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-%7@>CBA@>7@7BBCA4-48%<;;%<B@
```

And you set absolute offsets of 2 and 0:

FASTQ Quality Trimmer

FASTQ File: 7: FASTQ Trimmer on data 2

Keep reads with zero length: ☐

Trim ends: 5' and 3'

Window size: 1

Step Size: 1

Maximum number of bases to exclude from the window during aggr

0

Aggregate action for window: min score

Trim until aggregate score is: >=

Quality Score: 0.0

Execute

Quality Filtering

Filter FASTQ

FASTQ File:

7: FASTQ Trimmer on data 2

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

Minimum Size:

0

Maximum Size:

0

A maximum size less than 1 indicates no limit.

Minimum Quality:

0.0

Maximum Quality:

0.0

A maximum quality less than 1 indicates no limit.

Maximum number of bases allowed outside of quality range:

0

This is paired end data:

Quality Filter on a Range of Bases

Add new Quality Filter on a Range of Bases

Execute

Quality Filter on a Range of Bases

Quality Filter on a Range of Bases 1

Define Base Offsets as:

Absolute Values

Use Absolute for fixed length reads (Illumina, SOLiD)
Use Percentage for variable length reads (Roche/454)

Offset from 5' end:

0

Values start at 0, increasing from the left

Offset from 3' end:

0

Values start at 0, increasing from the right

Aggregate read score for specified range:

min score

Keep read when aggregate score is:

>=

Quality Score:

0.0

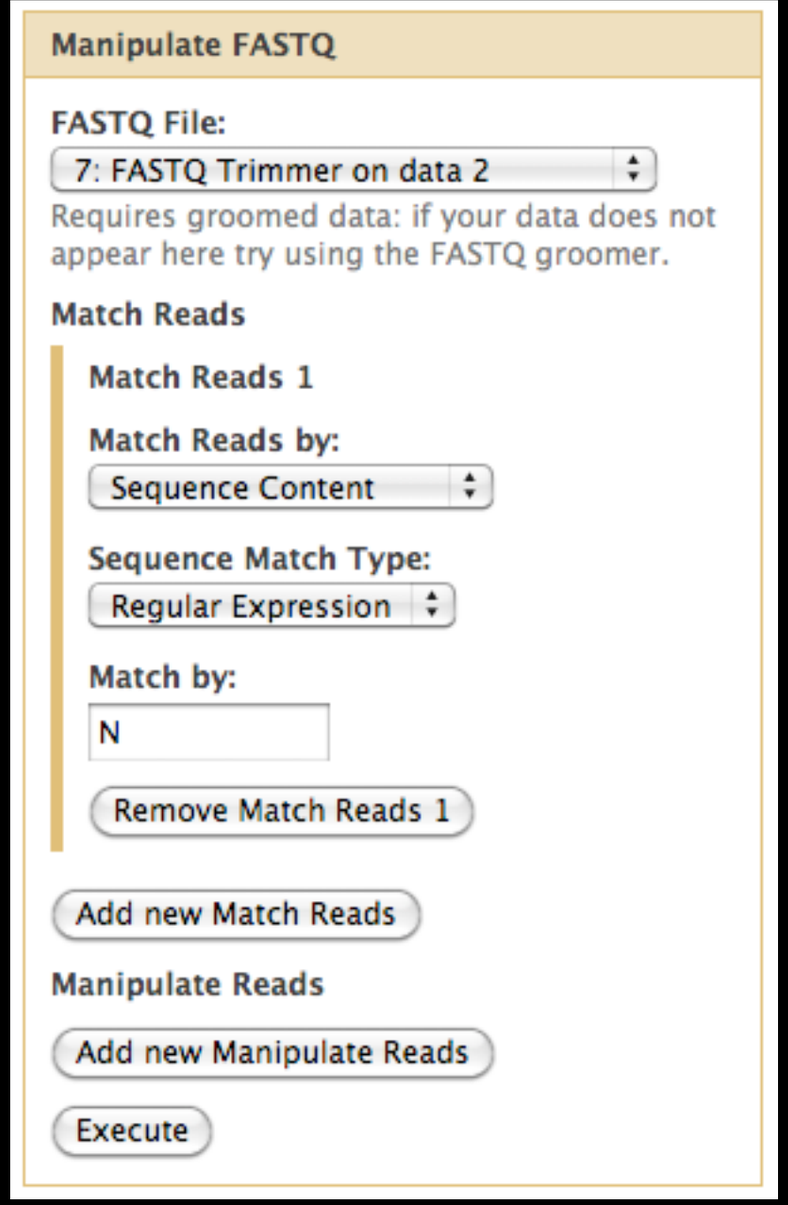
Remove Quality Filter on a Range of Bases 1

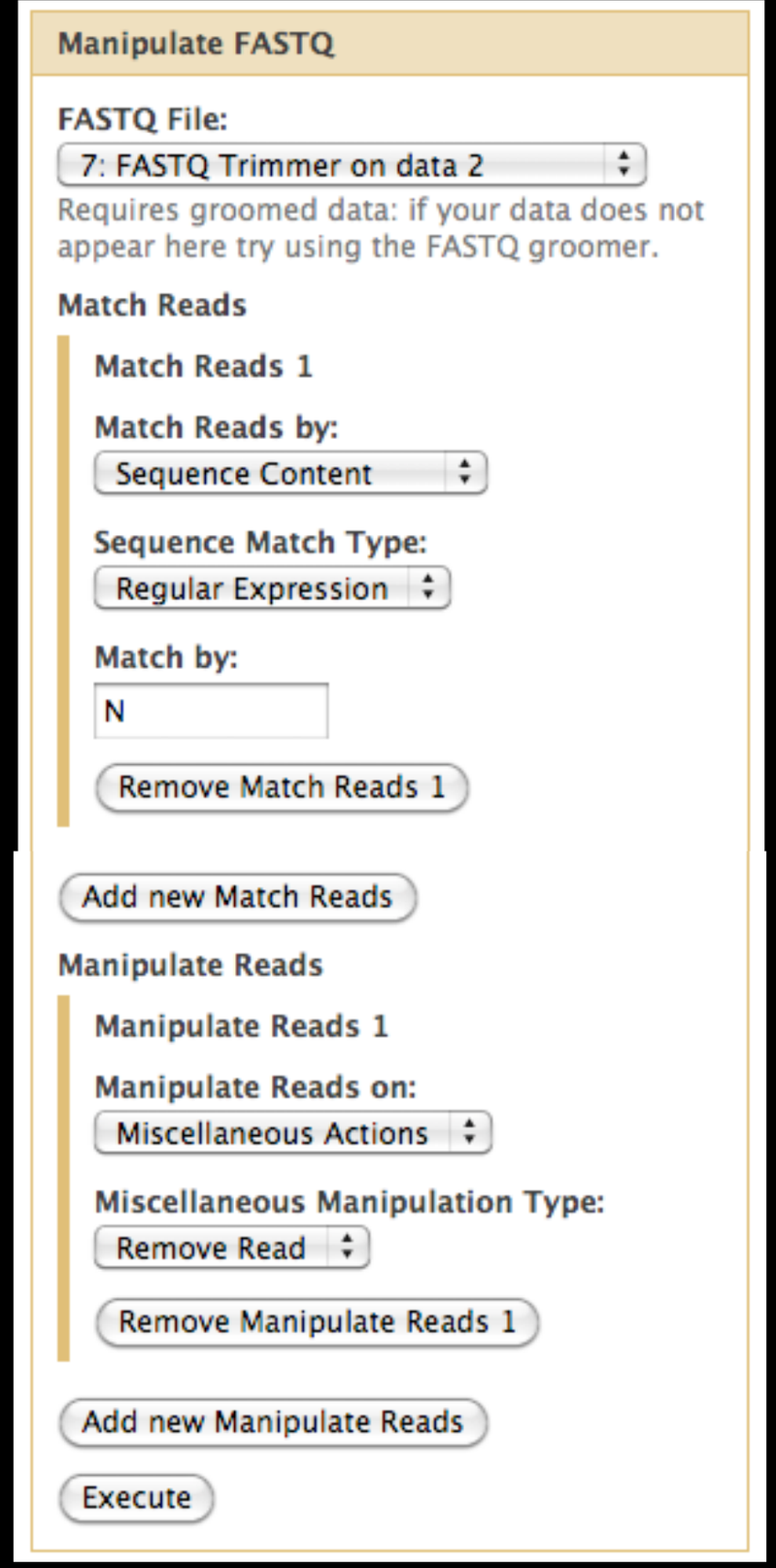
Add new Quality Filter on a Range of Bases

Execute

Manipulate FASTQ

- Remove reads with N's?





