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:

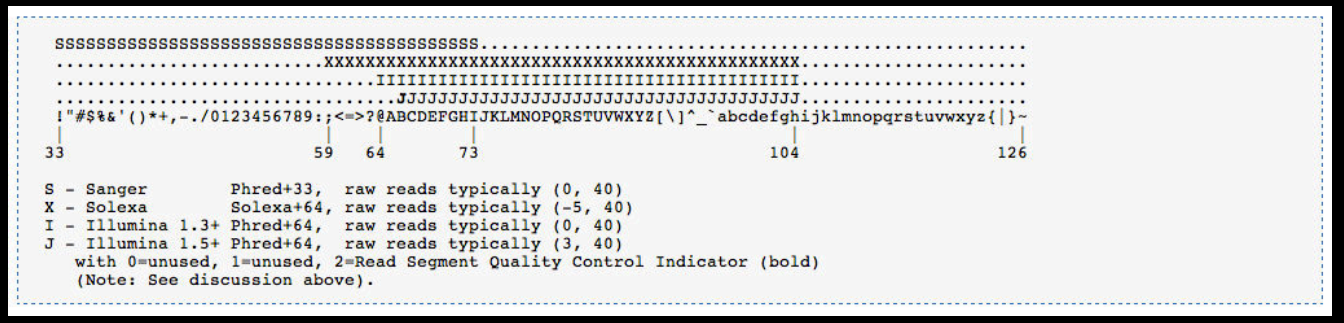
- reads and quality scores (FASTQ)

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>>>>>>CCCCCCC65



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

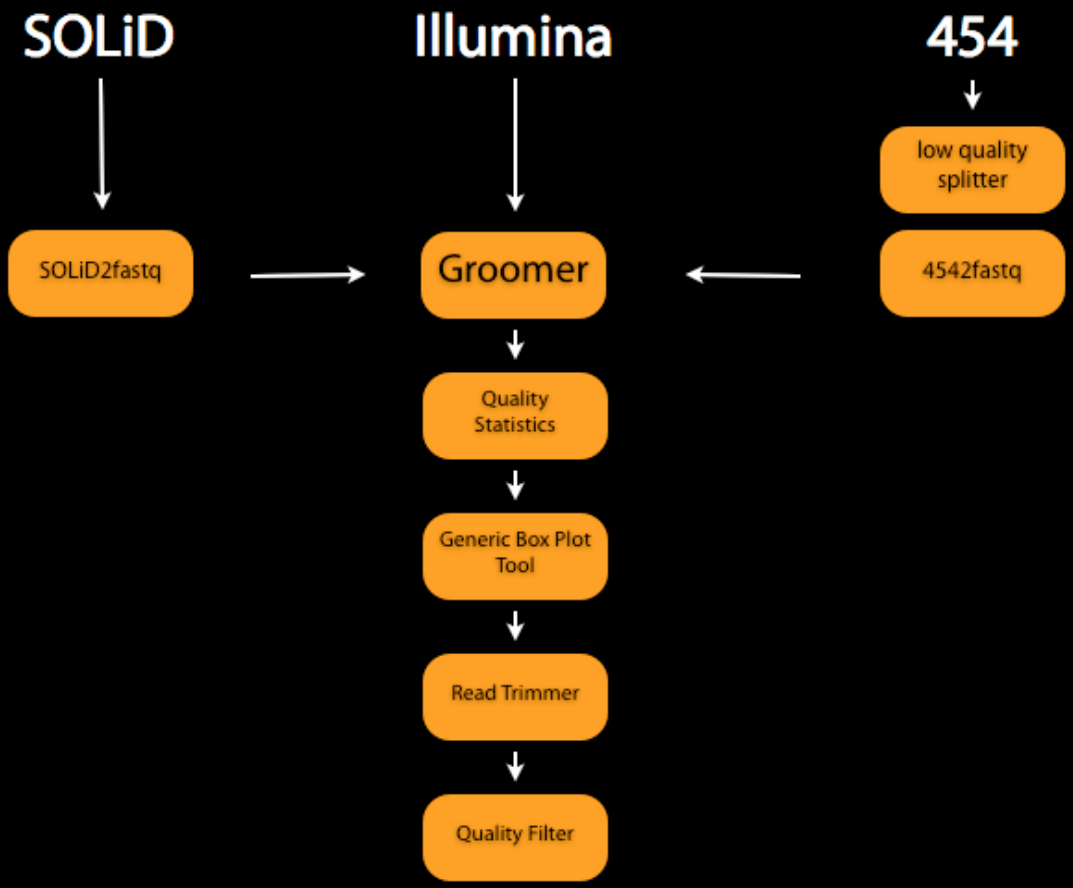
Using Galaxy to Analyze HTS Data

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

Galaxy for Sequencing Facilities

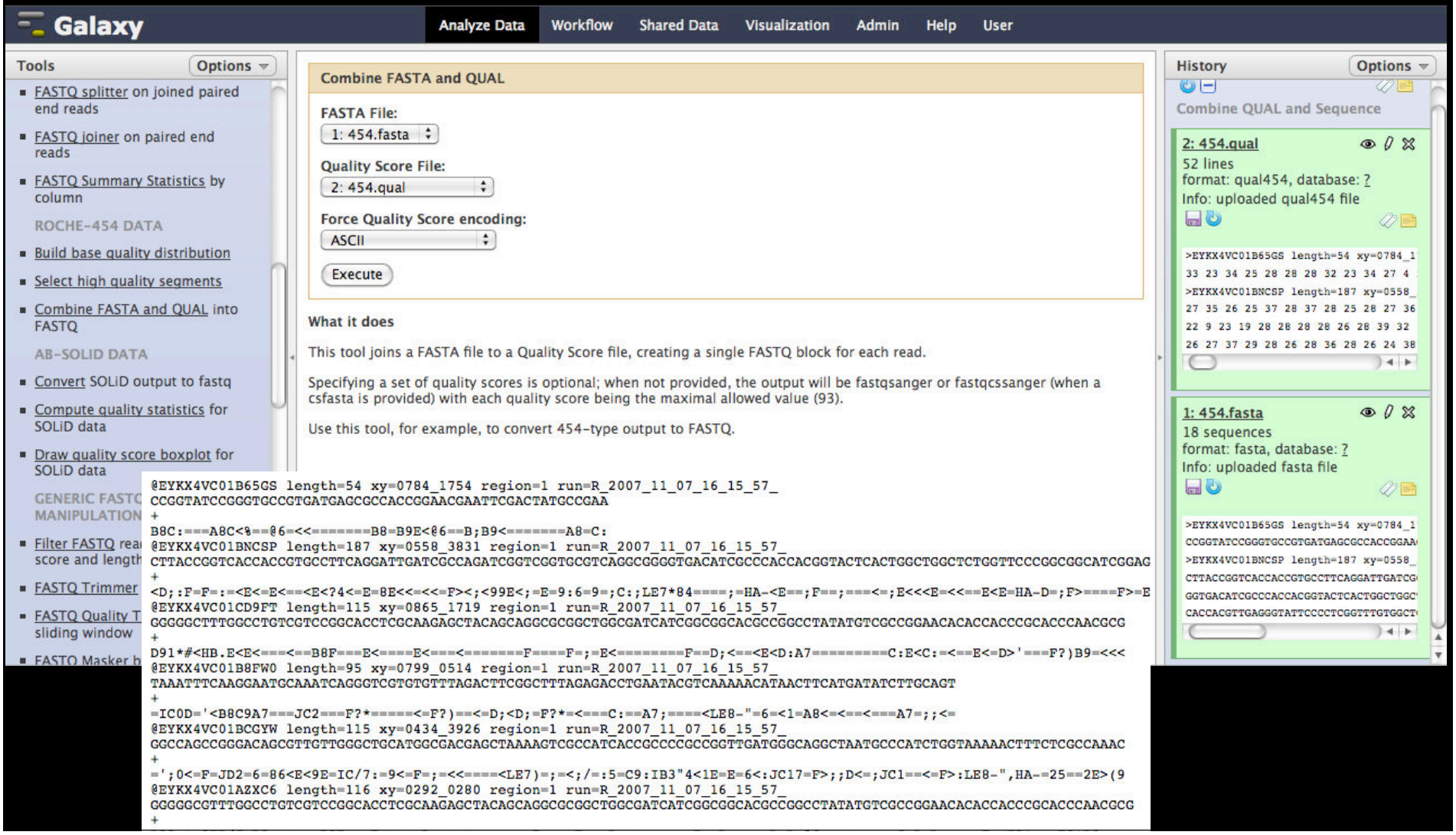
Galaxy exercises: ChIP-seq and RNA-seq

Prepare and Quality Check

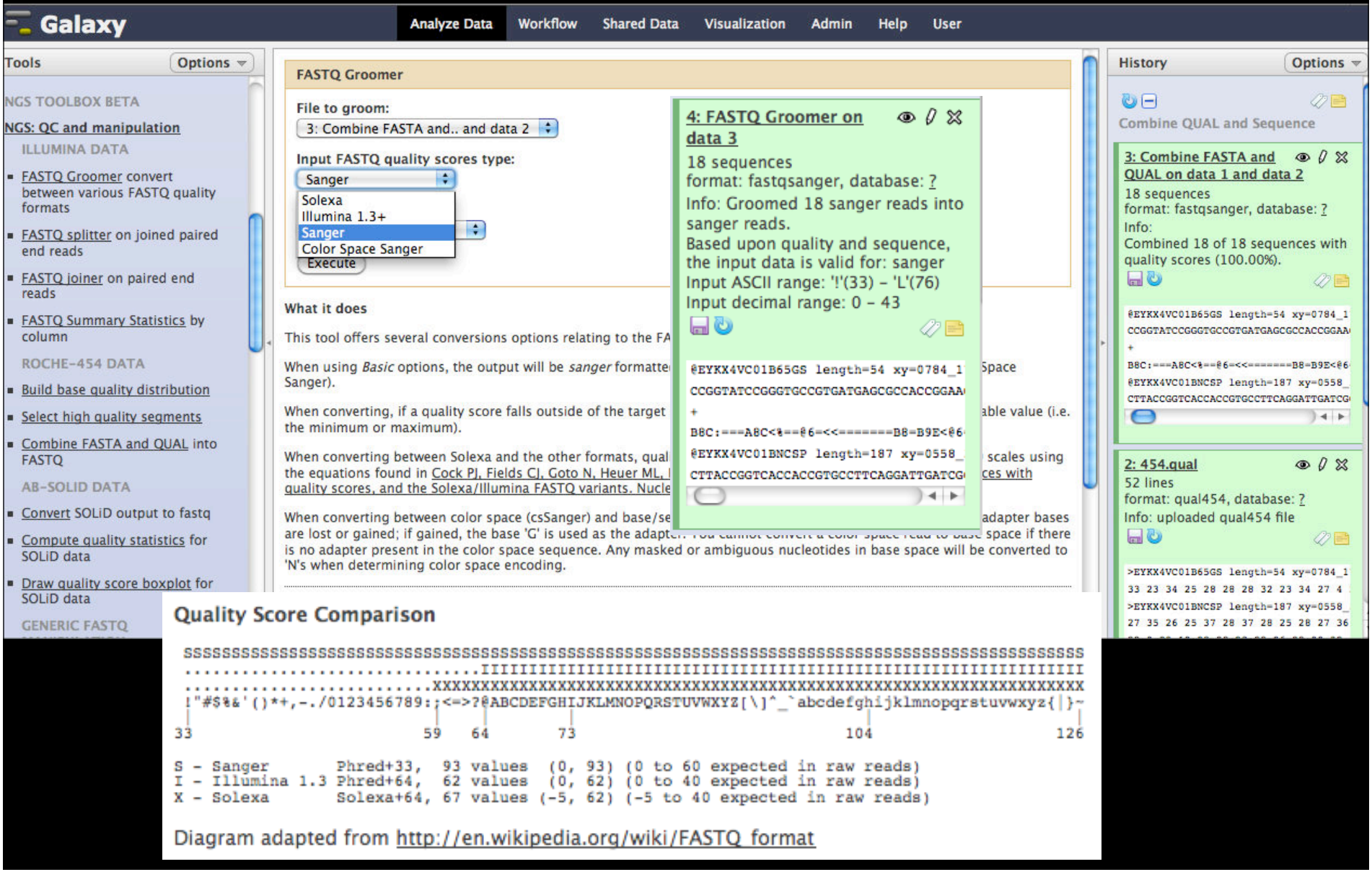


Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5.

Combining Sequences and Qualities



Grooming --> Sanger



Quality Score Comparison


```
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

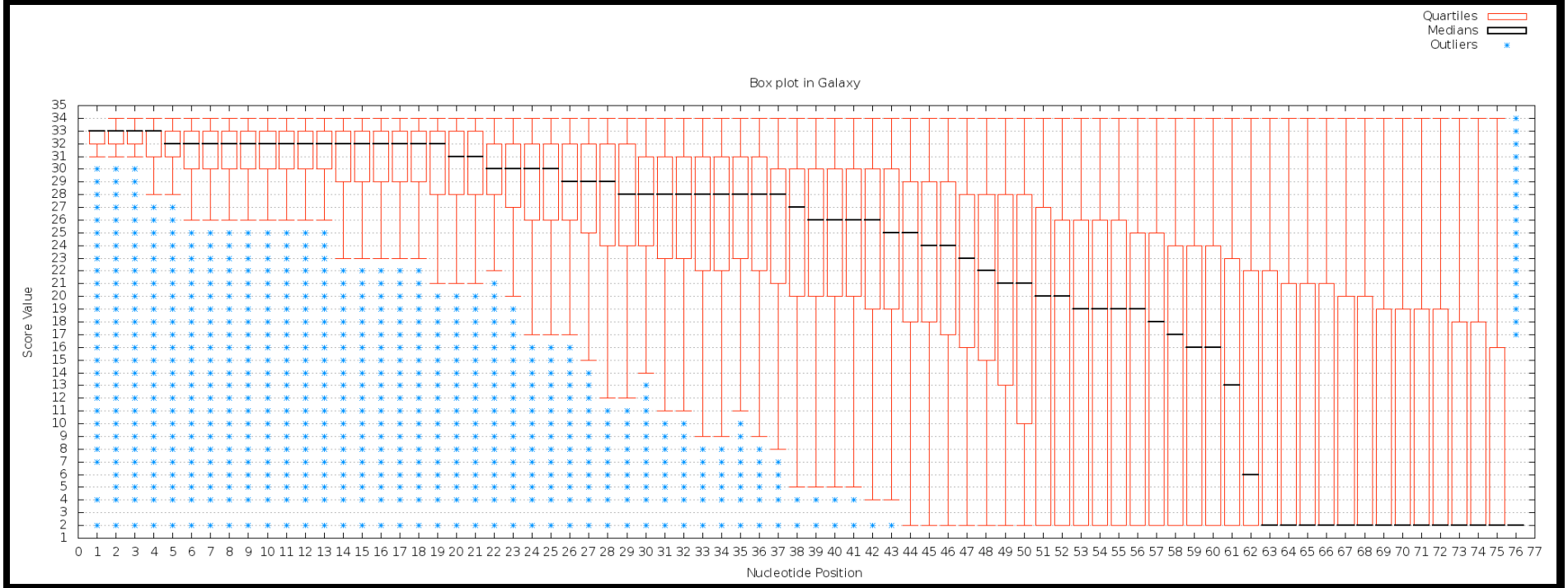
NGS: QC and manipulation

ILLUMINA DATA

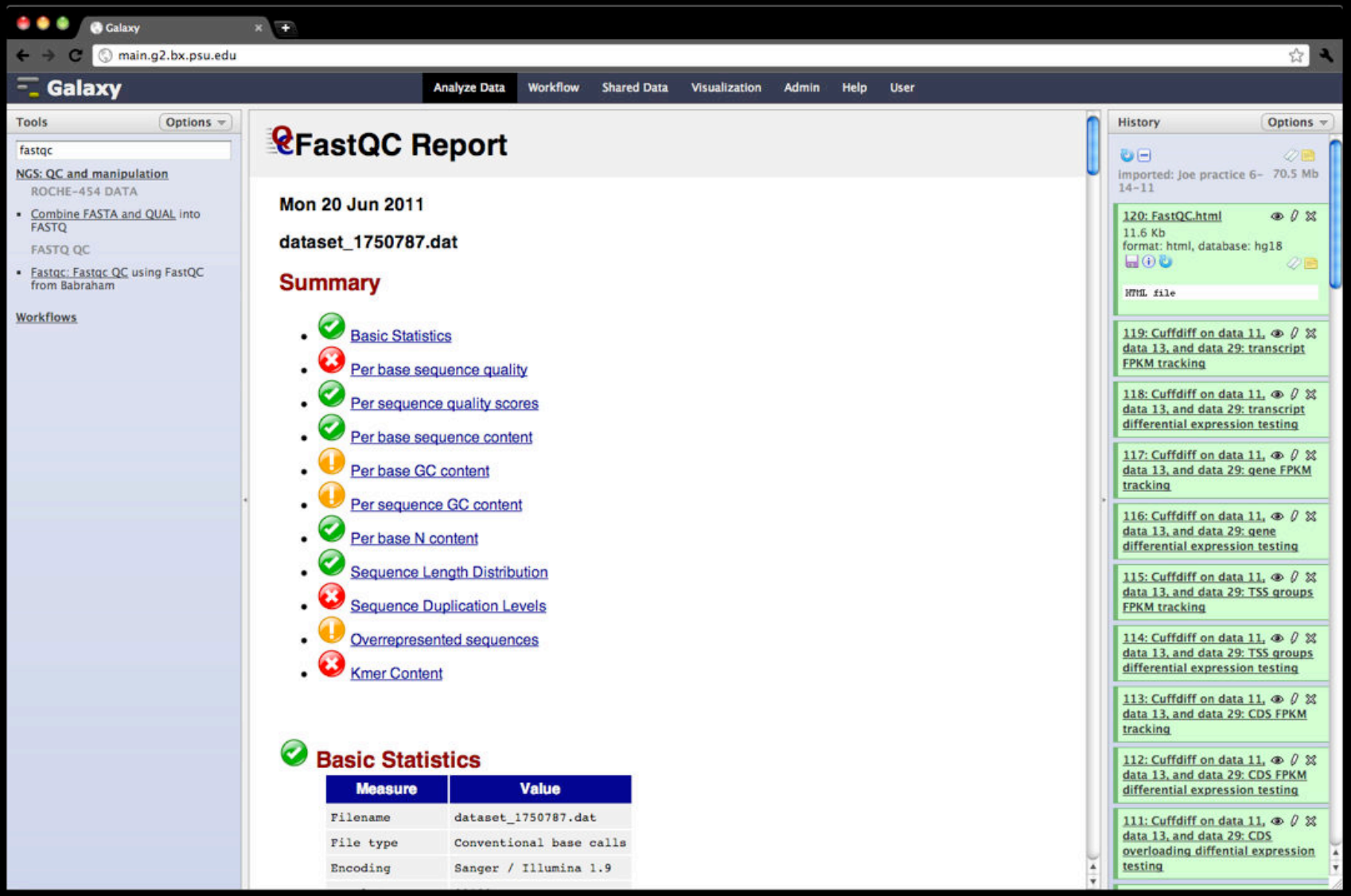
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

Graph/Display Data

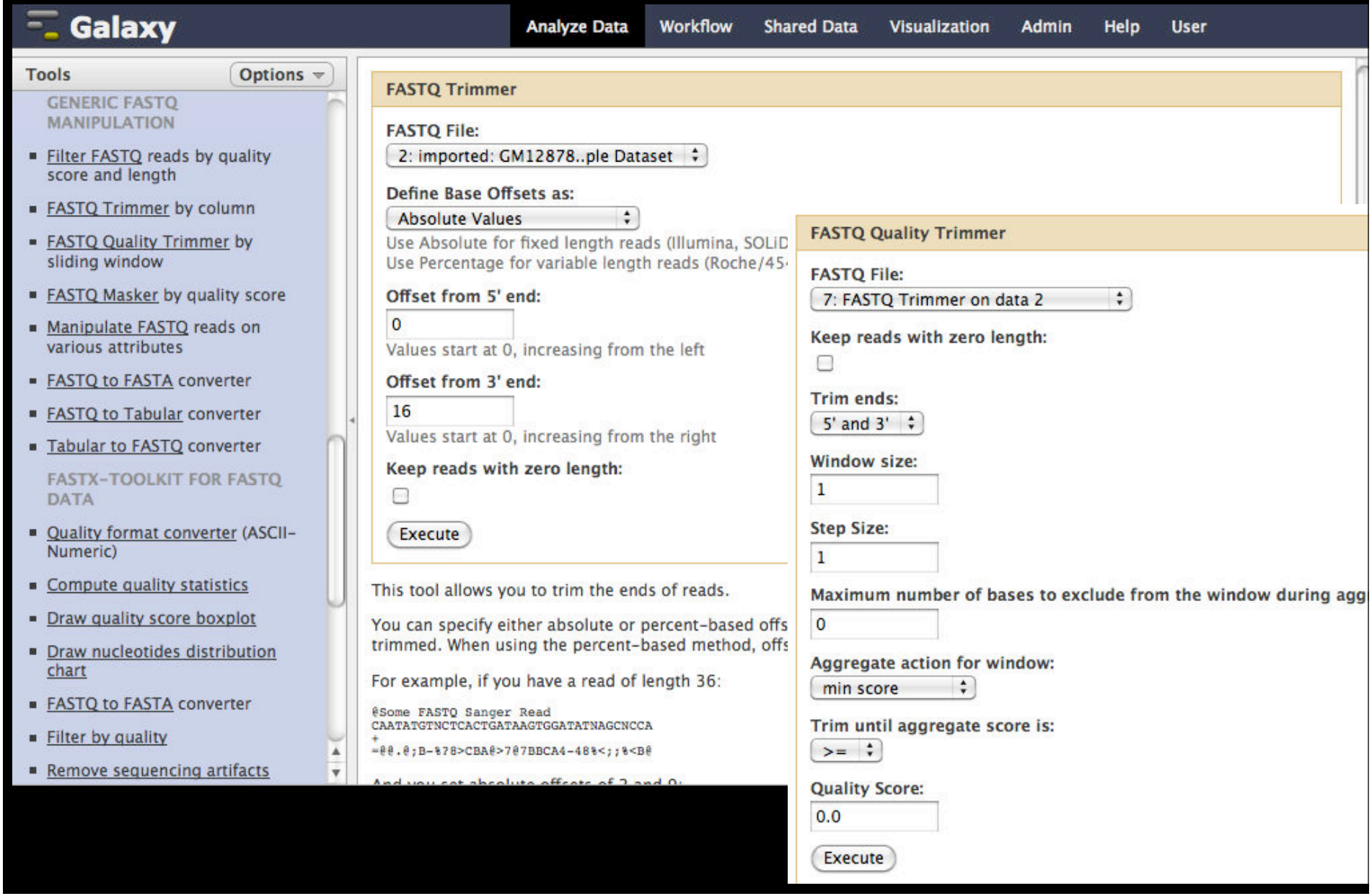
- Histogram of a numeric column
- Scatterplot of two numeric columns
- Plotting tool for multiple series and graph types
- Boxplot of quality statistics

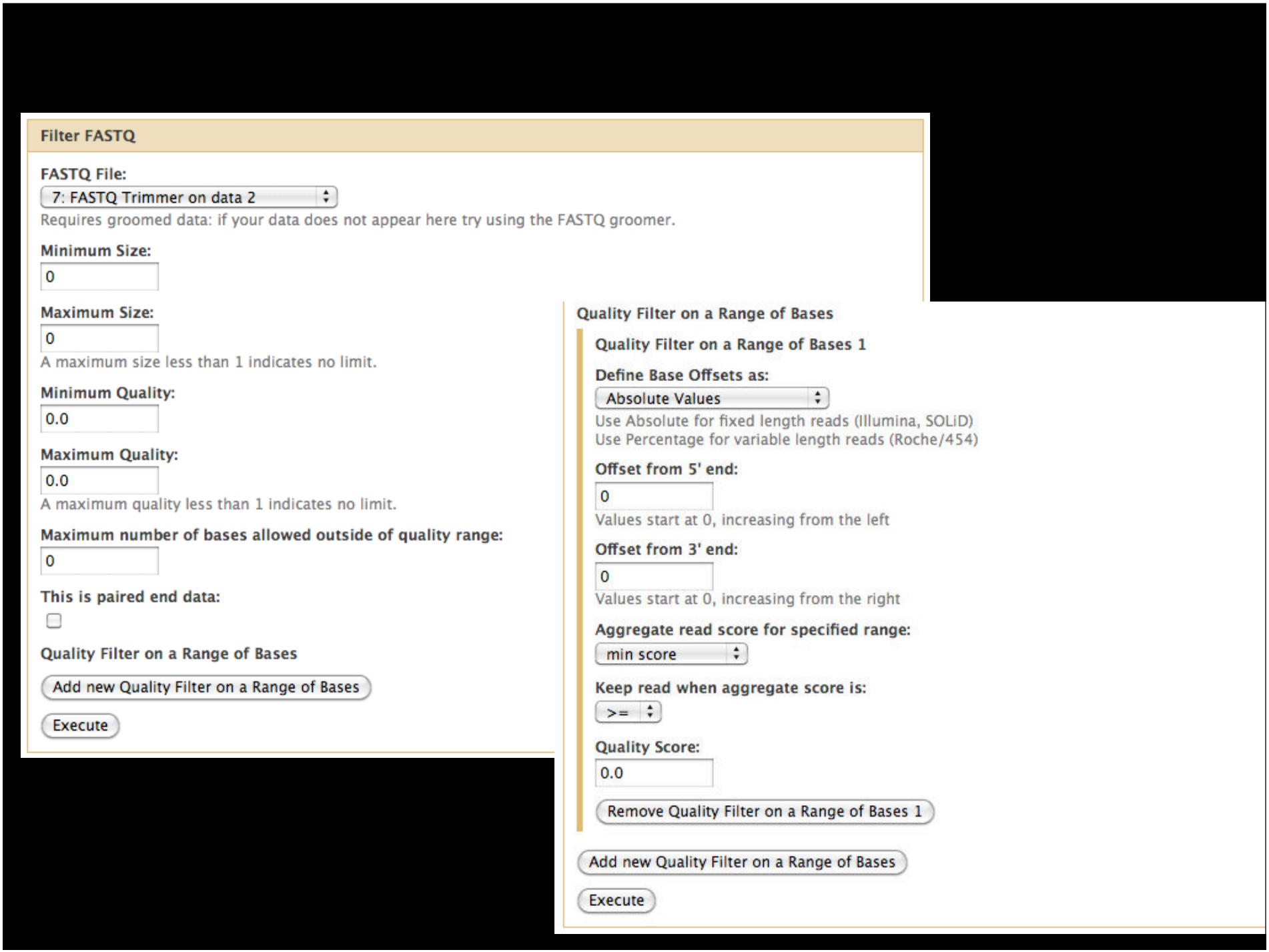


FastQC

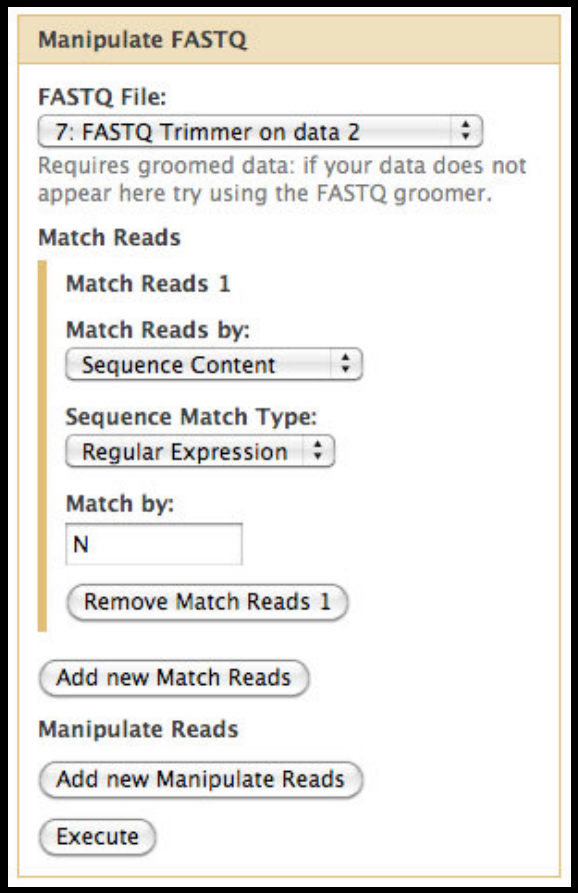

Galaxy
main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Tools
Options
fastqc
NGS: QC and manipulation
ROCHE-454 DATA
Combine FASTA and QUAL into FASTQ
FASTQ QC
Fastqc: Fastqc QC using FastQC from Babraham
Workflows
FastQC Report
Mon 20 Jun 2011
dataset_1750787.dat
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

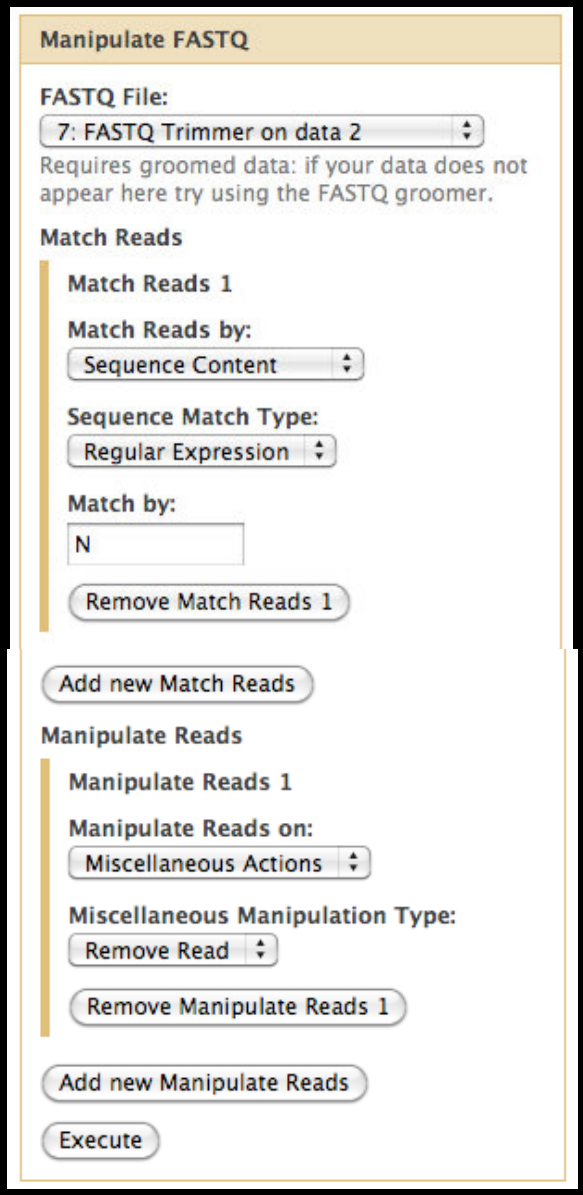
Read Trimming




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





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
- longer reads: LASTZ

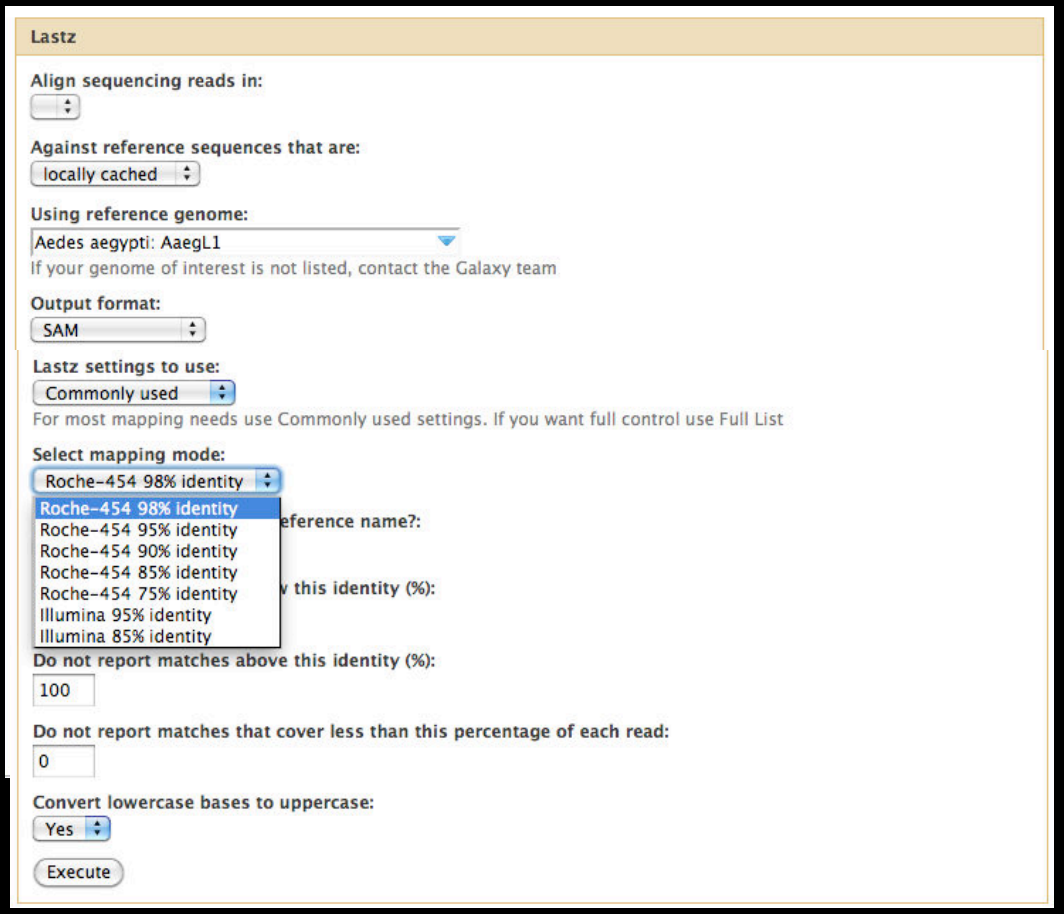