Available NGS Analysis Toolsets

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

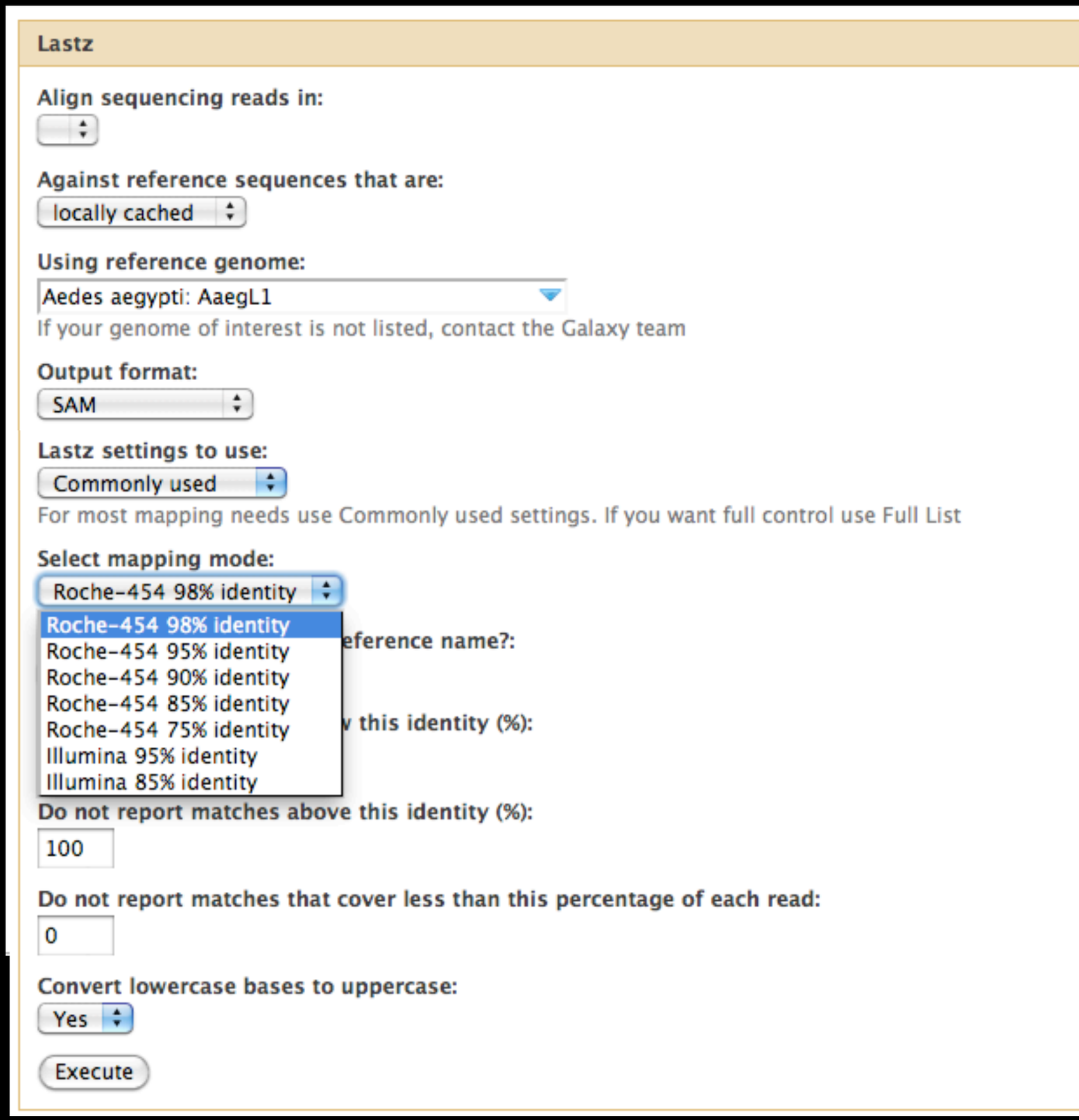
Mapping NGS Data

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

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

Mappers

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



Lastz

Align sequencing reads in:
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

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

Do you want to modify the reference name?:
No

Do not report matches below this identity (%):
0

Do not report matches above this identity (%):
100

Do not report matches that cover less than this percentage of each read:
0

Convert lowercase bases to uppercase:
Yes

Execute

What it does

LASTZ is a high performance pairwise sequence aligner derived from BLASTZ. It is written by Bob Harris in Webb Miller's laboratory at Penn State University. Special scoring sets were derived to improve runtime performance and quality. This Galaxy version of LASTZ is geared towards aligning short (Illumina/Solexa, AB/SOLiD) and medium (Roche/454) reads against a reference sequence. There is excellent, extensive documentation on LASTZ available here.

Input formats