Metagenomics

- Megablast

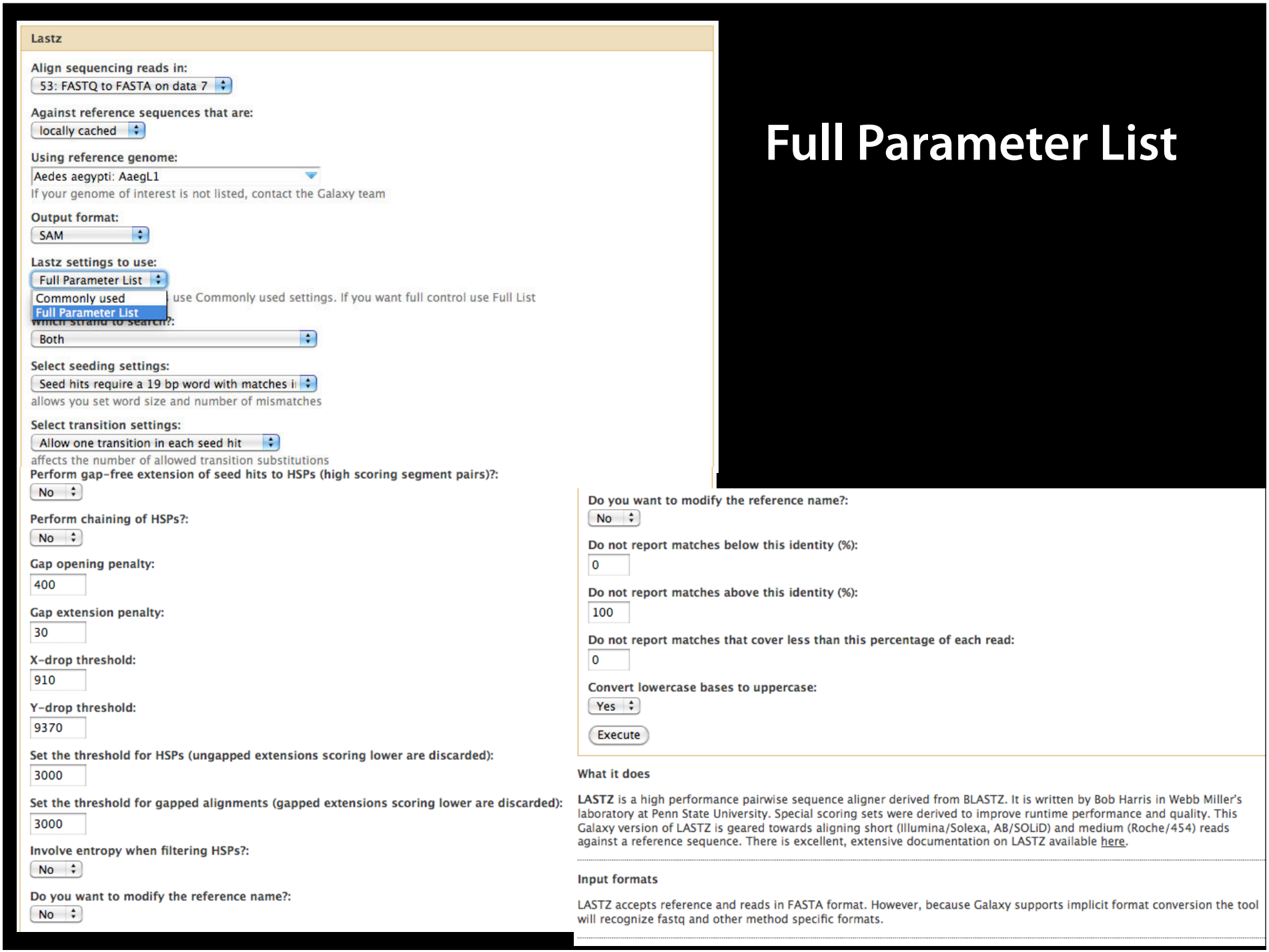
RNA / gapped-reads mapper

- Tophat

Commonly Used/Default Parameters



Full Parameter List



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

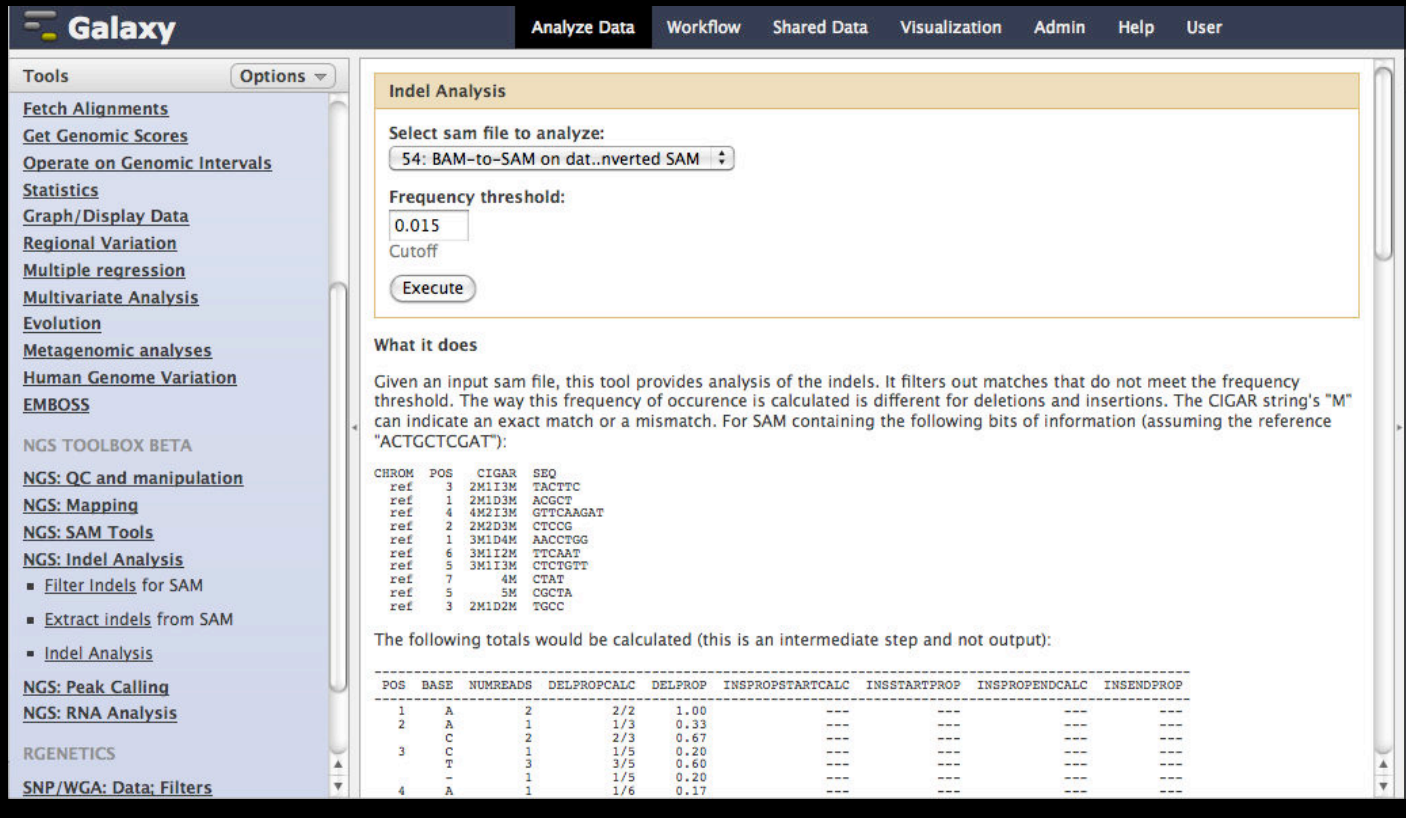
SNPs & INDELs

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
- 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/

Unified Genotyper

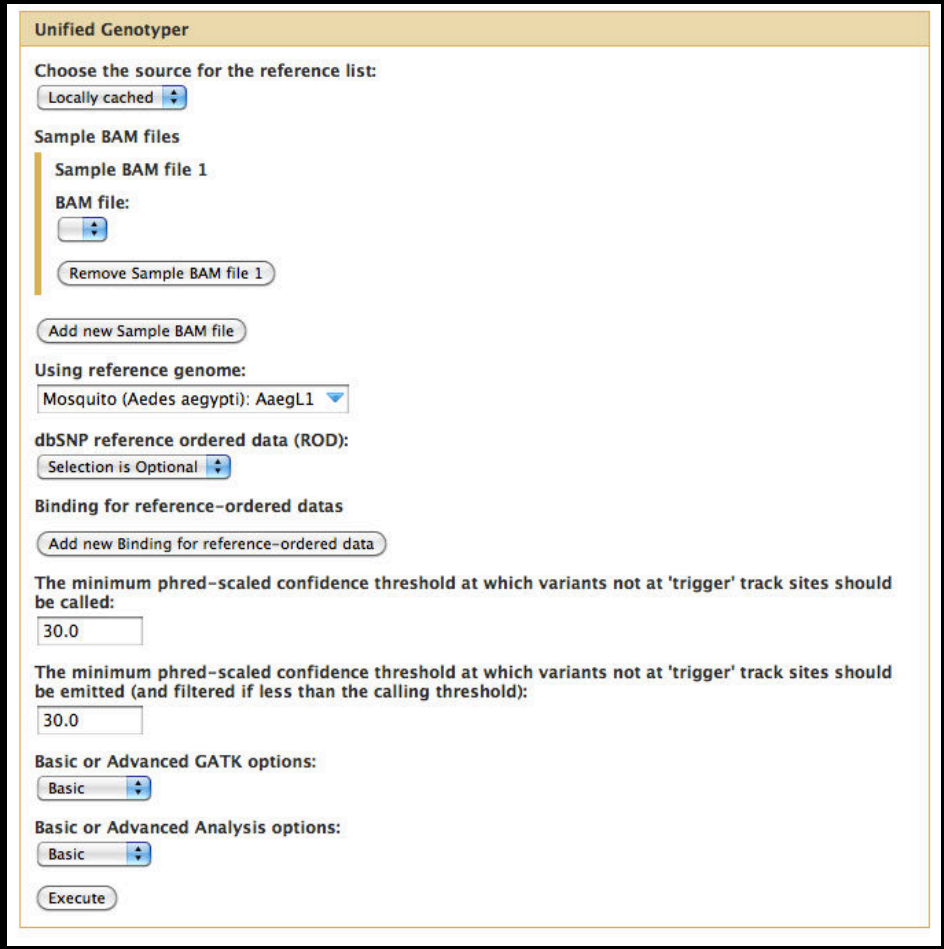
Inputs

- 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 for Sequencing Facilities

Galaxy exercises: ChIP-seq and RNA-seq

Peak Calling / ChIP-seq analysis

Punctate binding

- transcription factors

Diffuse binding

- histone modifications
- PolII

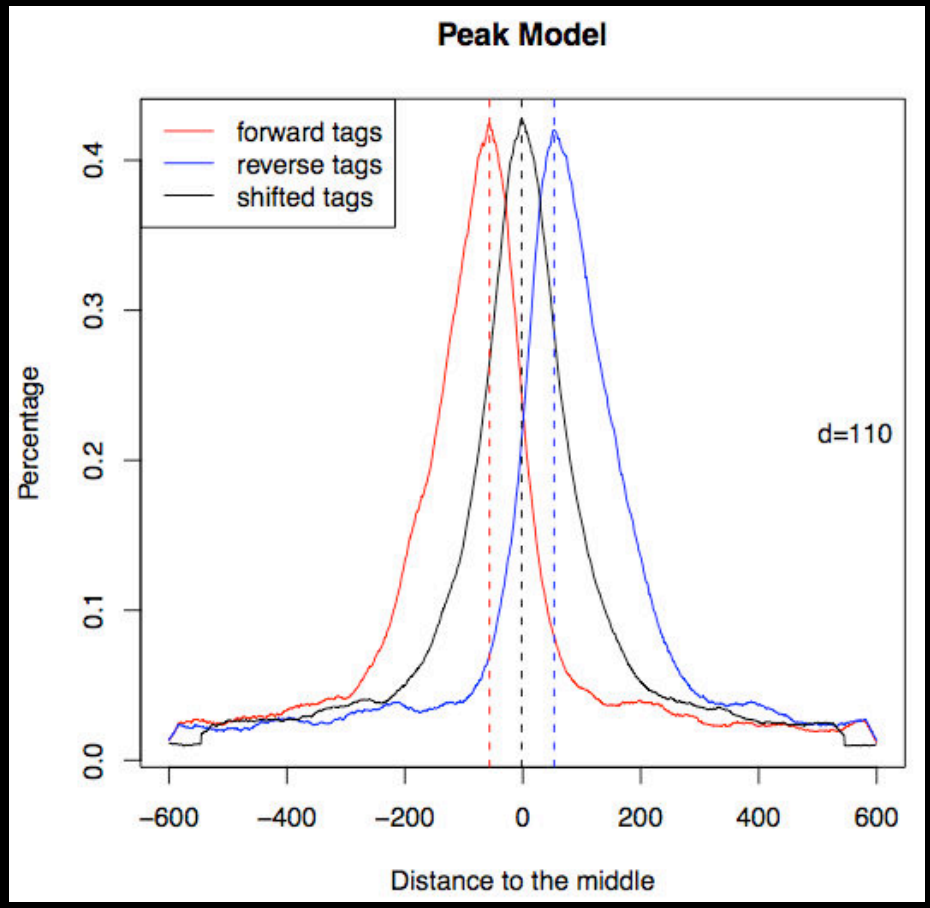
Punctate Binding --> MACS

Inputs

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

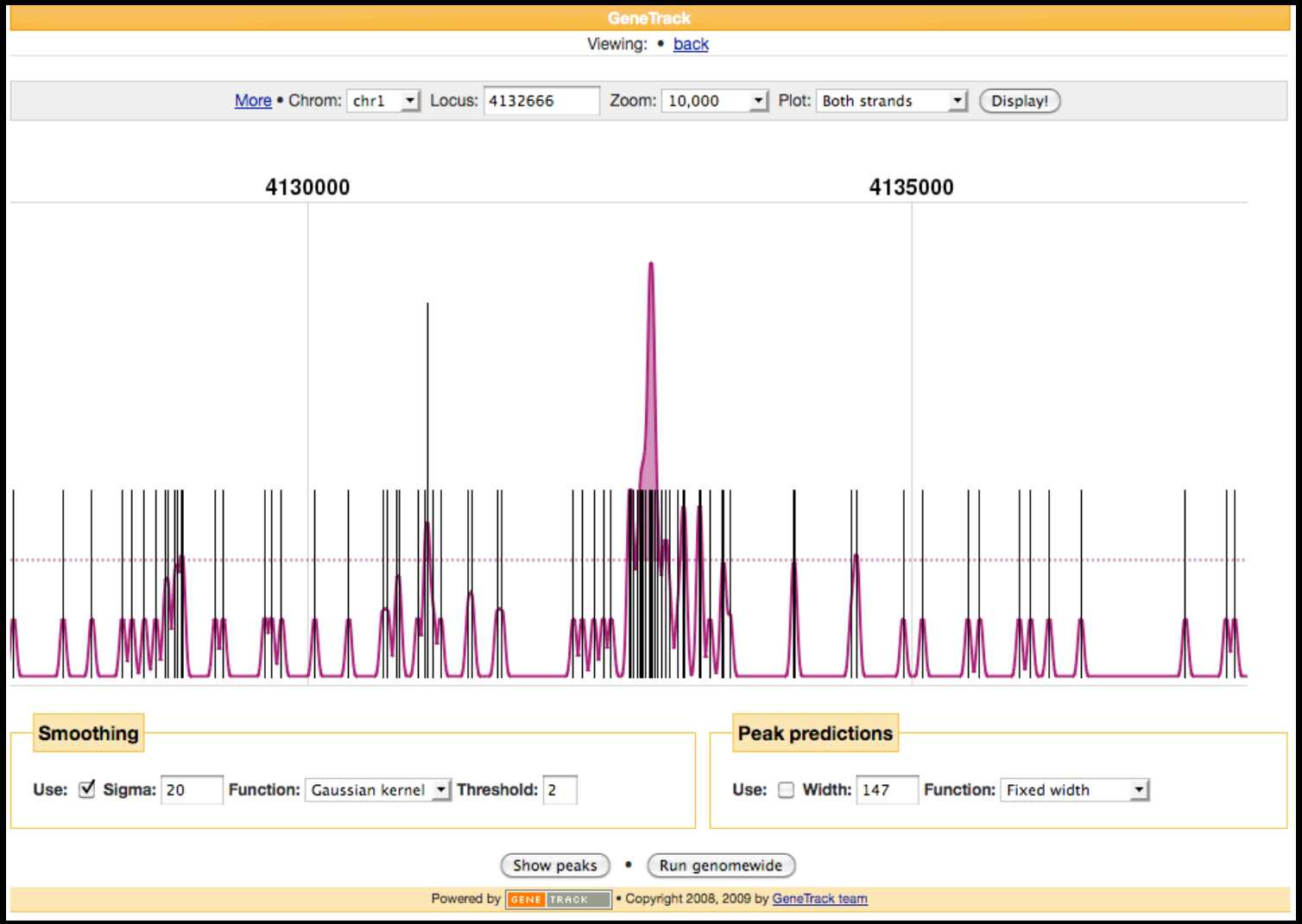
Outputs

- 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

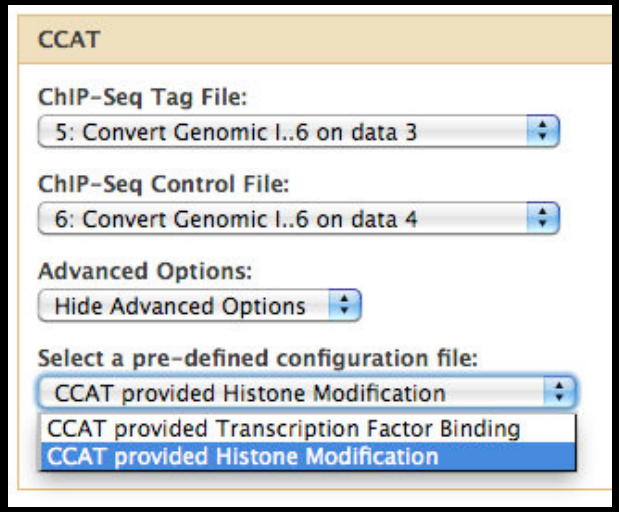
MACS --> GeneTrack

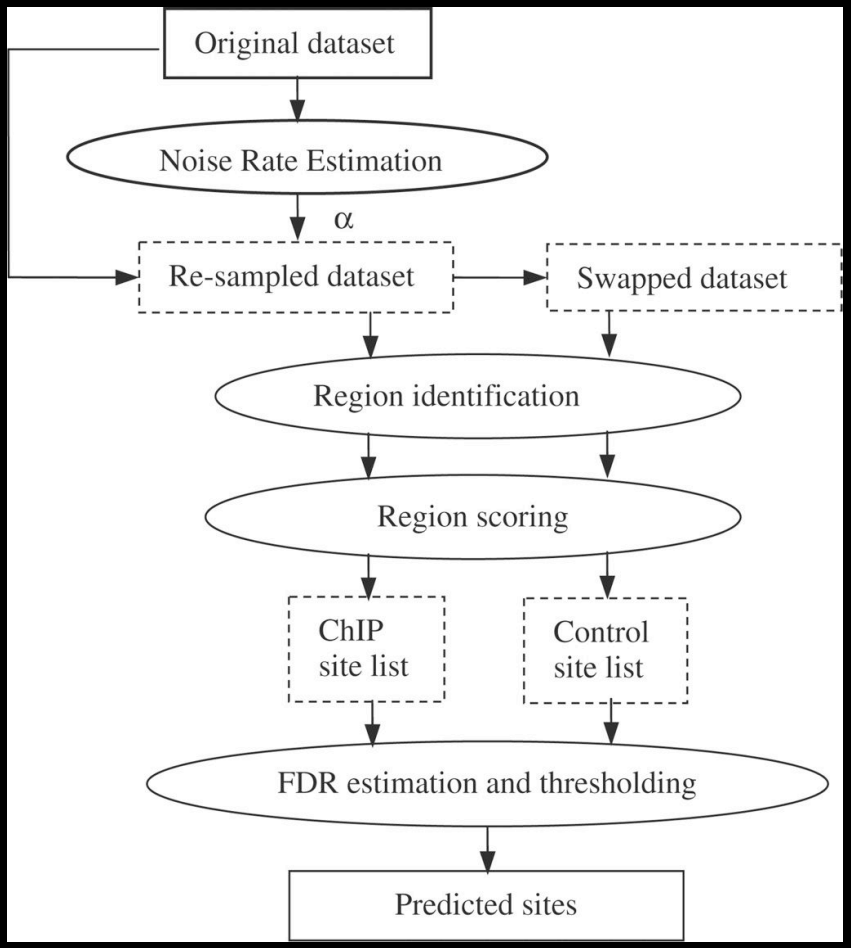


Albert I, Wachi S, Jiang C, Pugh BF. GeneTrack--a genomic data processing and visualization framework. Bioinformatics. 2008 May 15;24(10):1305-6. Epub 2008 Apr 3.