LASTZ accepts reference and reads in FASTA format. However, because Galaxy supports implicit format conversion the tool will recognize fastq and other method specific formats.

Available NGS Analysis Toolsets

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

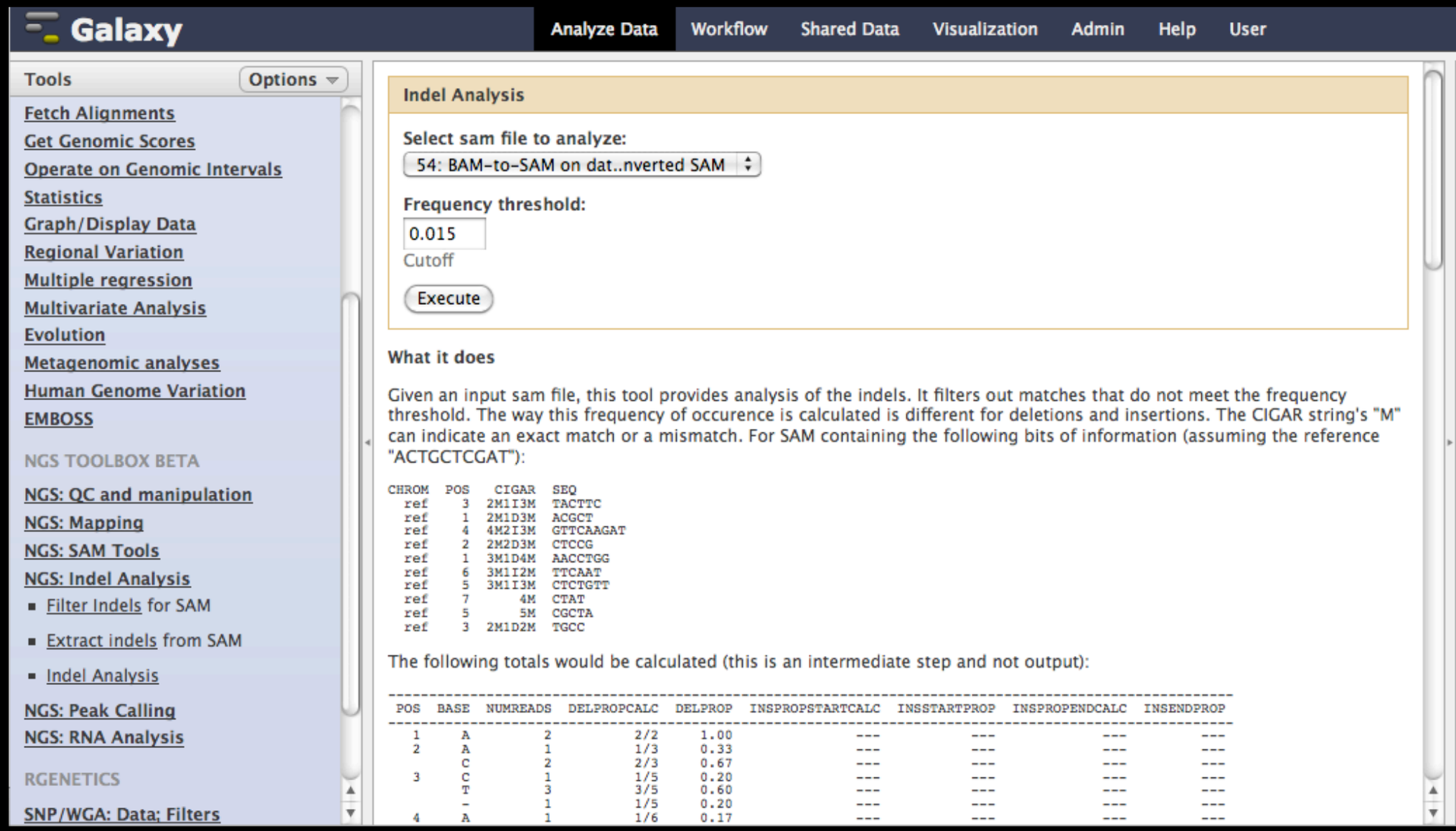
SNPs & INDELs

- SNPs from Pileup
  - Generate
  - Filter

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

INDELS



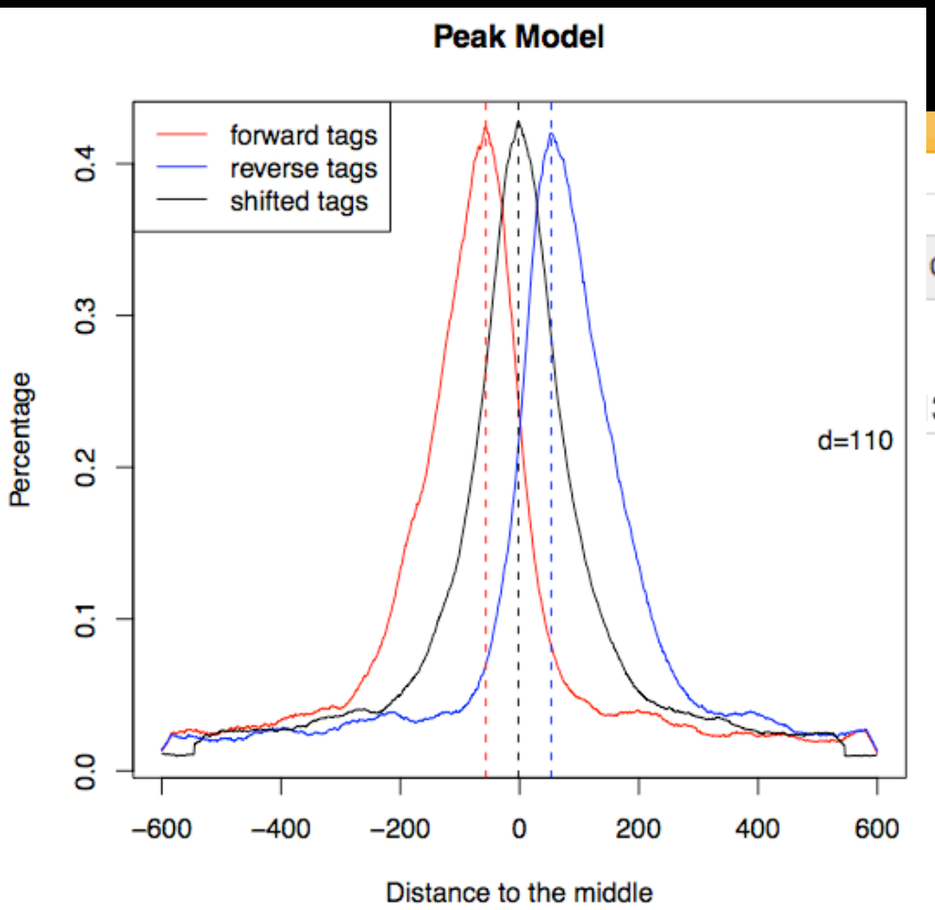
Available NGS Analysis Toolsets

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

Peak Calling / ChIP-seq analysis

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

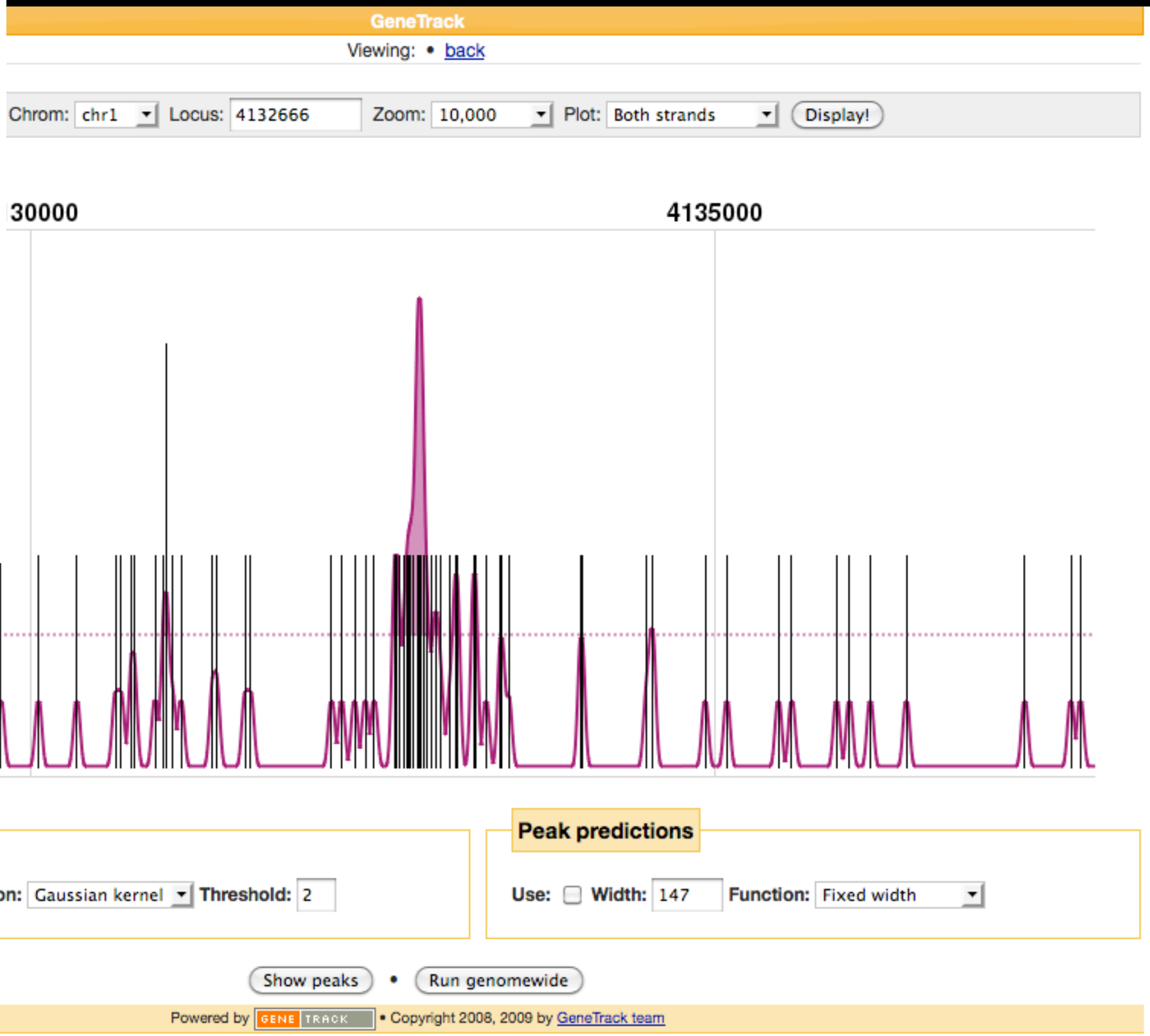
Punctate Binding



MACS

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

GeneTrack



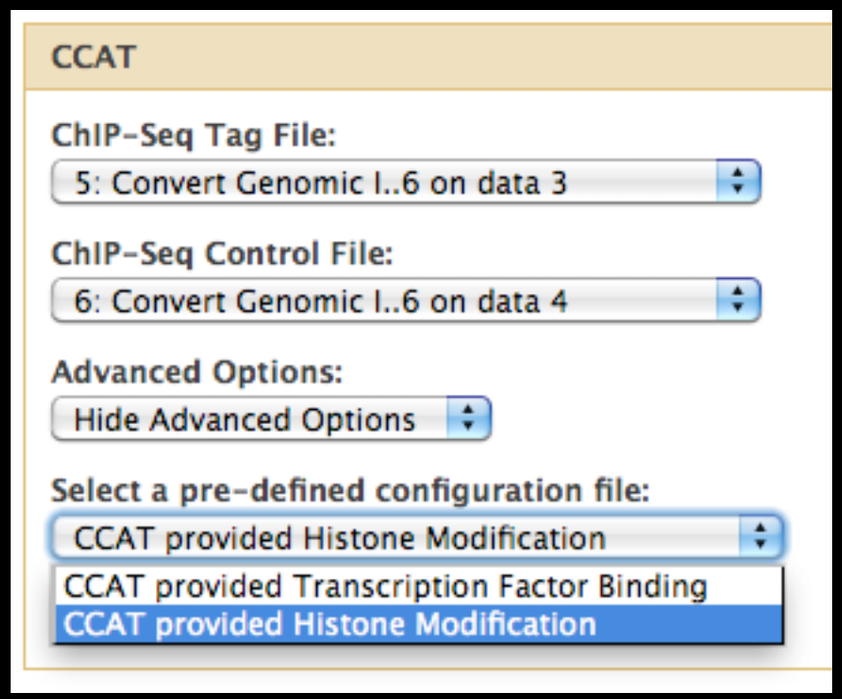
Albert I, Wachi S, Jiang C, Pugh BF. GeneTrack--a genomic data processing and visualization framework. Bioinformatics. 2008 May 15;24(10):1305-6. Epub 2008 Apr 3.

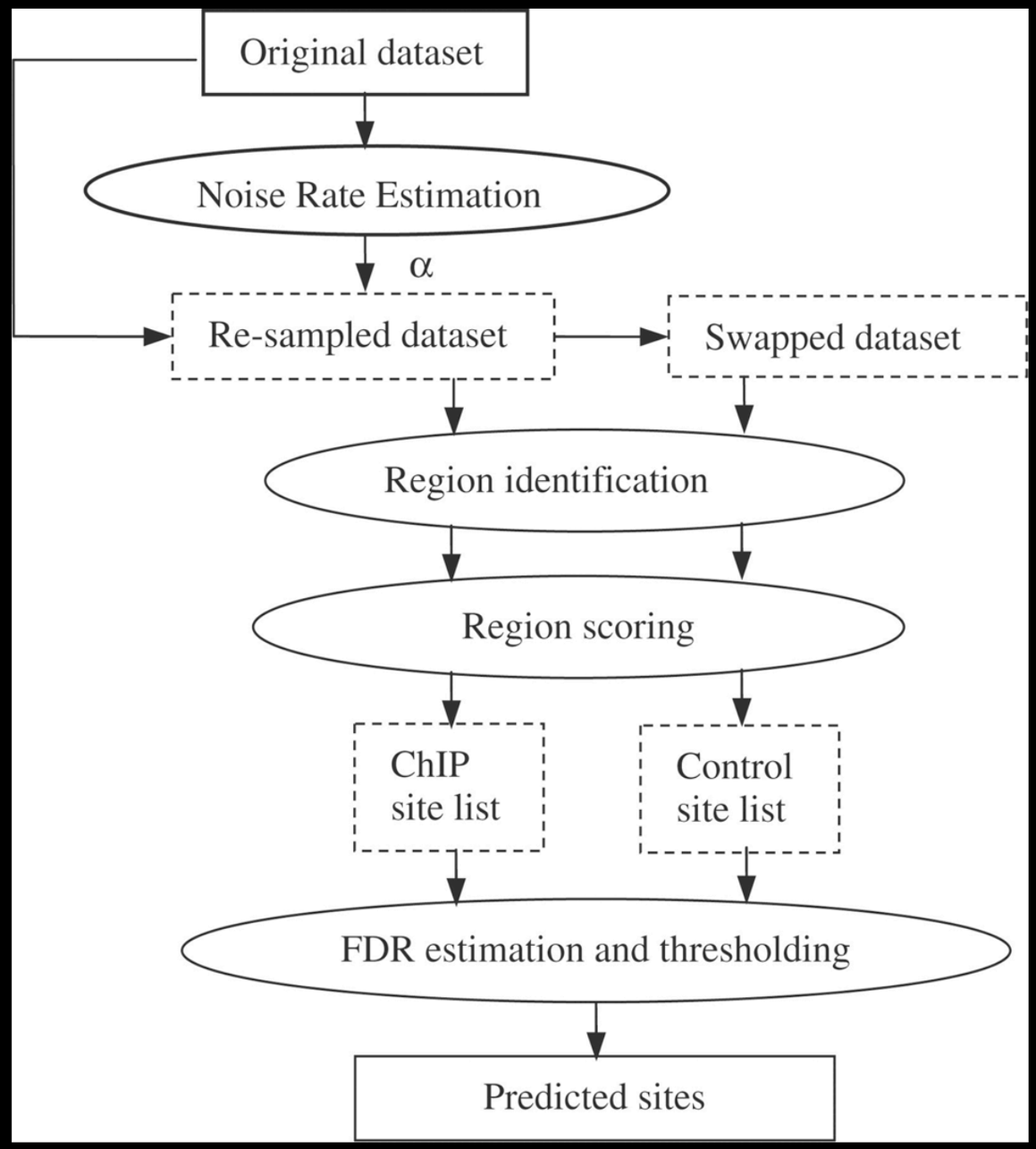
MACS

- Inputs
  - Enriched Tag file
  - Control / Input file (optional)
- Outputs
  - Called Peaks
  - Negative Peaks (when control provided)
  - Shifted Tag counts (wig, convert to bigWig for visualization)

Diffuse Binding

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





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.