Diffuse Binding

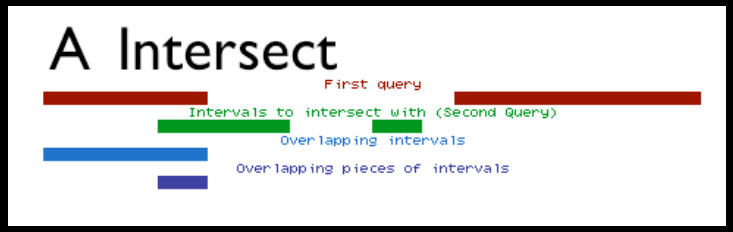
CCAT (Control-based ChIP-seq Analysis Tool)

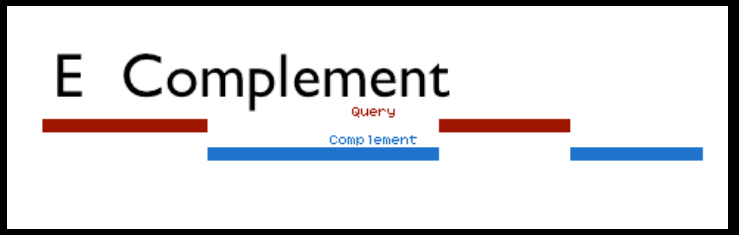


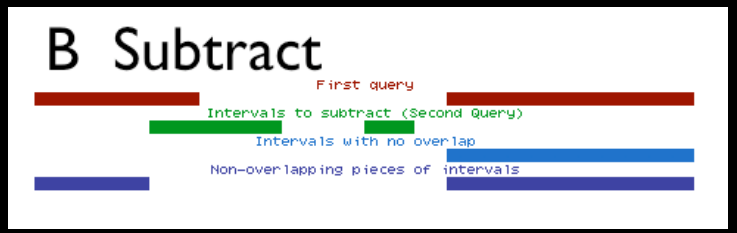


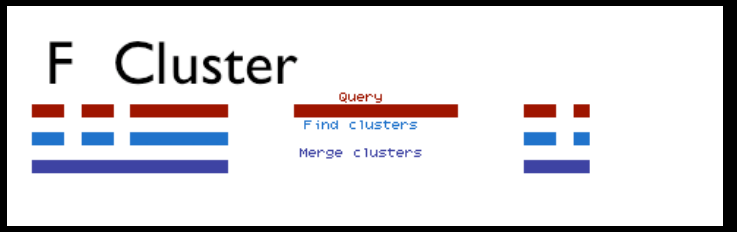
Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F, Sung WK. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics. 2010 May 1;26(9):1199-204.

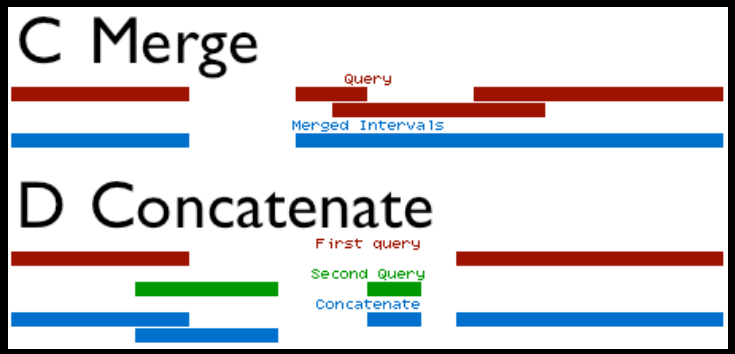
I have Peaks, now what?











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)

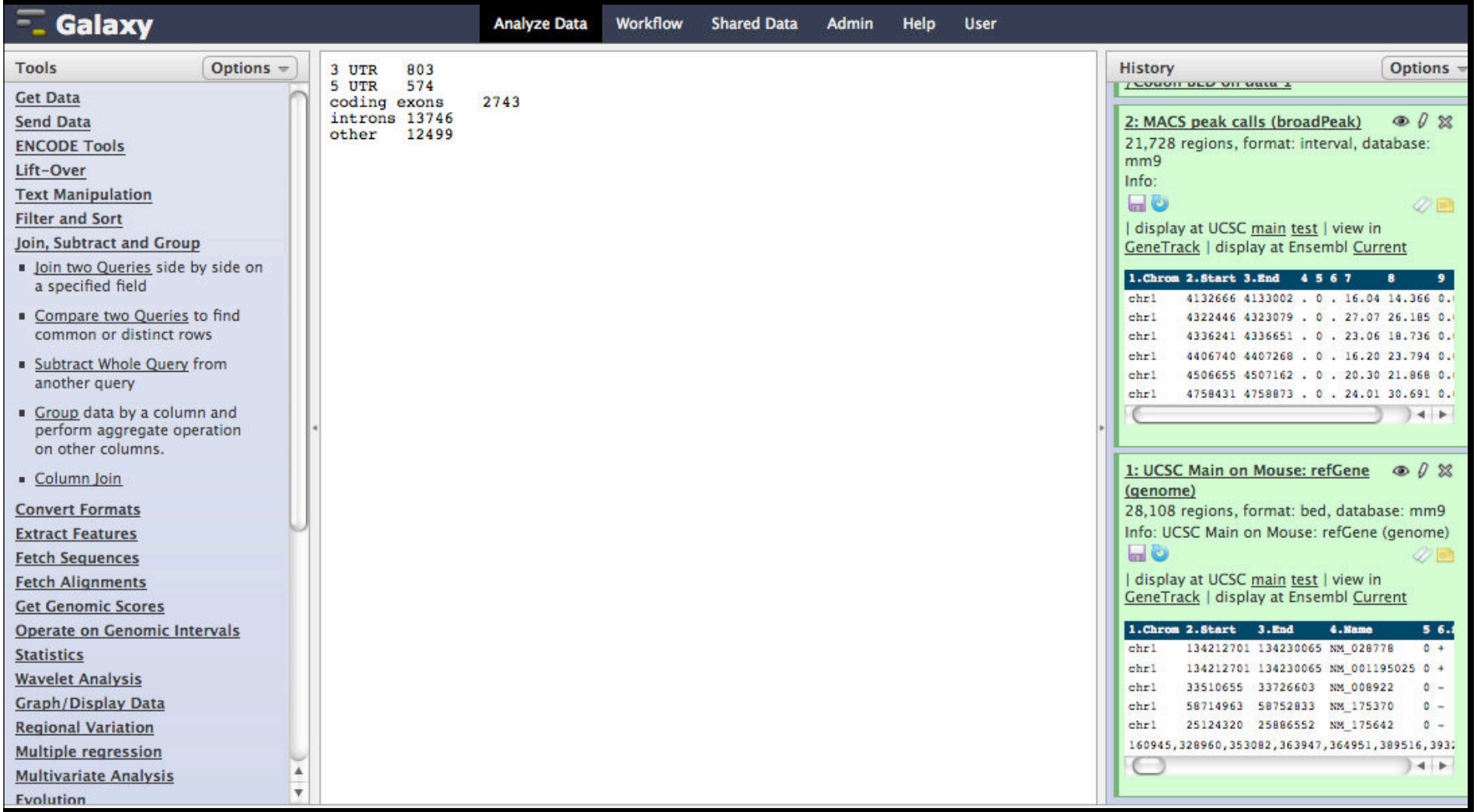
Create an Identifier column for the 'other' type (Add column)

Concatenate the exons/introns and other files

Determine which Peaks overlap the region types (Join)

Calculate counts for each region type (Group)

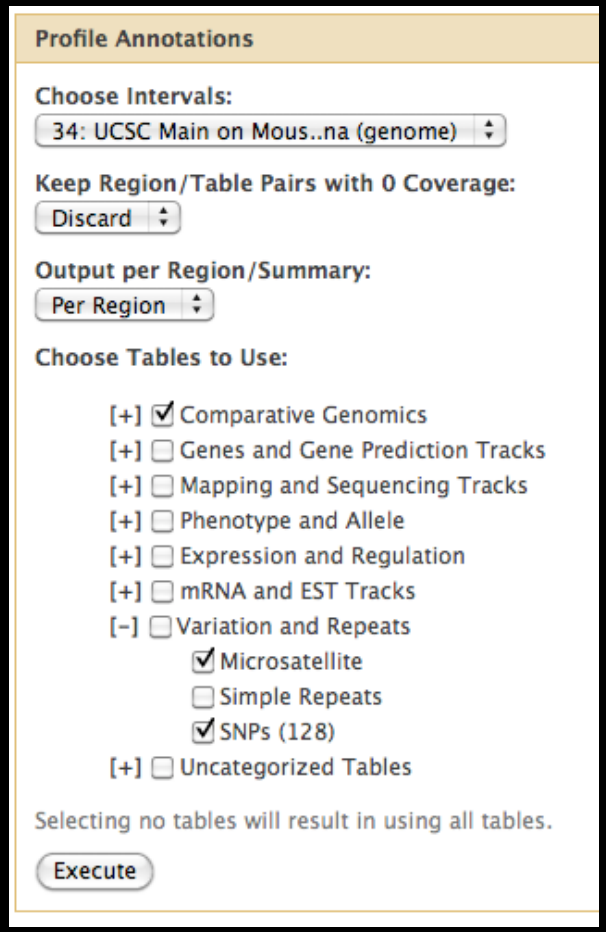
Secondary Analysis



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

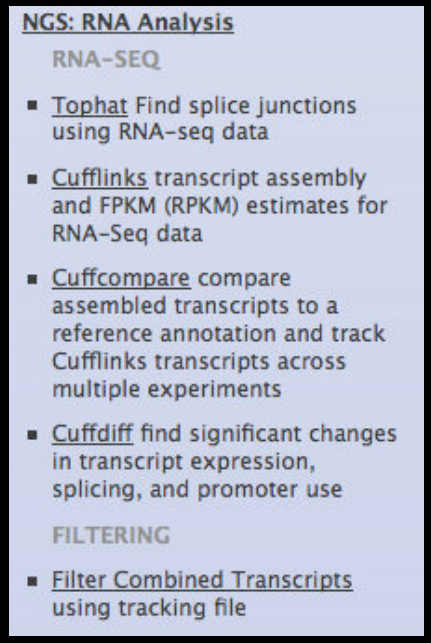
Galaxy for Sequencing Facilities

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

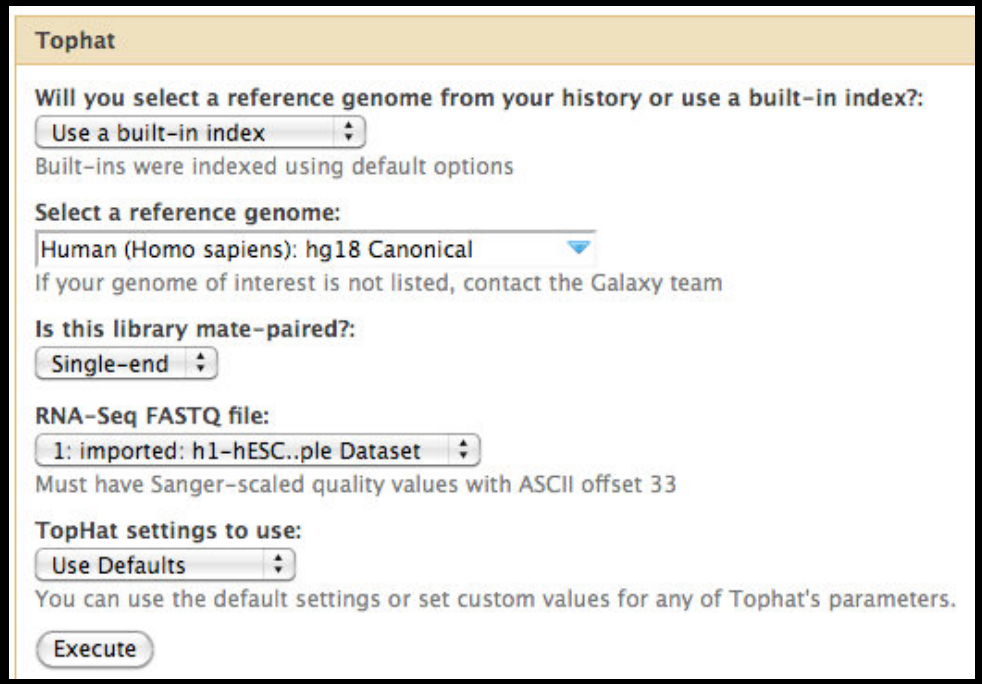
TopHat

Map RNA (FASTQ) to a reference Genome

- gapped mapper

Outputs

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



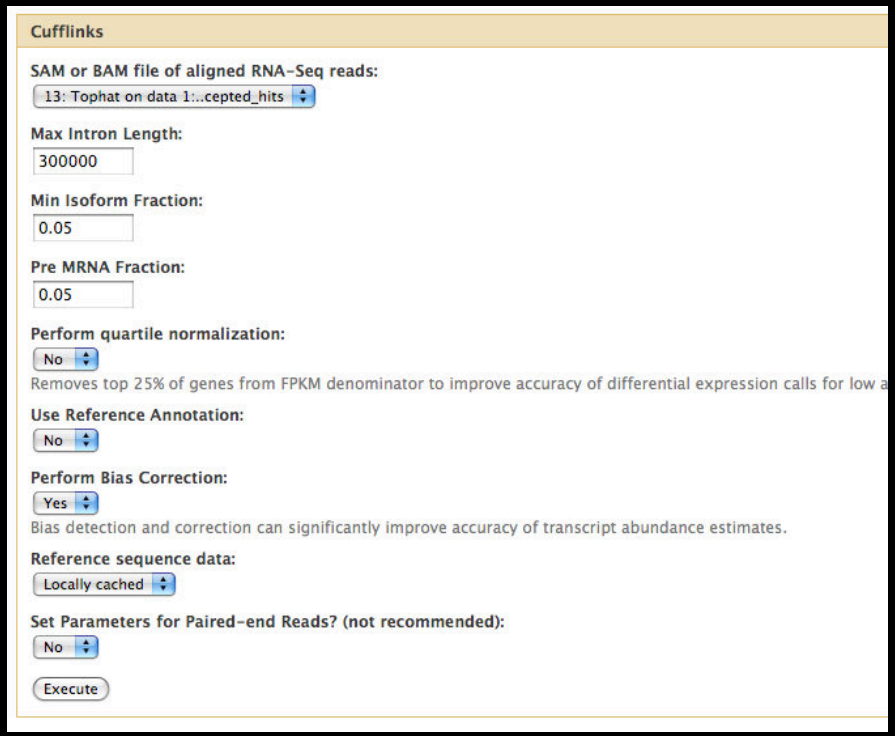
Cufflinks

Goal: transcript assembly and quantitation

Input: aligned RNA-Seq reads, usually from TopHat

Outputs

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



Cuffcompare

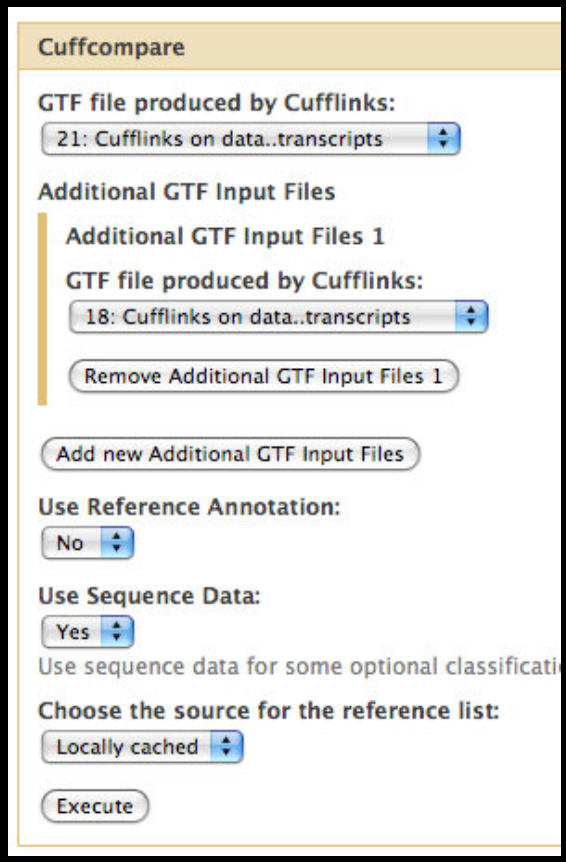
Goals

- 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

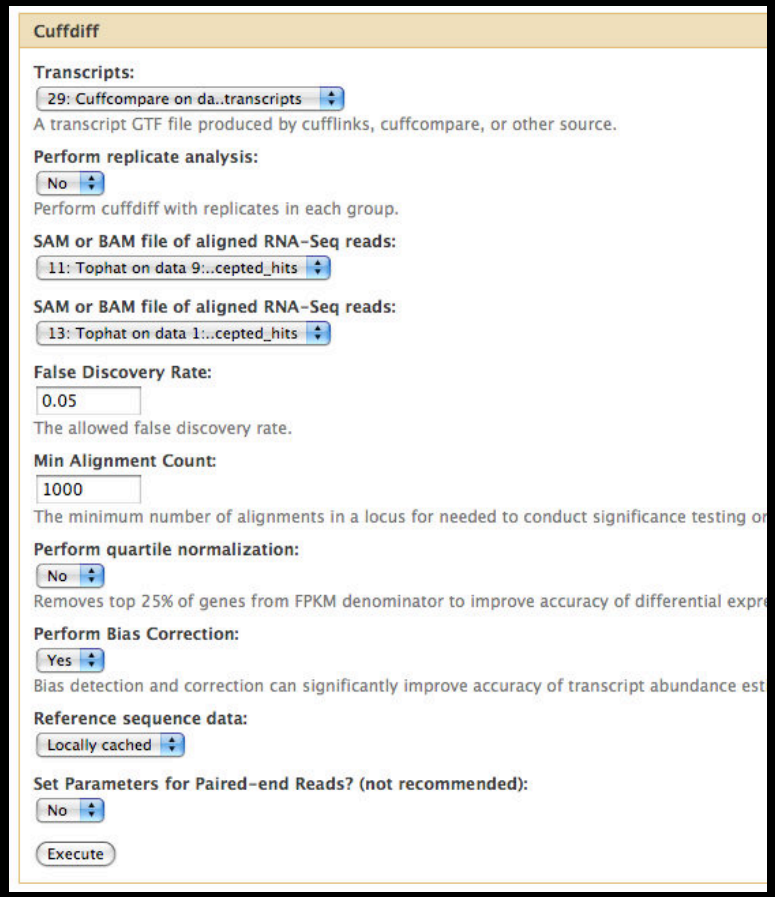
- differential expression testing
- transcript quantitation

Inputs

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

Outputs

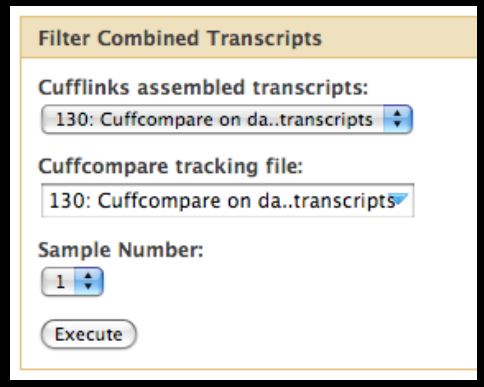
- differential expression tests for transcripts, genes, splicing, promoters, CDS
- quantitation values for most elements

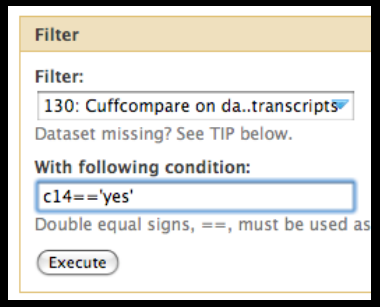


Next Steps

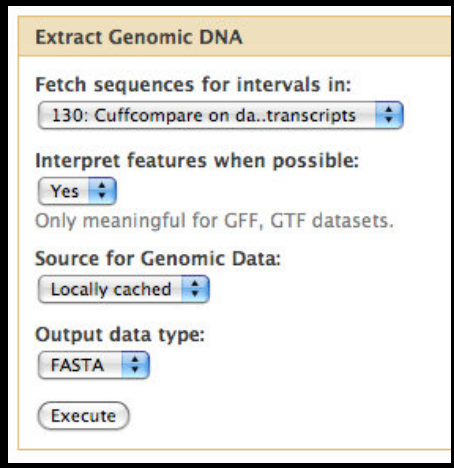
Filtering

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

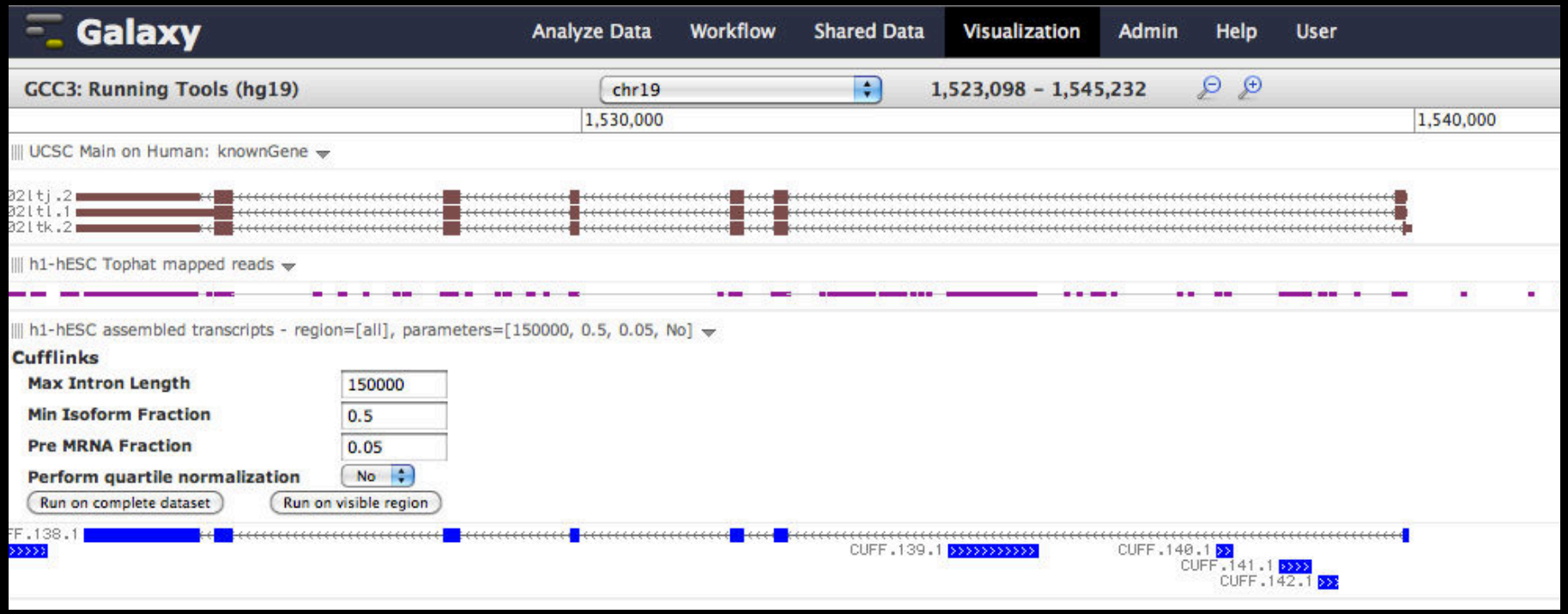


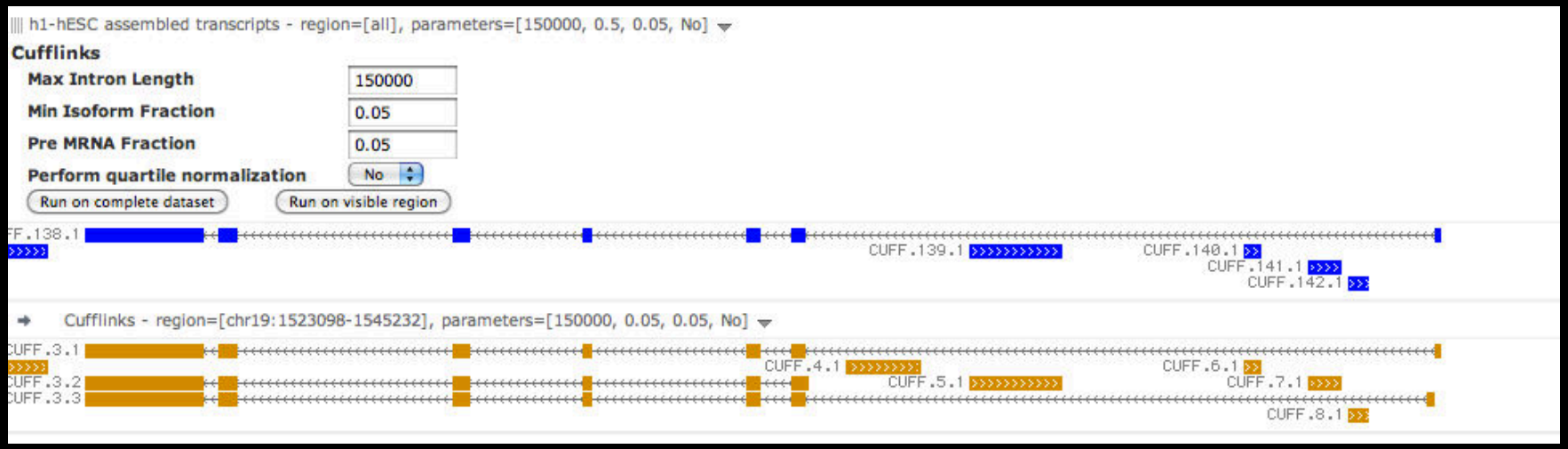


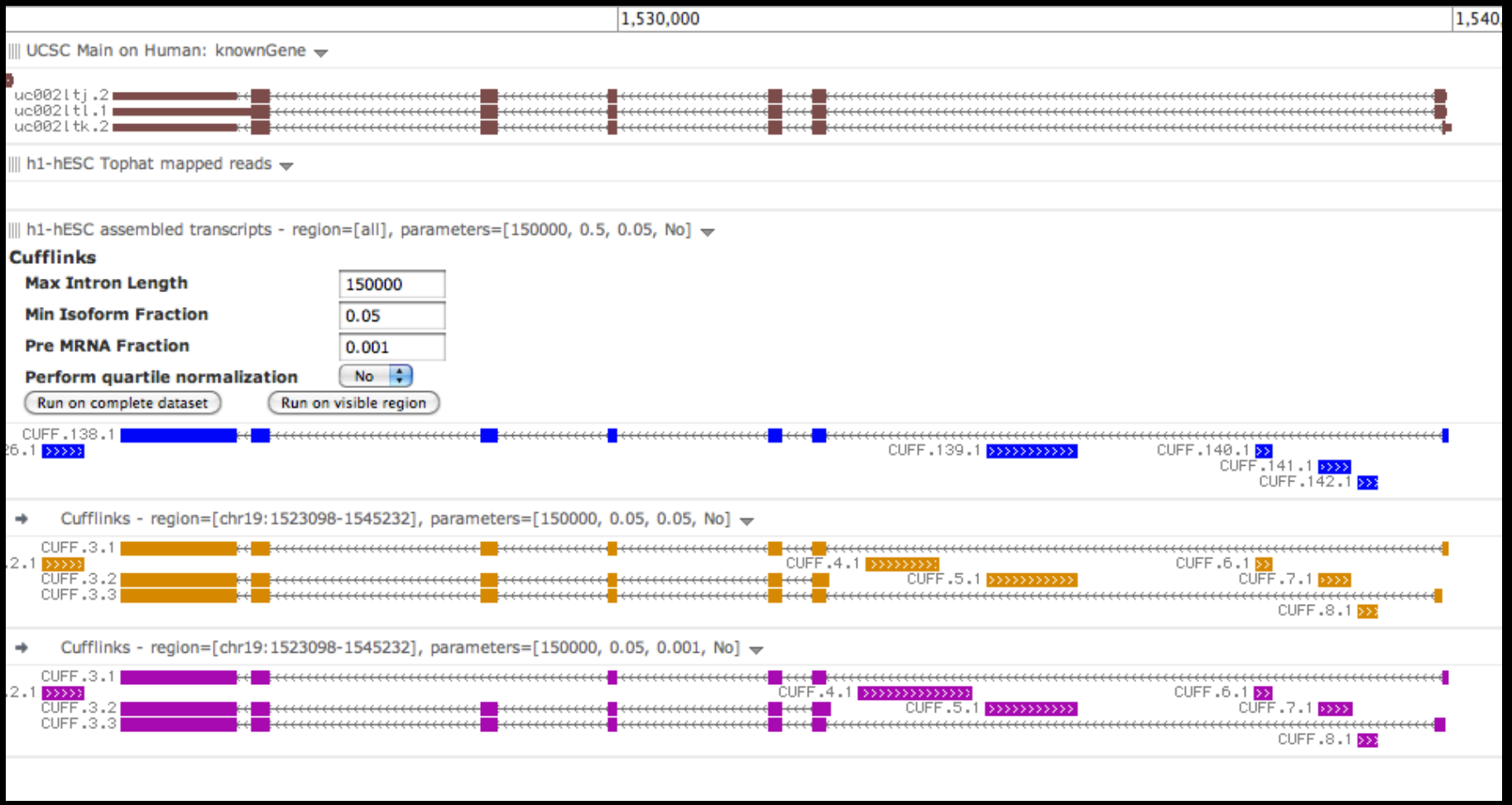