ChIP-seq Exercise

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

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

Accessible Page | ChIP seq exercise

ChIP-seq exercises

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

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

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

Mapping reads and peak calling

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

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

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

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

Calling peaks with a control sample

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

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

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

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

Create a workflow and reuse

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

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

Galaxy Dataset | G1E CTCF

Galaxy Dataset | G1E input

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

Identify differential binding sites

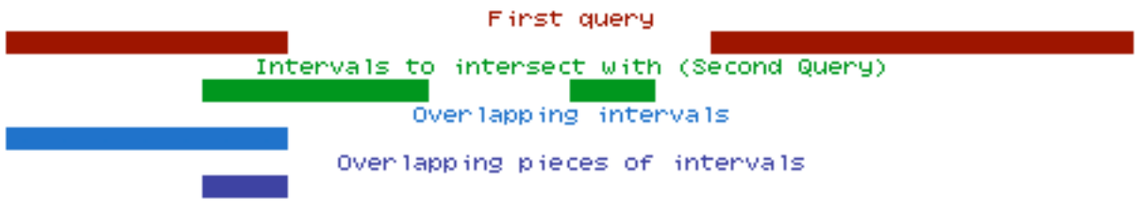
G1E is a model for erythropoiesis, the G1E line is a GATA1 null derived line which can be induced to differentiate by estradiol treatment (thus G1E-ER4). Here we will use Galaxy to identify sites that have differential binding across the two developmental stages.

I have Peaks, now what?

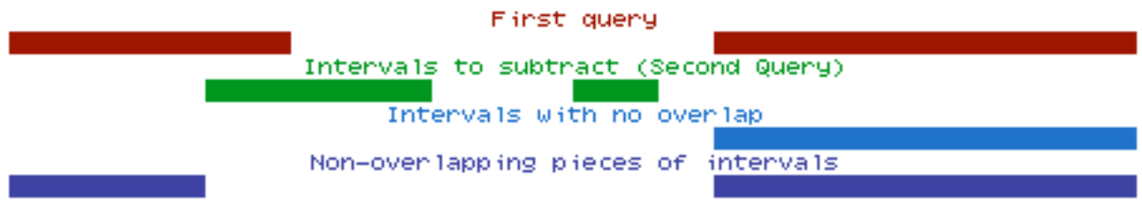
- Visualize
- Share
- Compare to existing Annotations
  - Interval Operations
  - Annotation Profiler

Genomic Interval Operations

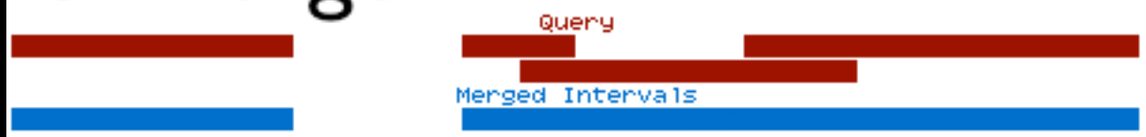
A Intersect



B Subtract



C Merge



D Concatenate

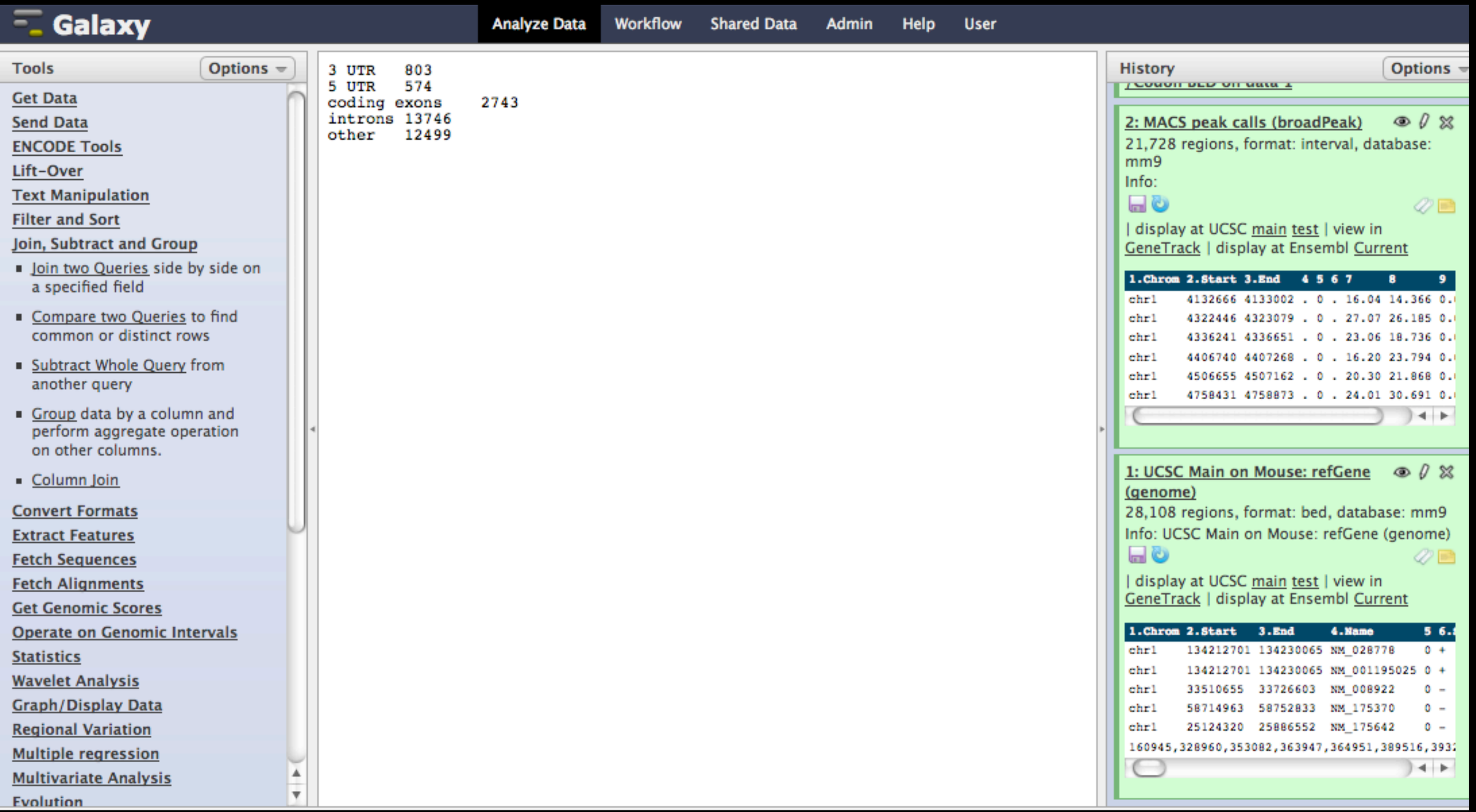
E Complement

F Cluster

Secondary Analysis

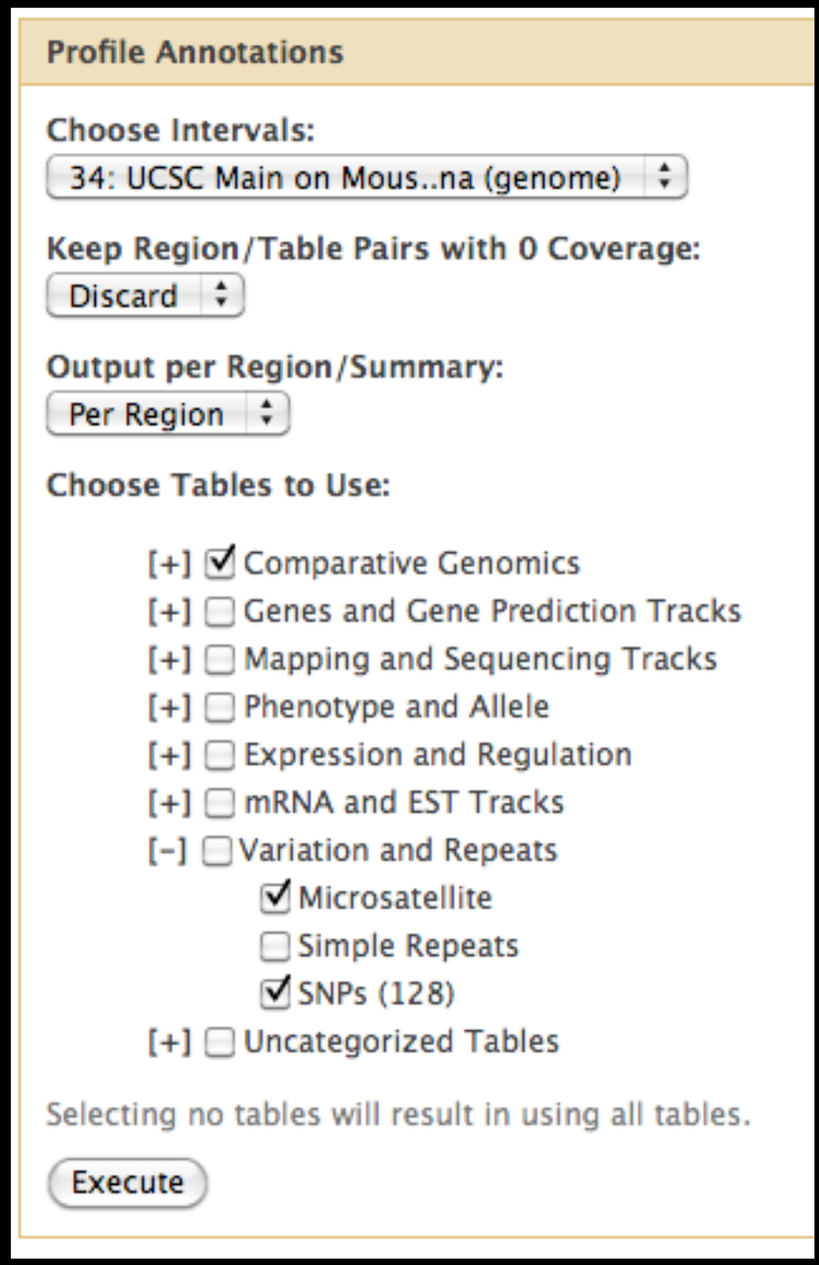
- A simple goal: determine number of peaks that overlap a) coding exons, b) 5-UTRs, c) 3-UTRs, d) introns and d) other regions
- Get Data
  - Import Peak Call data
  - Retrieve Gene location data from external data resource
- Extract exon and intron data from Gene Data (Gene BED To Exon/Intron/Codon BED expander x4)
- Create an Identifier column for each exon type (Add column x4)
- Create a single file containing the 4 types (Concatenate)
- Complement the exon/intron intervals
- Force complemented file to match format of Gene BED expander output (convert to BED6)