Extract transcript sequences and profile sequences for function



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.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
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.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
CUFF.8.1
Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.001, No]
CUFF.3.1
CUFF.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
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

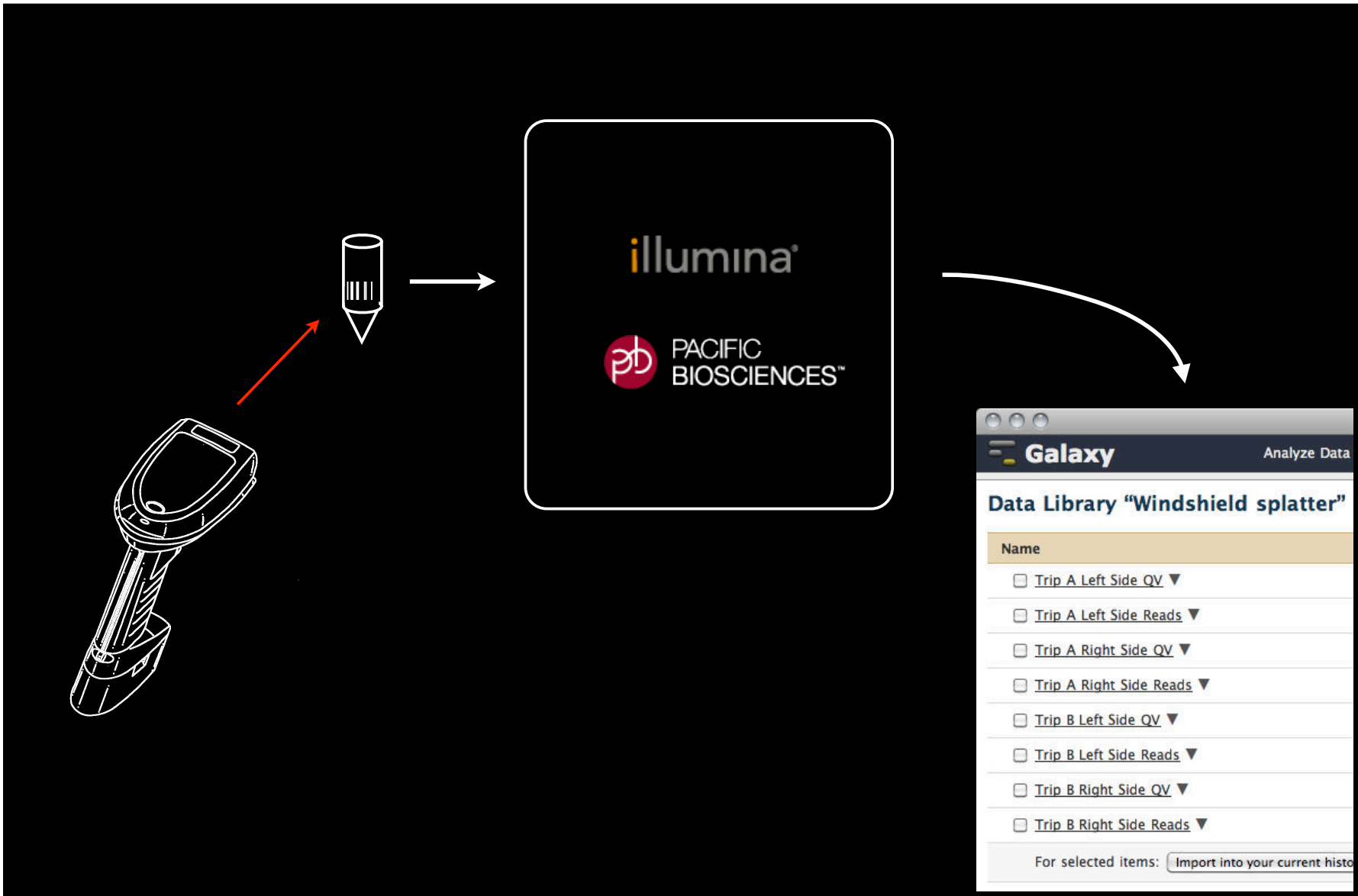
High-throughput Sequencing (HTS) Data

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

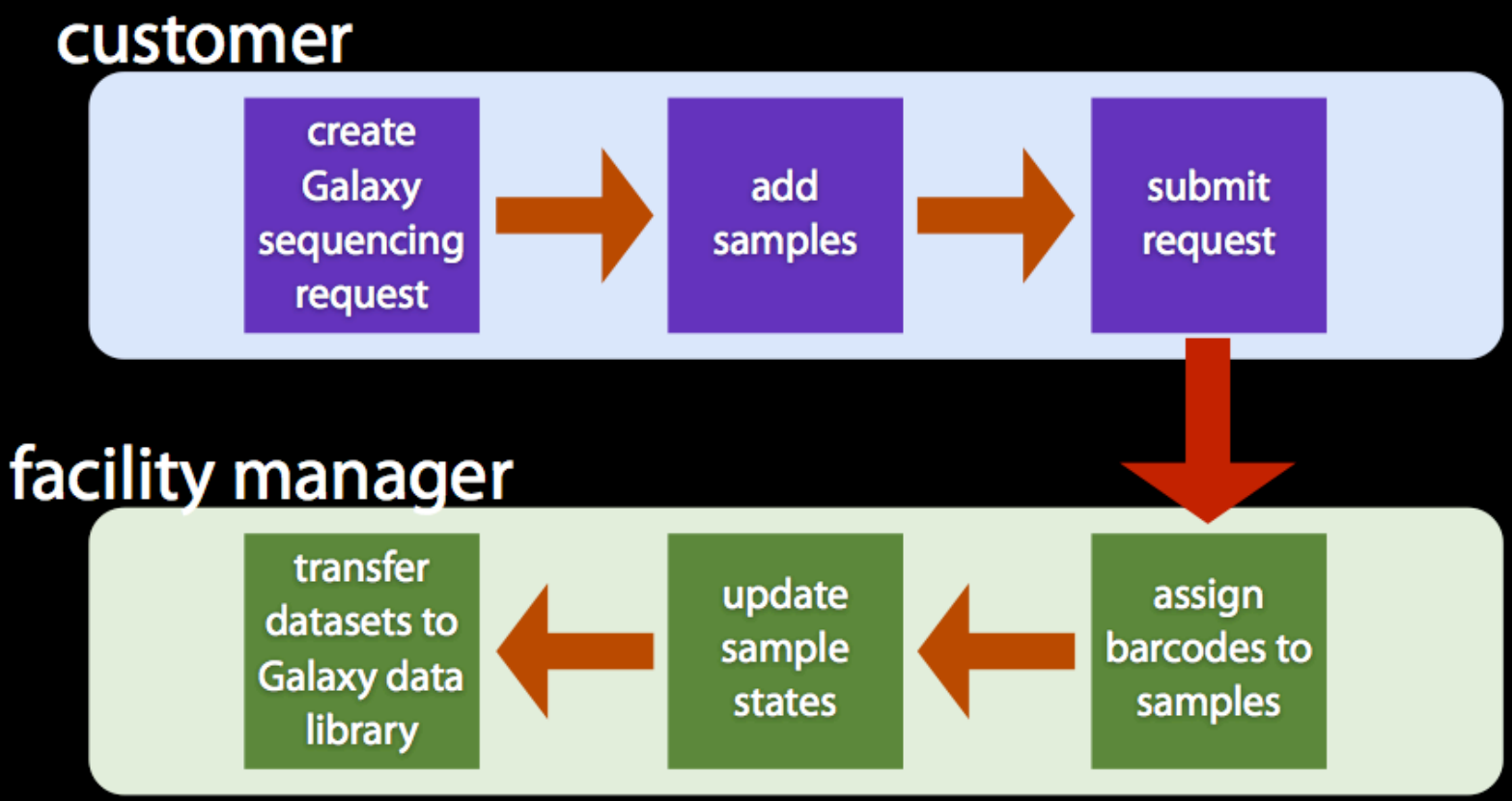
Built-in system for tracking sequencing requests

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

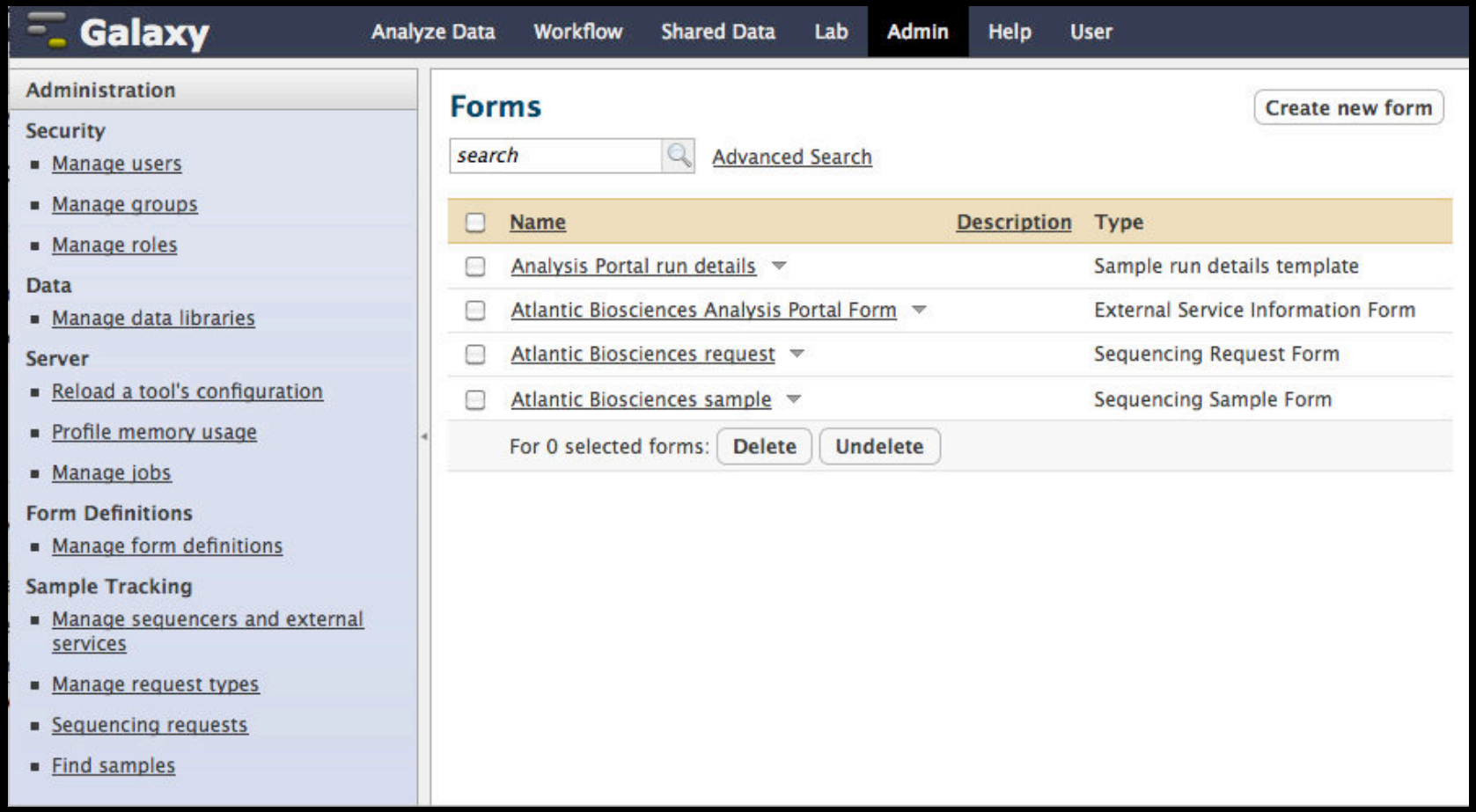
Create and submit a sequencing request

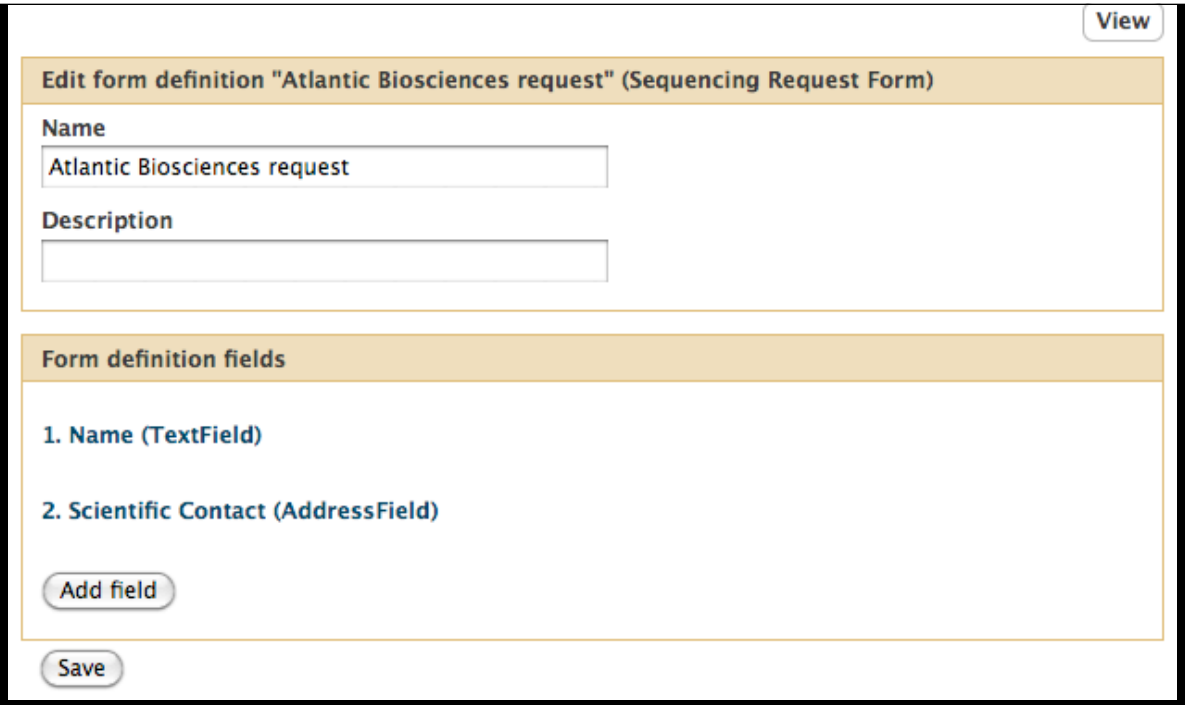
Edit and resubmit a rejected sequencing request

Obtain datasets at the end of a sequencing run

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

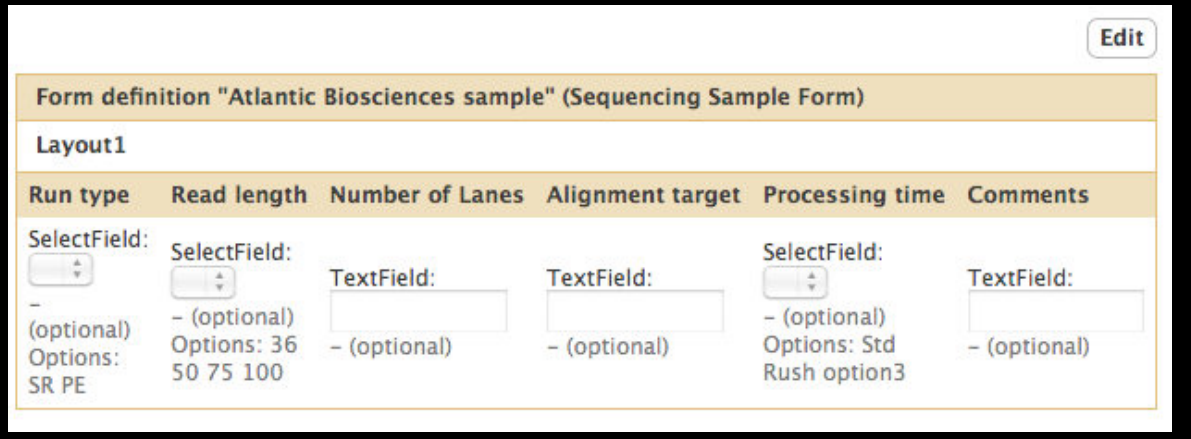
Configure Available Request / Sample Options



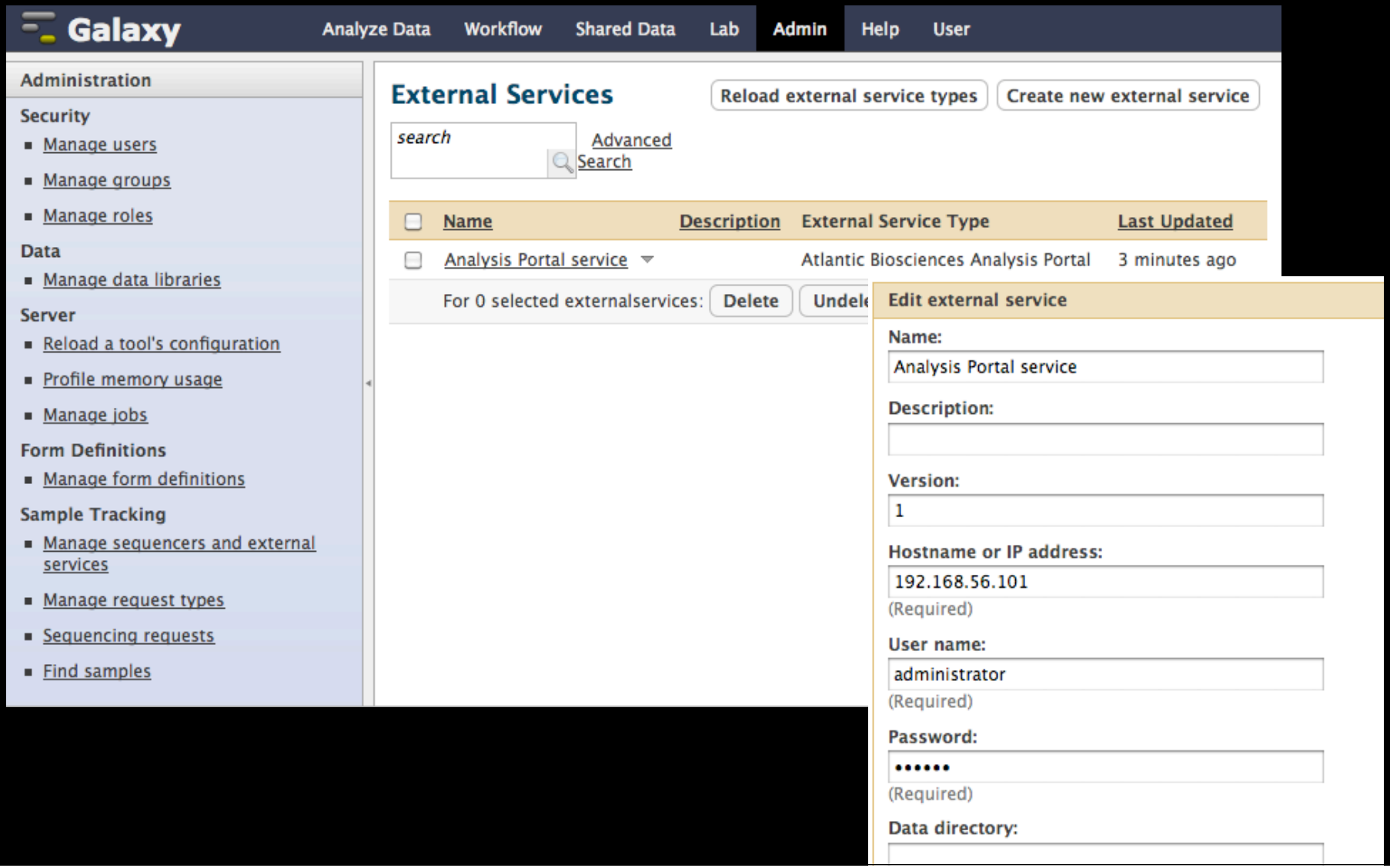


Configurations can be

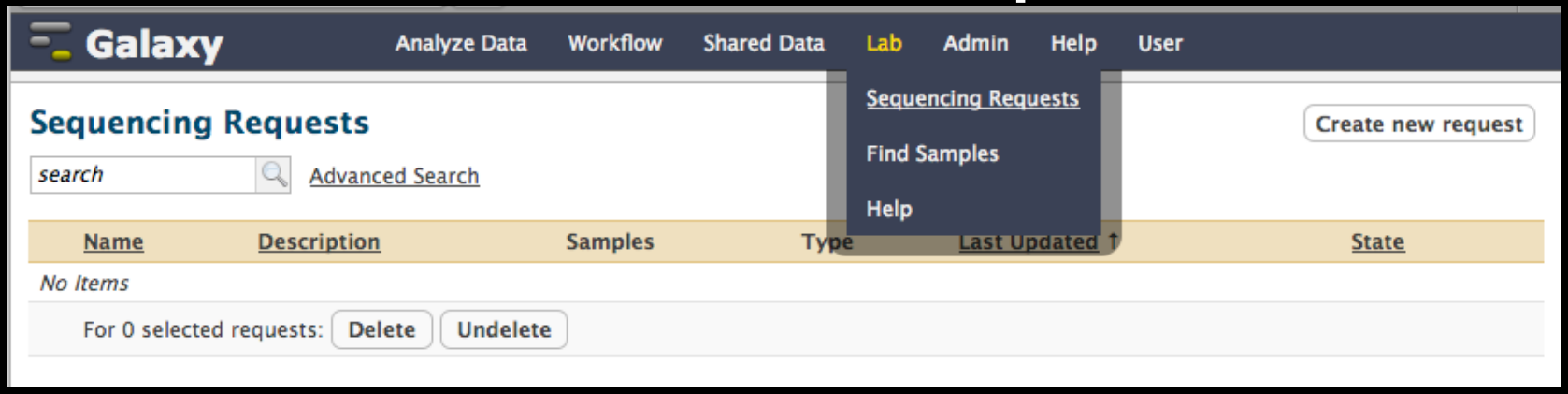
- custom-built
- loaded from provided configuration files

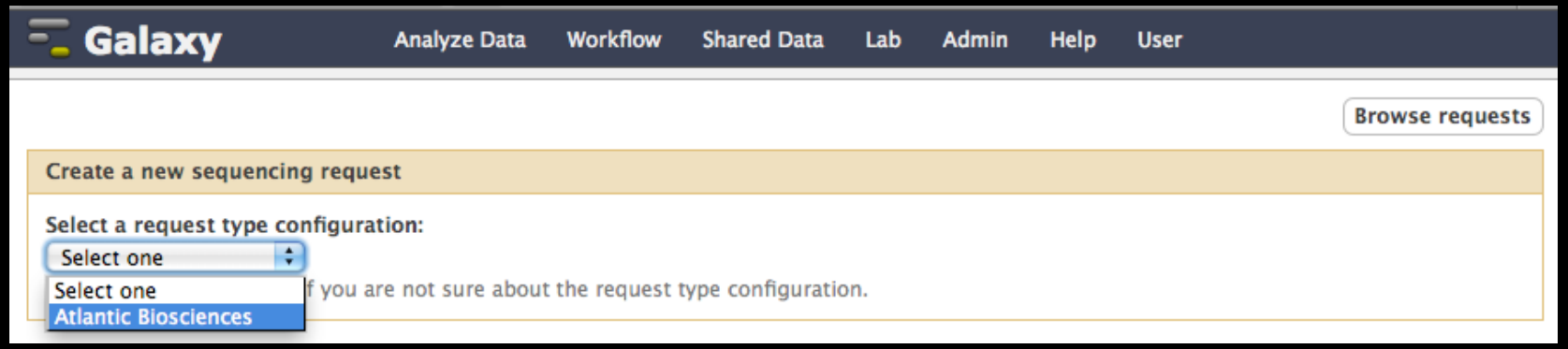


Configure the Sequencer

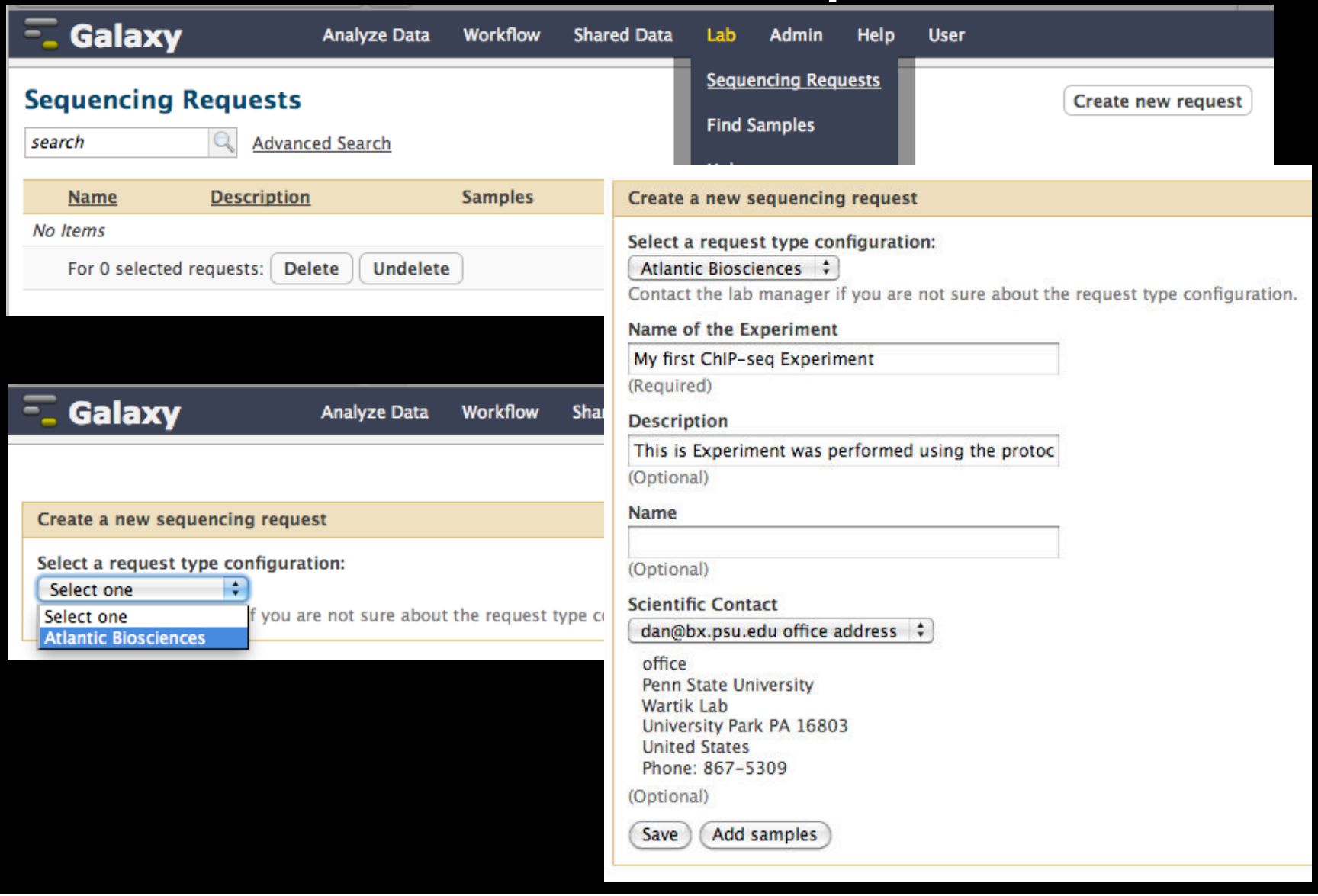


User Creates a Request

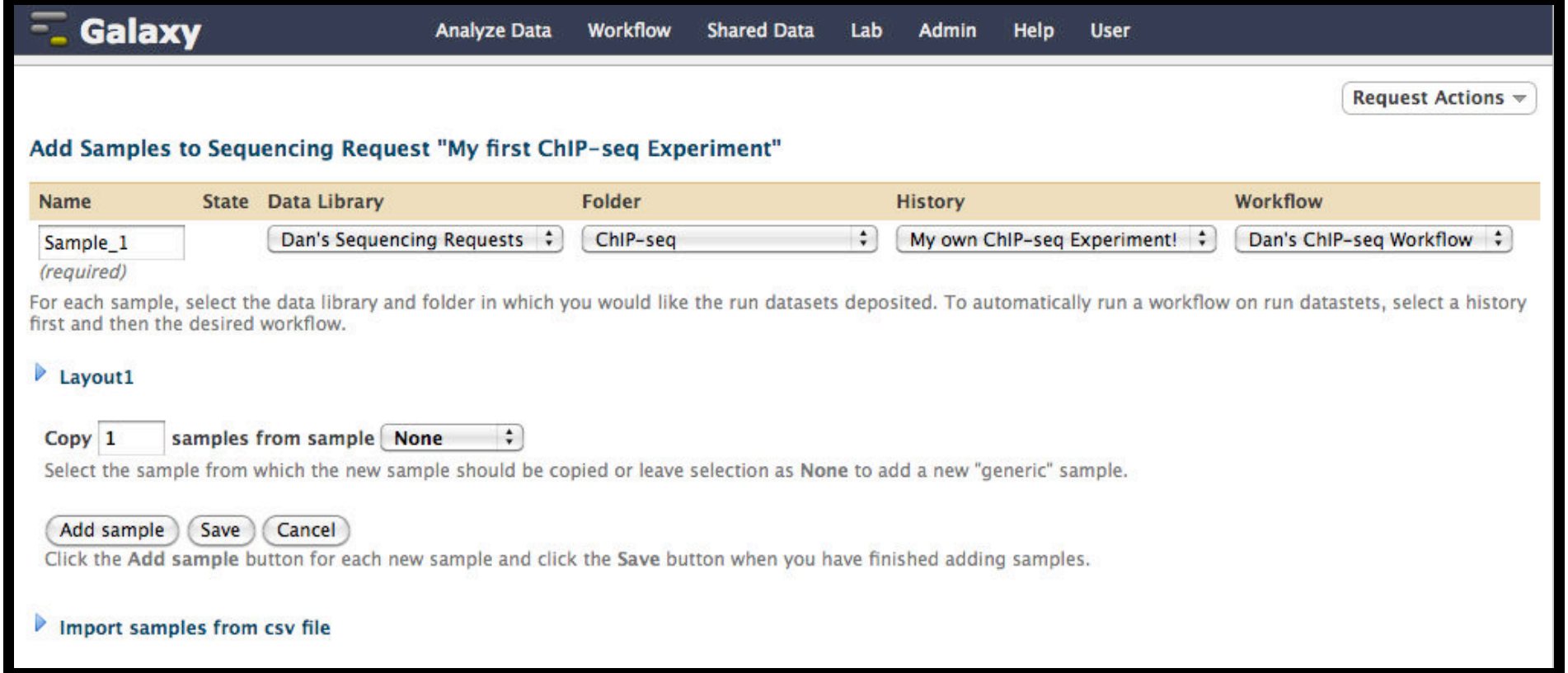




User Creates a Request

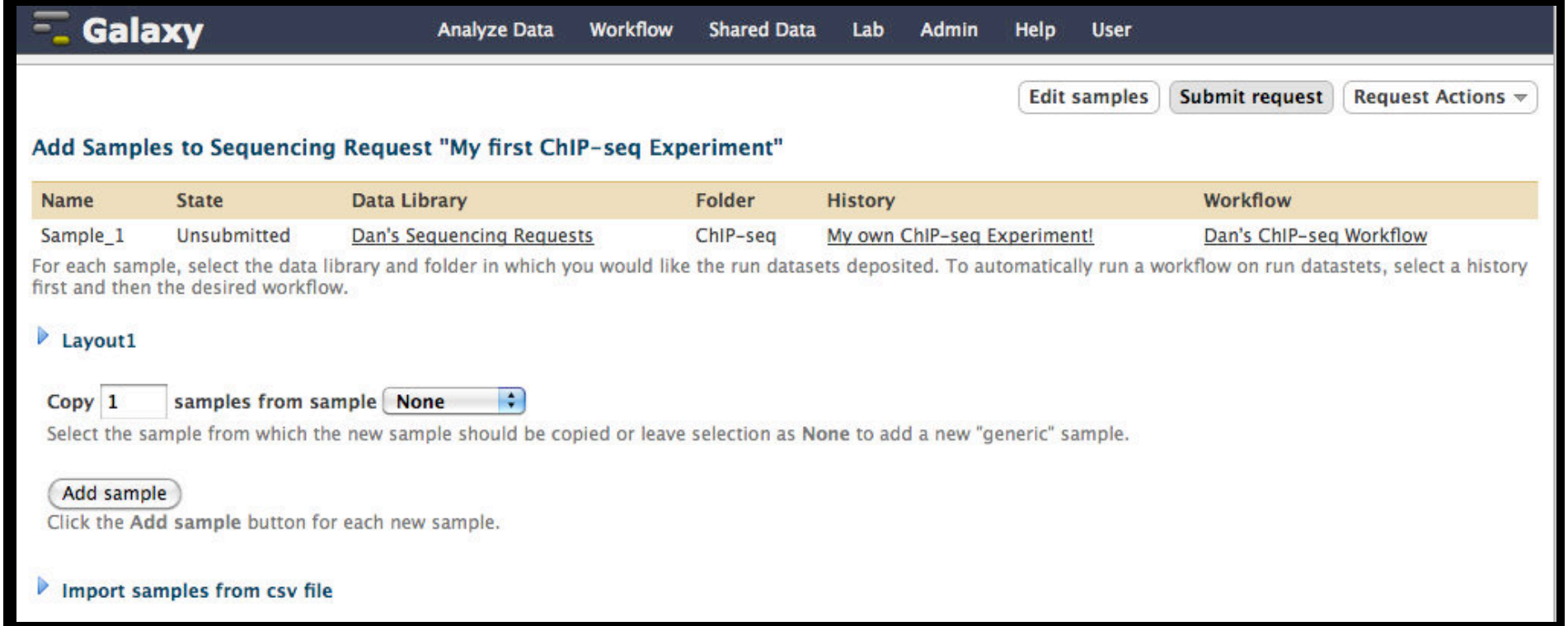


User Adds a Sample

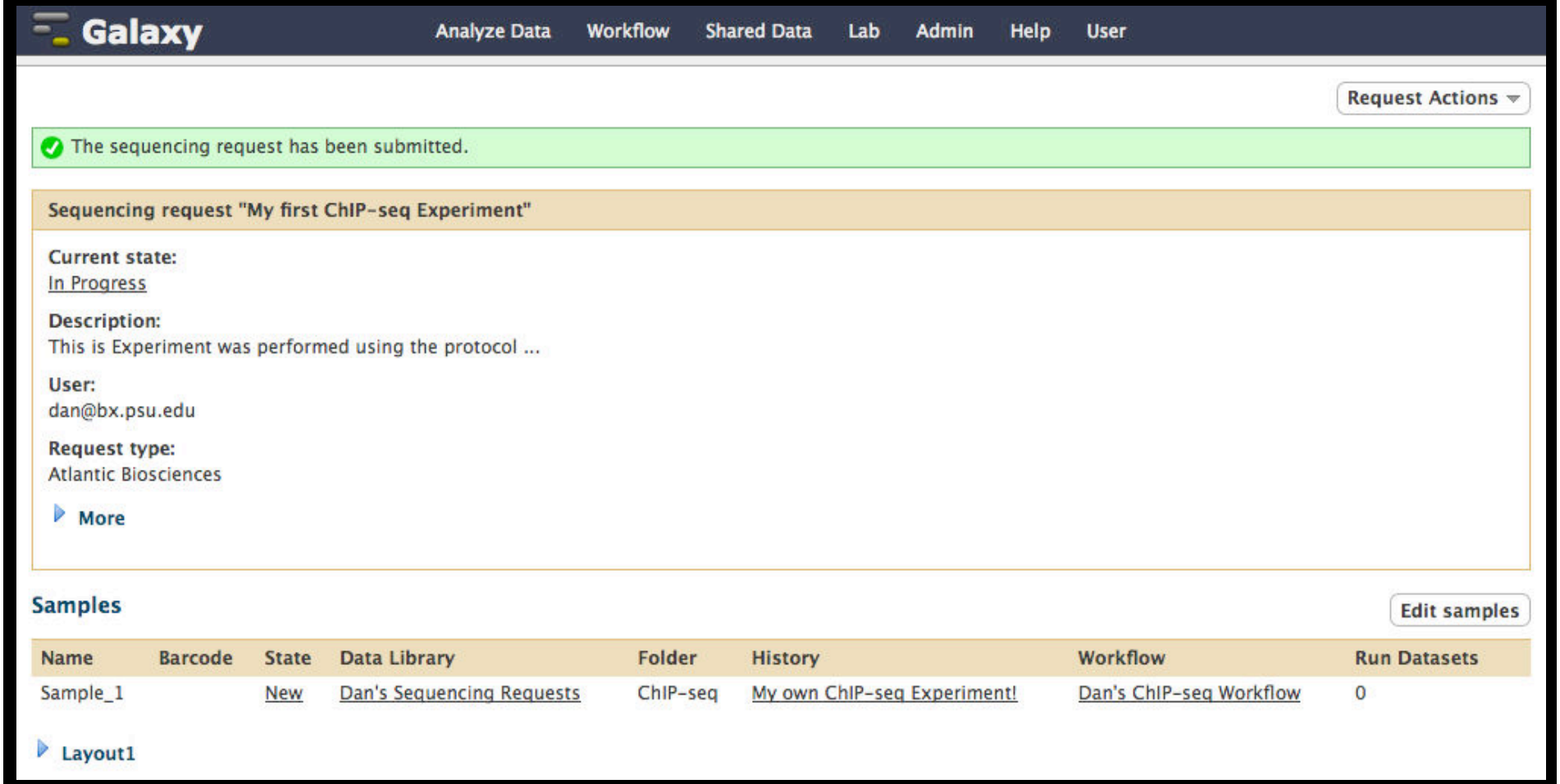


History
Workflow to run

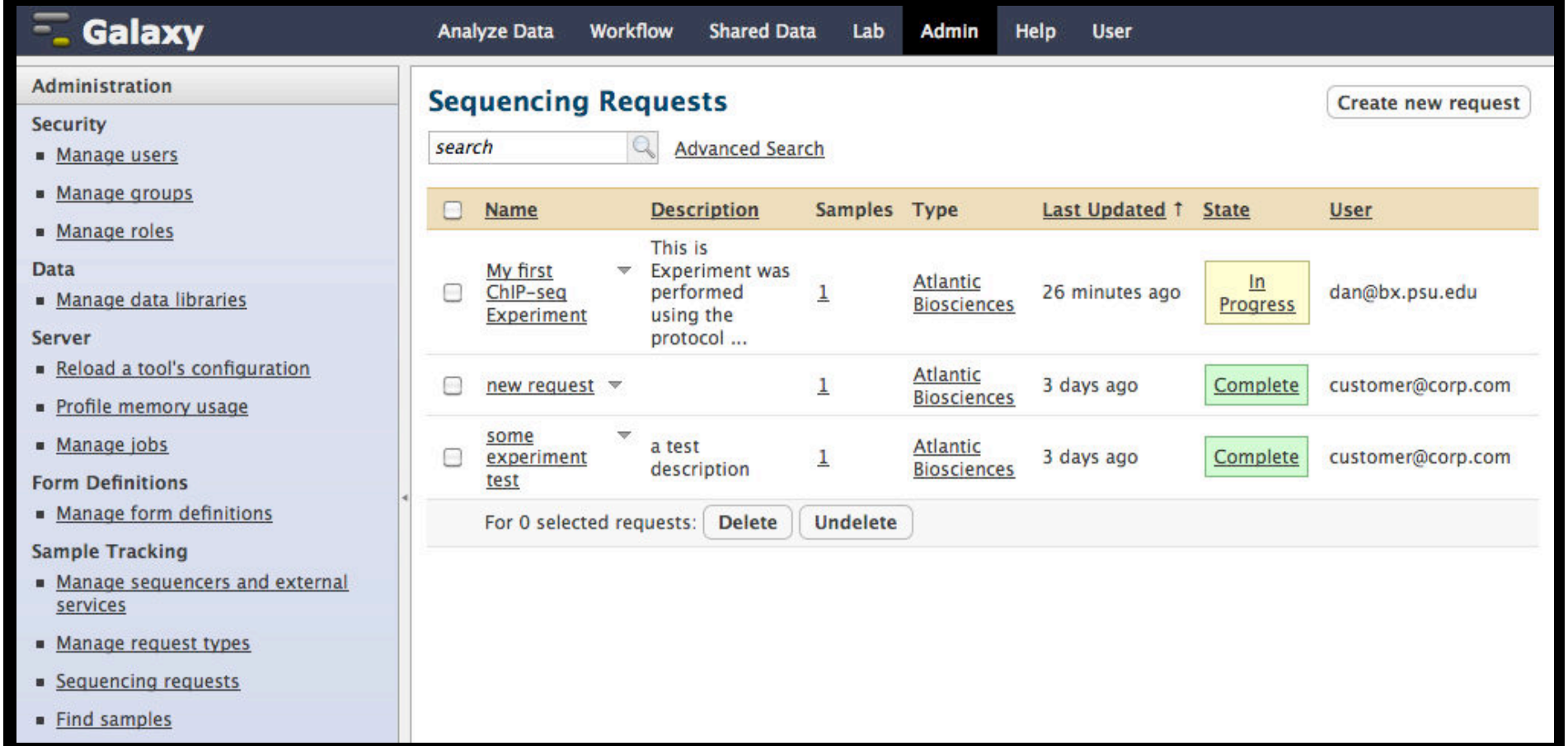
Samples Added, Submit Request



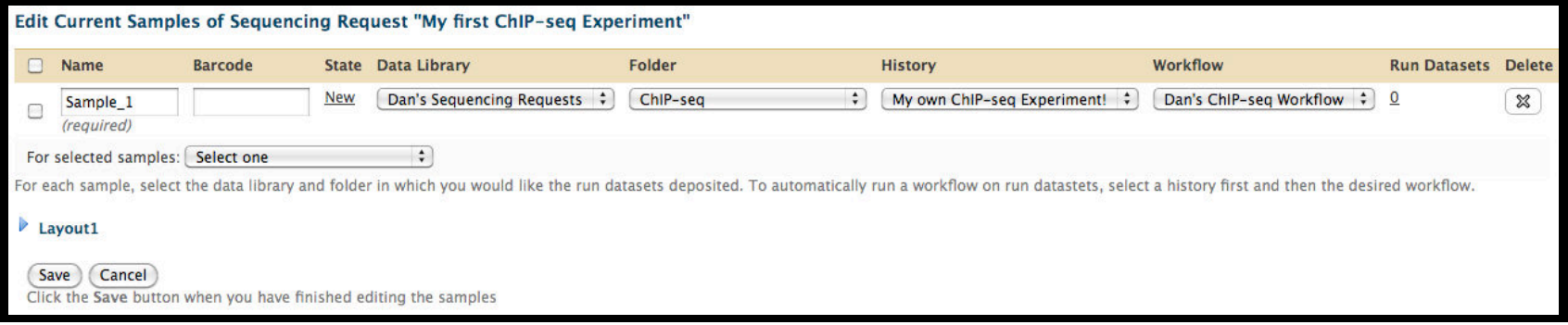
Samples enter "New" state

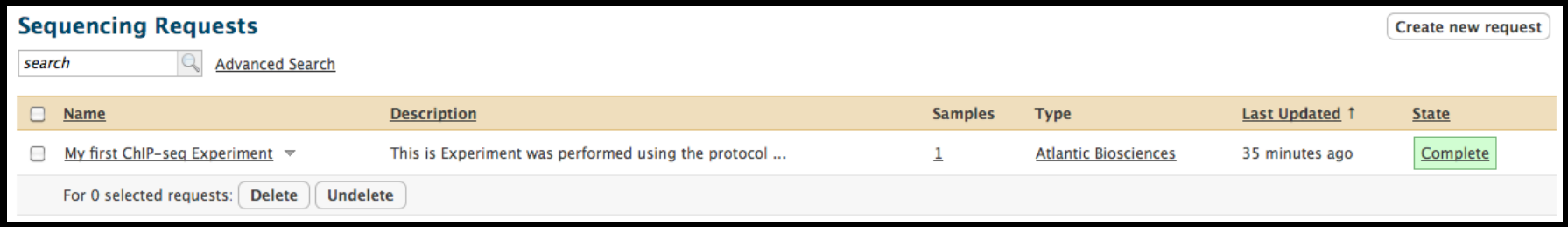


Sequencing Facility is informed of Request



Sequencing Facility Receives Samples





Facility

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

User can watch progress of sequencing request

Sequencing Finished

Datasets are transferred from sequencer into Galaxy

- library
- user's history