- Create an Identifier column for the 'other' type (Add column)
- Concatenate the exons/introns and other files
- Determine which Peaks overlap the region types (Join)
- Calculate counts for each region type (Group)

Secondary Analysis



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



Available NGS Analysis Toolsets

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

RNA-seq

- TopHat
- Cufflinks
- Cuffcompare
- Cuffdiff

NGS: RNA Analysis

RNA-SEQ

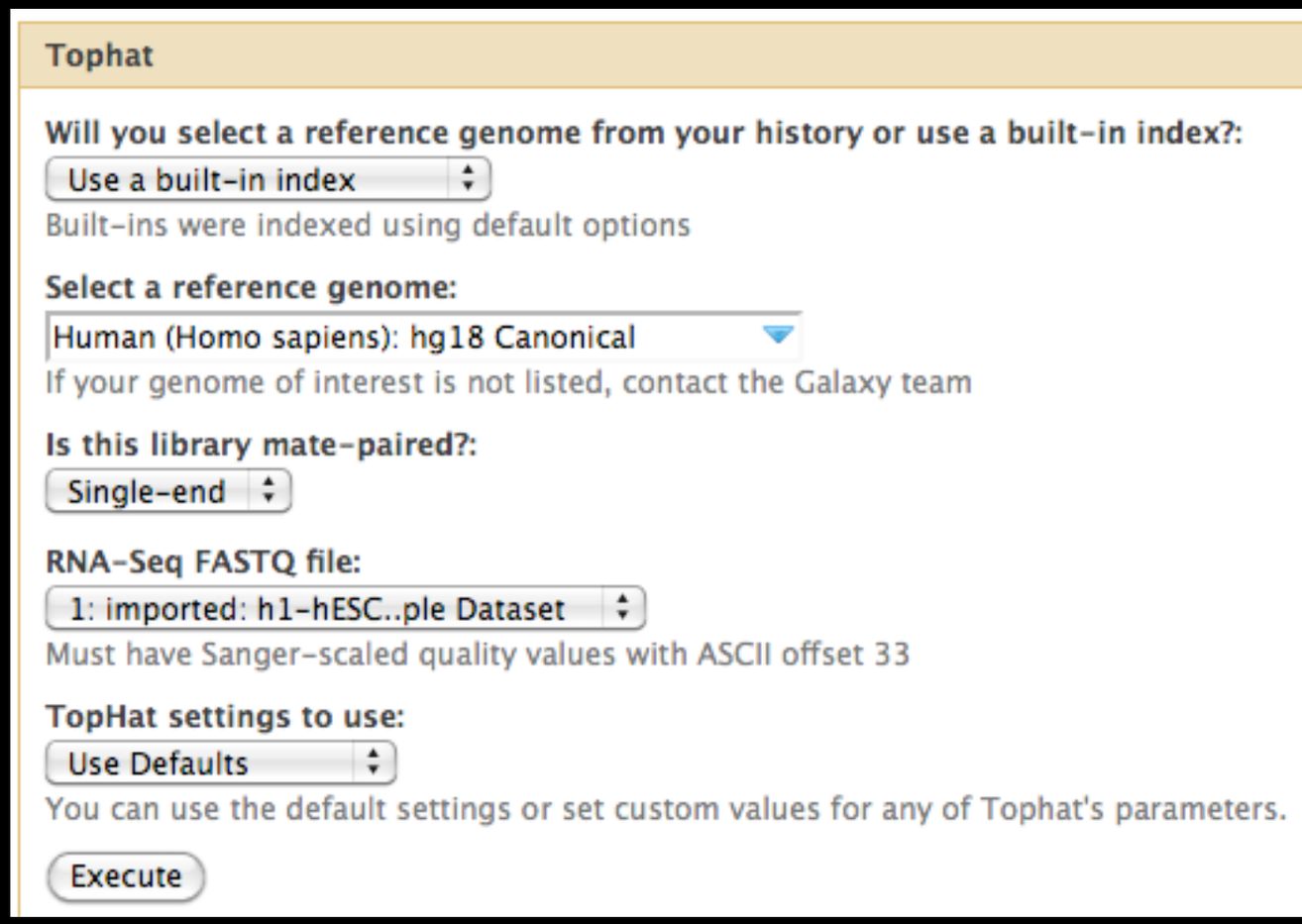
- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript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

FILTERING

- Filter Combined Transcripts using tracking file

TopHat

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



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

Cufflinks

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

Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621

Cuffcompare

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

Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621

Cuffdiff

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

Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621

RNA-seq Tutorial

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

Published Pages | jeremy | Galaxy RNA-seq Analysis Exercise

RNA-seq Analysis Exercise

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

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

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

and

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

Understanding and preprocessing the reads

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

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

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

Map processed reads

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

http://main.g2.bx.psu.edu/u/jeremy/p/galaxy-rna-seq-analysis-exercise

Overview

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

Sample Tracking System

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

Sequencing Facility Managers

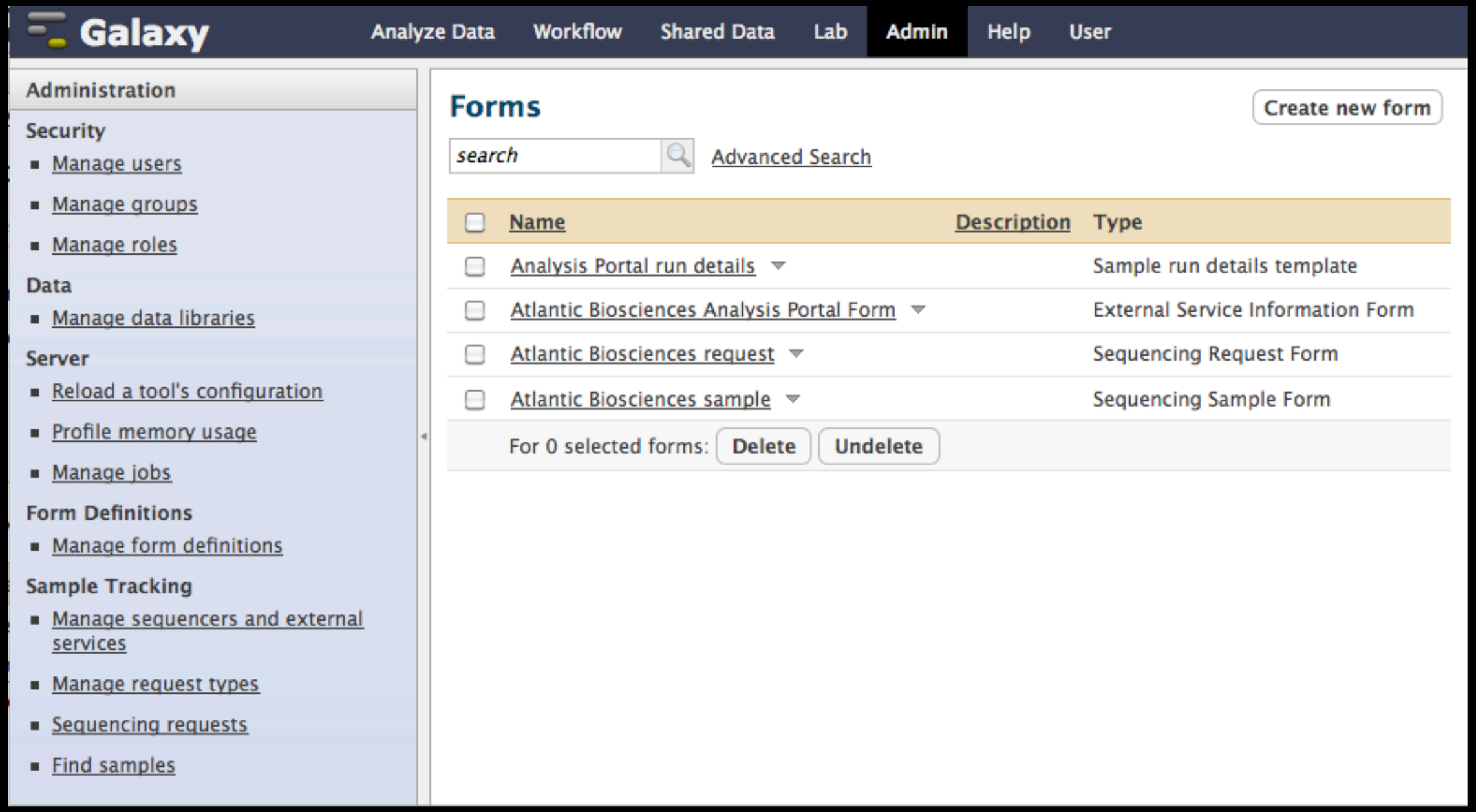
- Setup the Galaxy sample tracking system according to the core facility workflow. [Once per request type]
- Create and submit a sequencing request on behalf of another user.
- Reject an incomplete or erroneous sequencing request.
- Receive samples and assign them tracking barcodes.
- Setup data transfer from the sequencer
- Transfer the datasets from the sequencer to Galaxy at the end of the sequence run.

} Can be automated

Sequencing Facility Users

- Create and submit a sequencing request.
- Edit and resubmit a rejected sequencing request.
- Obtain datasets at the end of a sequencing run.
- Select Libraries and Histories, and Workflows to populate and run on sequenced samples.

Configure Available Request / Sample Options



- Uses Galaxy forms, do this once

View

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

Name

Atlantic Biosciences request

Description

Form definition fields

1. Name (TextField)

2. Scientific Contact (AddressField)

Add field

Save

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

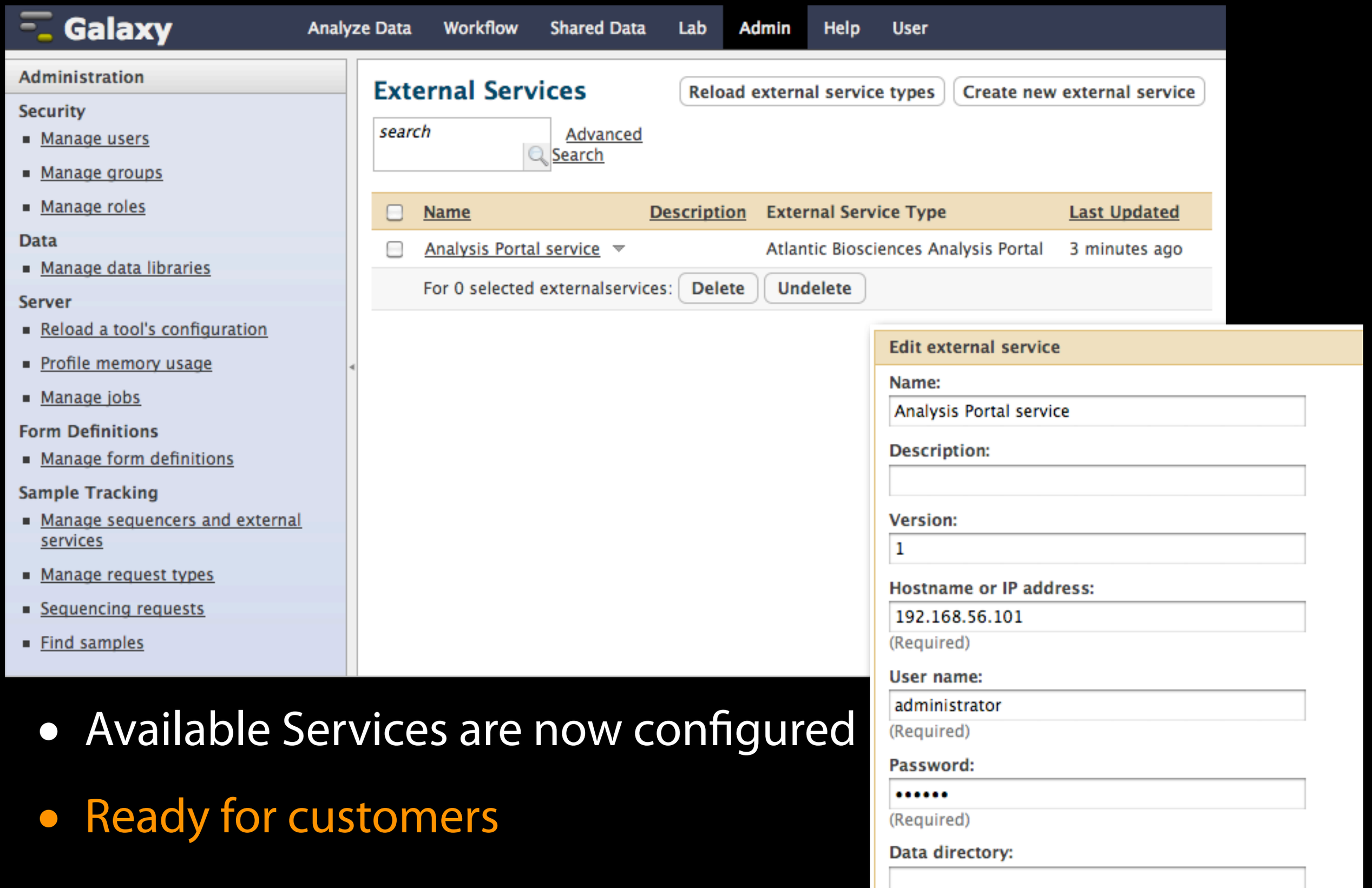
Edit

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

Layout1

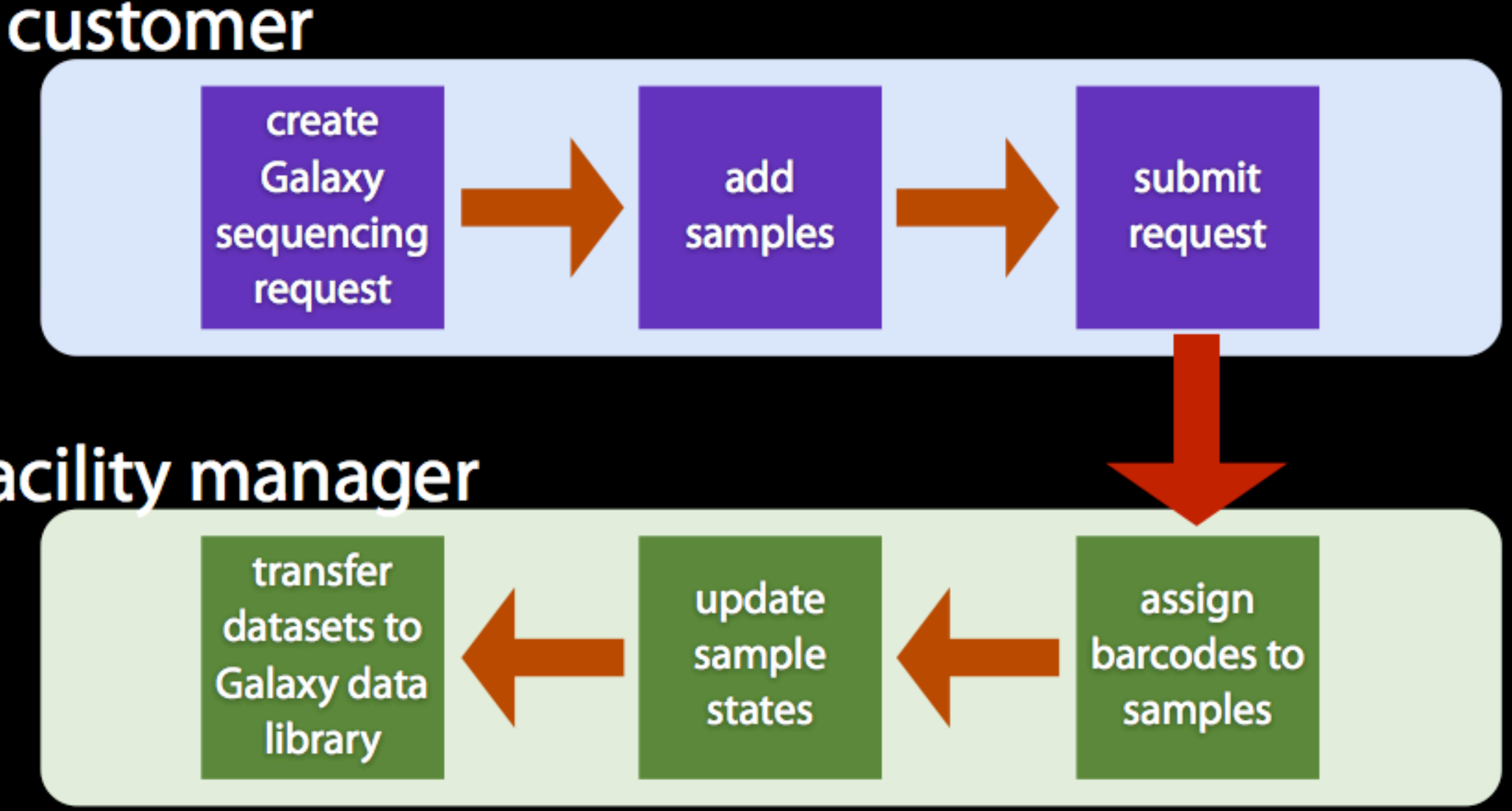
Run type	Read length	Number of Lanes	Alignment target	Processing time	Comments
SelectField: – (optional) Options: SR PE	SelectField: – (optional) Options: 36 50 75 100	TextField: – (optional)	TextField: – (optional)	SelectField: – (optional) Options: Std Rush option3	TextField: – (optional)

Configure the Sequencer

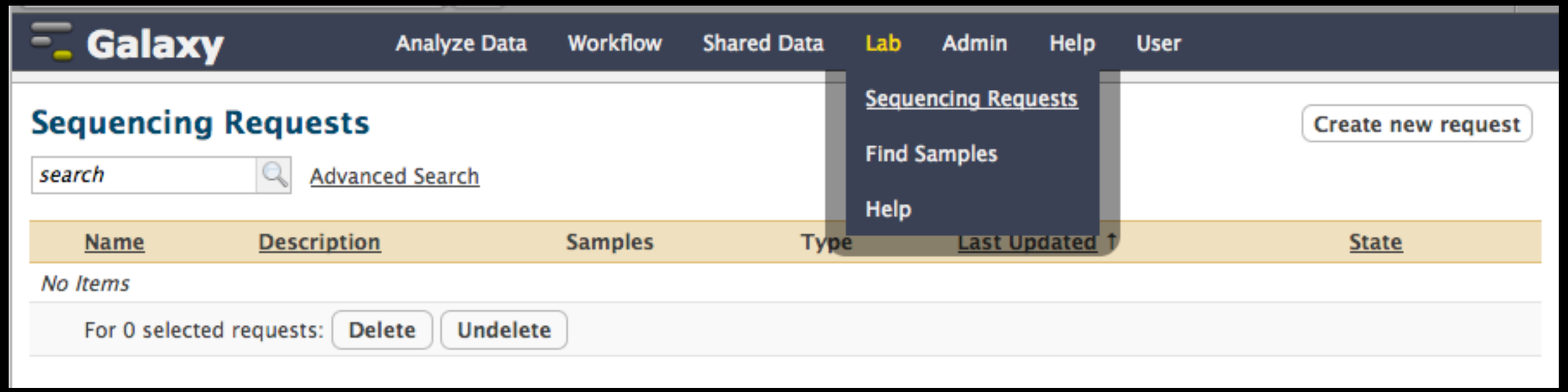


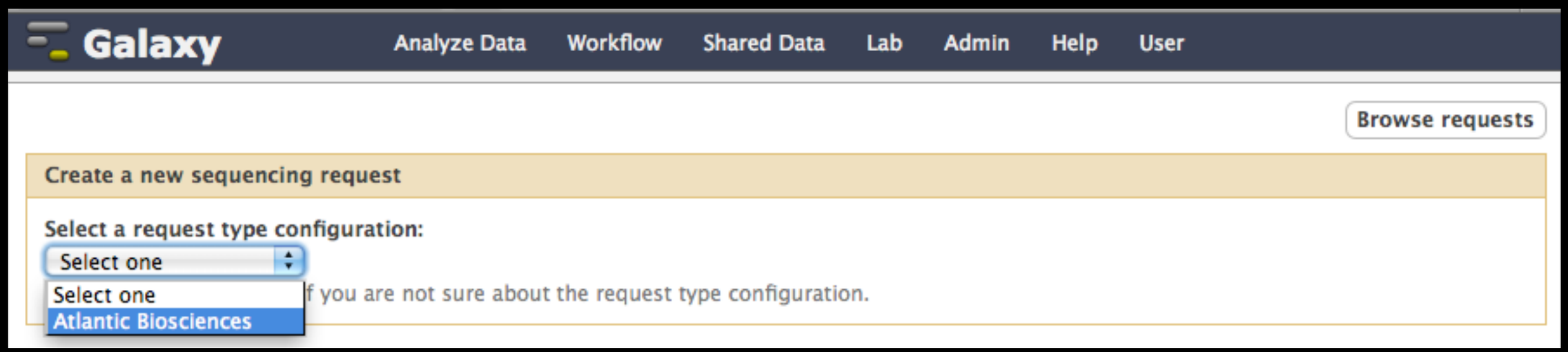
- Available Services are now configured
- Ready for customers

How does it all work?