Galaxy Workflow is executed on Dataset

User is automatically emailed

Extending Sample Tracking with ngLims

An add-on written by community contributor Brad Chapman

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

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

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

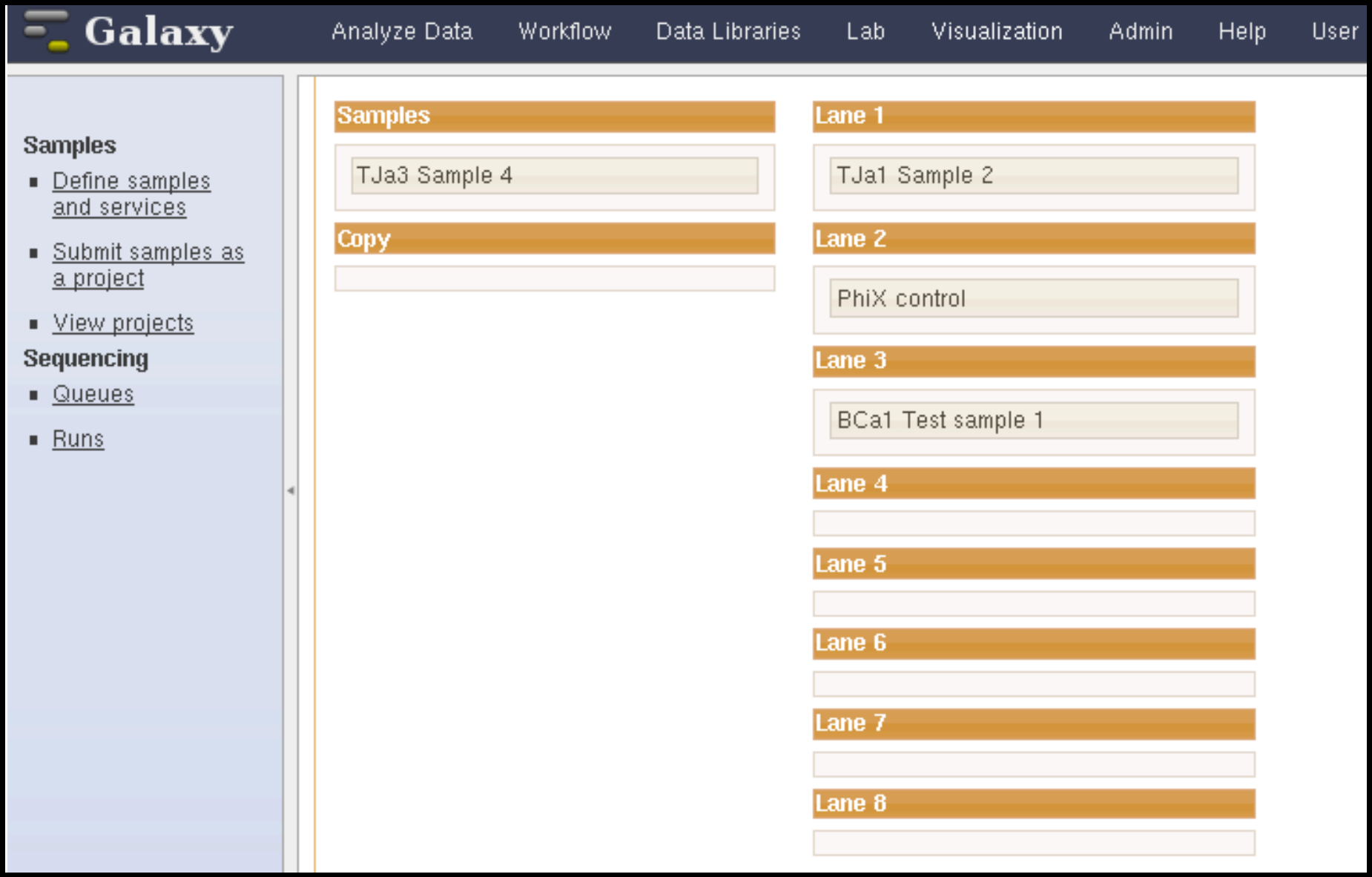
Sample tracking is completely extensible

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

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

Interaction with sequence instruments and secondary analysis is completely pluggable

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



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

EMORY

Enis Afgan

Dannon Baker

Dave Clements

Jeremy Goecks

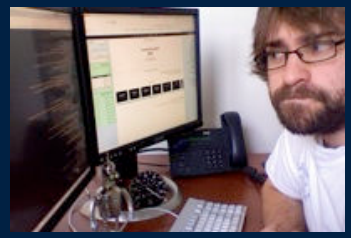
Kanwei Li

James Taylor

PENNSTATE

1855

Dan Blankenberg



Nate Coraor

Jennifer Jackson

Greg von Kuster

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-analysis-exercise

- 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