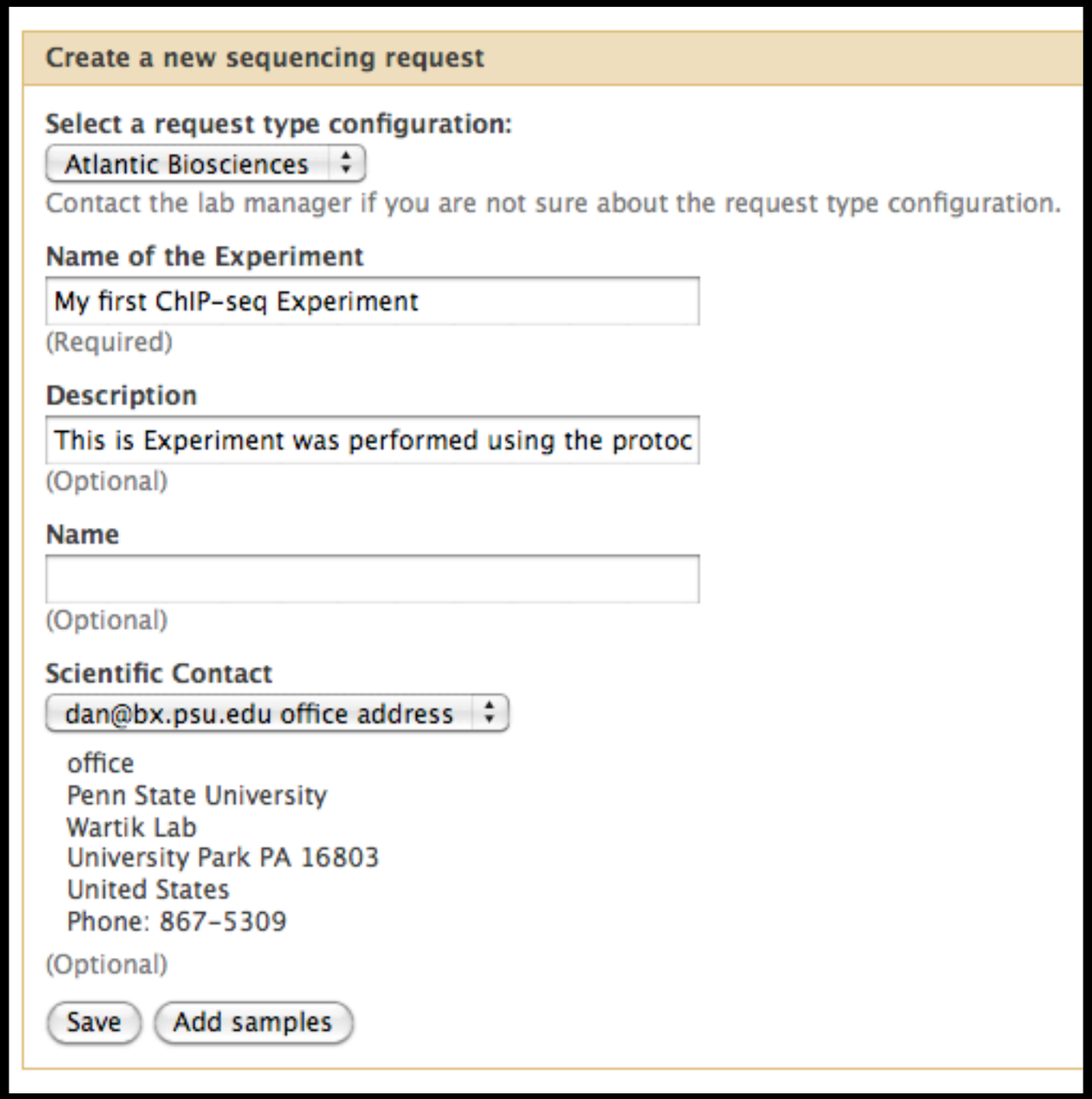


Customer Creates a Request

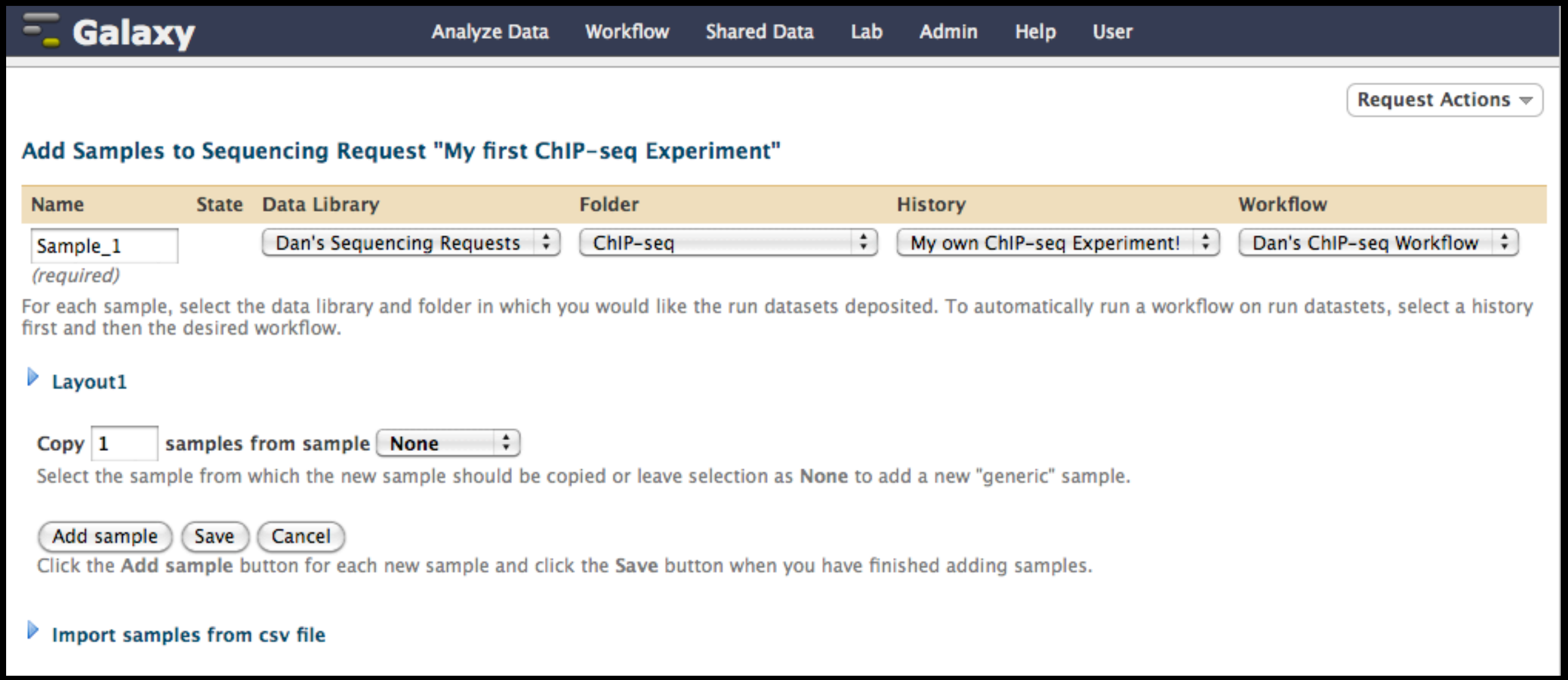




Customer Describes Request



Customer Adds a Sample



- Provide a name for the sample
- Data Library and Folder
- 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 datastets, select a history first and then the desired workflow.

Layout1

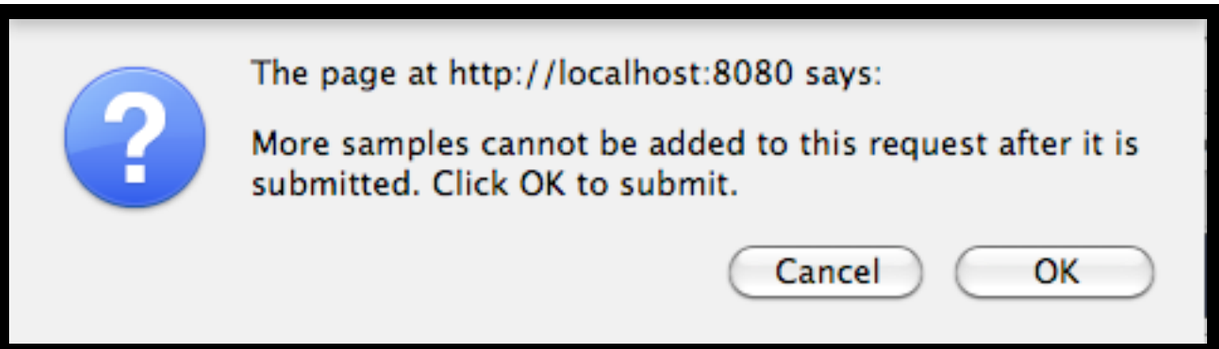
Copy 1 **samples from sample** None

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

Add sample

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

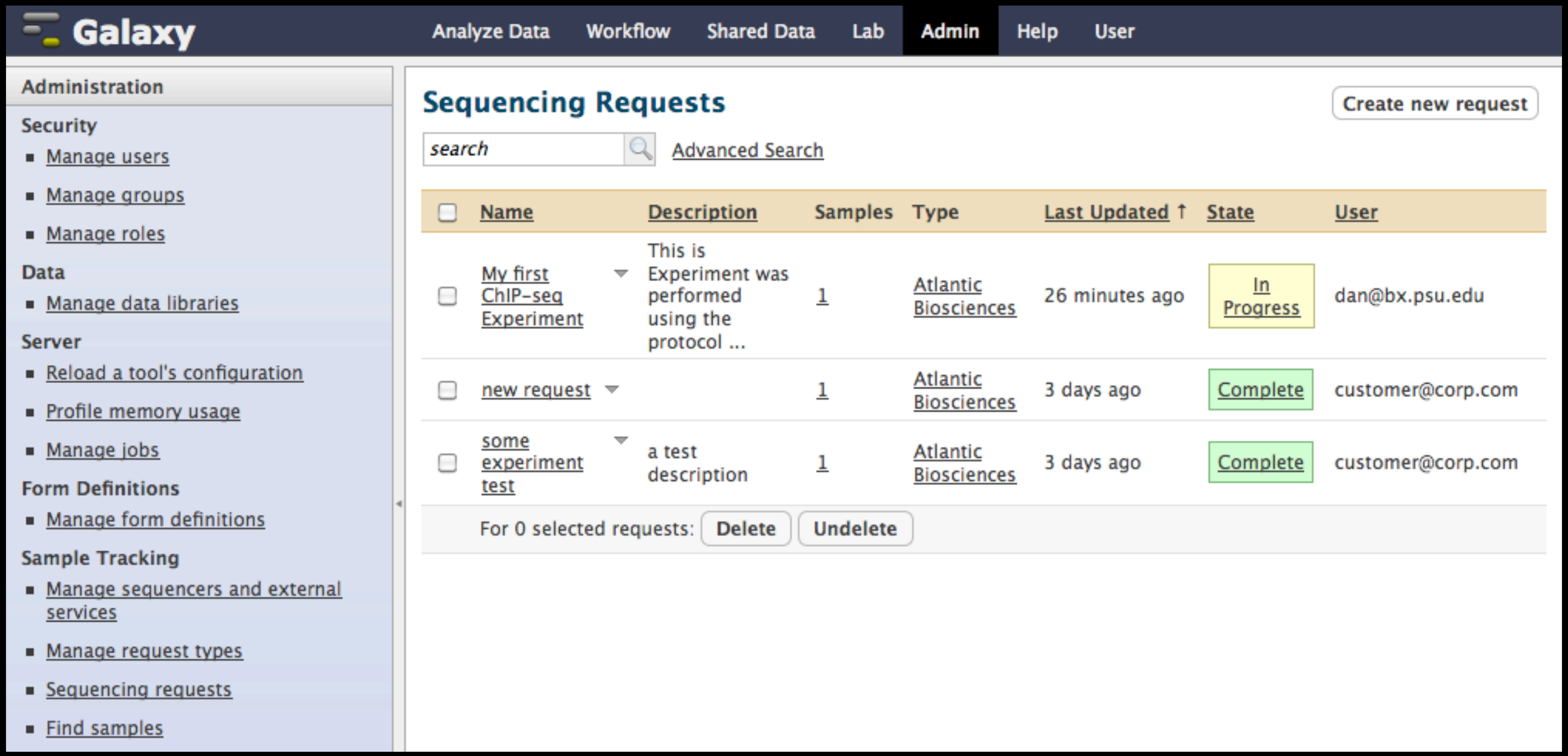
Edit samples

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

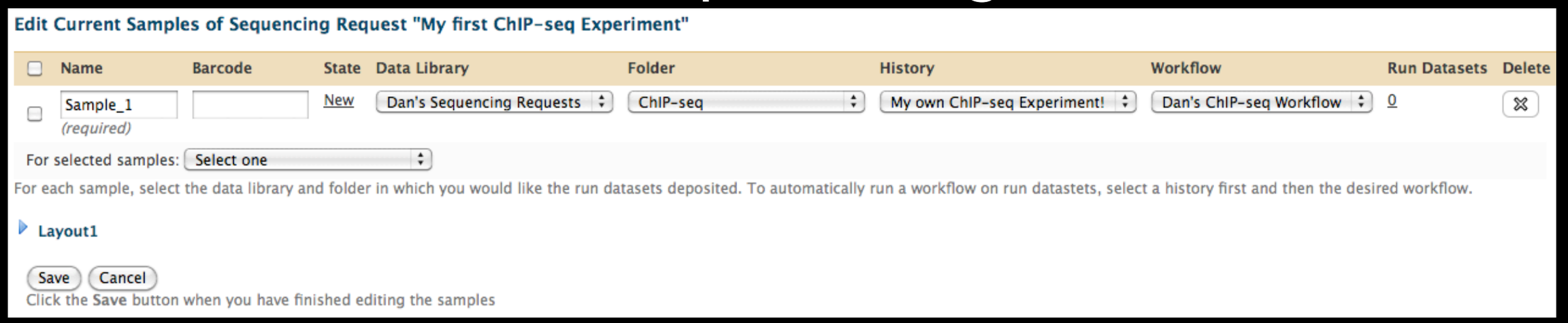
- Customer sends samples to sequencing facility

Sequencing Facility is informed of Request

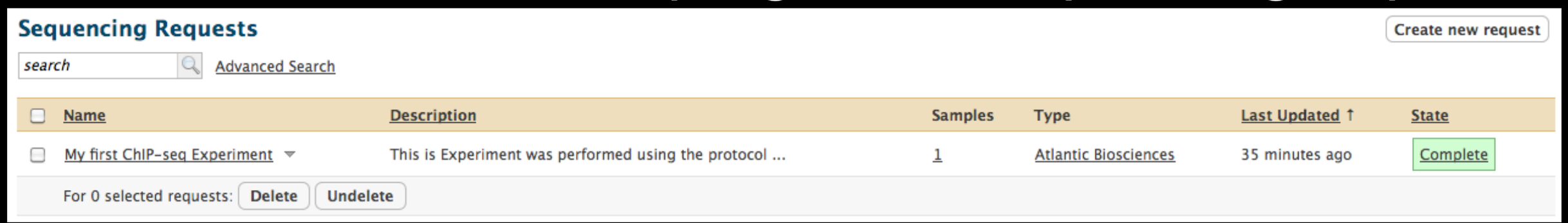


Sequencing Facility Receives Samples

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



- Customer 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
- Customer is automatically emailed

Extending Sample Tracking with ngLims

- An add-on written by community contributor Brad Chapman
- http://bitbucket.org/chapmanb/galaxy-central
- https://bitbucket.org/galaxy/galaxy-central/wiki/LIMS/nglims
- http://bcbio.wordpress.com/2011/01/11/next-generation-sequencing-information-management-and-analysis-system-for-galaxy/

Using Galaxy

- Use public Galaxy server: UseGalaxy.org
- Download Galaxy source: GetGalaxy.org
- Galaxy Wiki: GalaxyProject.org
- Screencasts: GalaxyCast.org
- Public Mailing Lists
  - galaxy-bugs@bx.psu.edu
  - galaxy-user@bx.psu.edu
  - galaxy-dev@bx.psu.edu

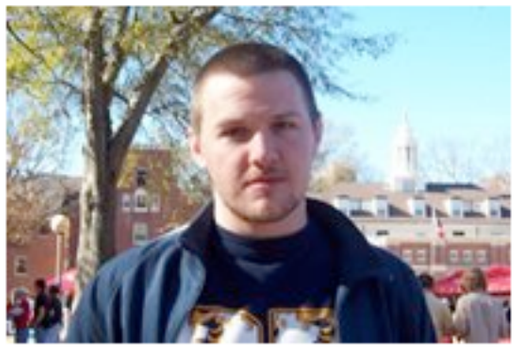
Acknowledgments

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

Galaxy Team

Enis Afgan | Emory

Guru Ananda | Penn State



Dannon Baker | Emory

James Taylor | Emory

Jeremy Goecks | Emory

Sergei Kosakovsky Pond | UCSD

Greg von Kuster | Penn State

Dave Clements | Emory

Nate Coraor | Penn State

+ Jennifer Jackson

Ross Lazarus | Harvard | BakerIDI

Kanwei Li | Emory

Anton Nekrutenko | Penn State

Kelly Vincent | Penn State

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

Galaxy 2011

Community Conference

25-26 May Lunteren, The Netherlands

http://galaxy.psu.edu/gcc2011