An Introduction to Galaxy

Daniel Blankenberg
The Galaxy Team
http://UseGalaxy.org

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

The Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and transparent computational biomedical research

What is Galaxy?

GUI for genomics

- for complete analyses: analyze, visualize, share, publish

A free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

Open source software that makes integrating your own tools and data and customizing for your own site simple

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

Galaxy Analysis Workspace

Galaxy

http://main.g2.bx.psu.edu/

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

Tools Options

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- EMBOSS

NGS TOOLBOX BETA

- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Indel Analysis
- NGS: Peak Calling

RGENETICS

- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models

Workflows

Map with Bowtie for Illumina

Will you select a reference genome from your history or use a built-in index?: Use a built-in index

Built-ins were indexed using default options

Select a reference genome: mm9

If your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?: Paired-end

Forward FASTQ file: 1: E18 PE.1 Reads

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file: 1: E18 PE.1 Reads

Must have Sanger-scaled quality values with ASCII offset 33

Maximum insert size for valid paired-end alignments (-X): 1000

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff): FR (for Illumina)

Bowtie settings to use: Commonly used

For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Suppress the header in the output SAM file: ☑

Bowtie produces SAM with several lines of header information by default

Execute

What it does

Bowtie is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.

History Options

Imported: SNP Pileup Analysis for Sample E18

- 15: Variants from sample E18, consensus different, in RefSeq Genes
- 14: UCSC mm9 RefSeq Genes
- 13: Variants from sample E18 where consensus base different than ref. base
- 10: Variants from sample E18
- 9: Generate pileup on data 8
- 8: SAM-to-BAM on data 7
- 7: Map with Bowtie for Illumina on data 6 and data 5
- 6: E18 PE.2 Reads Groomed, Trimmed
- 5: E18 PE.1 Reads Groomed, Trimmed
- 4: E18 PE.2 Reads Groomed
- 3: E18 PE.1 Reads Groomed
- 2: E18 PE.2 Reads
- 1: E18 PE.1 Reads

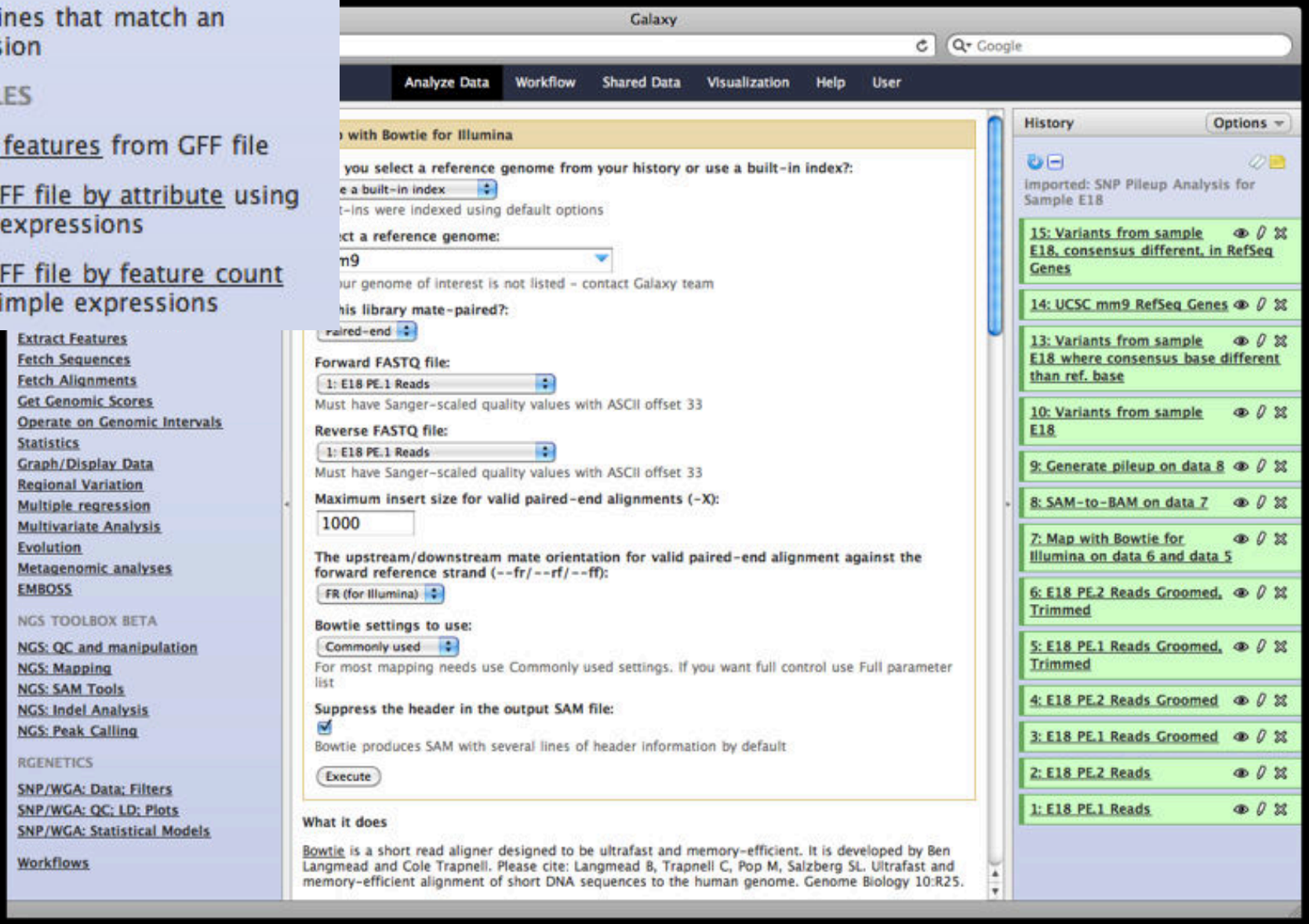
xy Analysis Workspace

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression

GFF FILES

- Extract features from GFF file
- Filter GFF file by attribute using simple expressions
- Filter GFF file by feature count using simple expressions



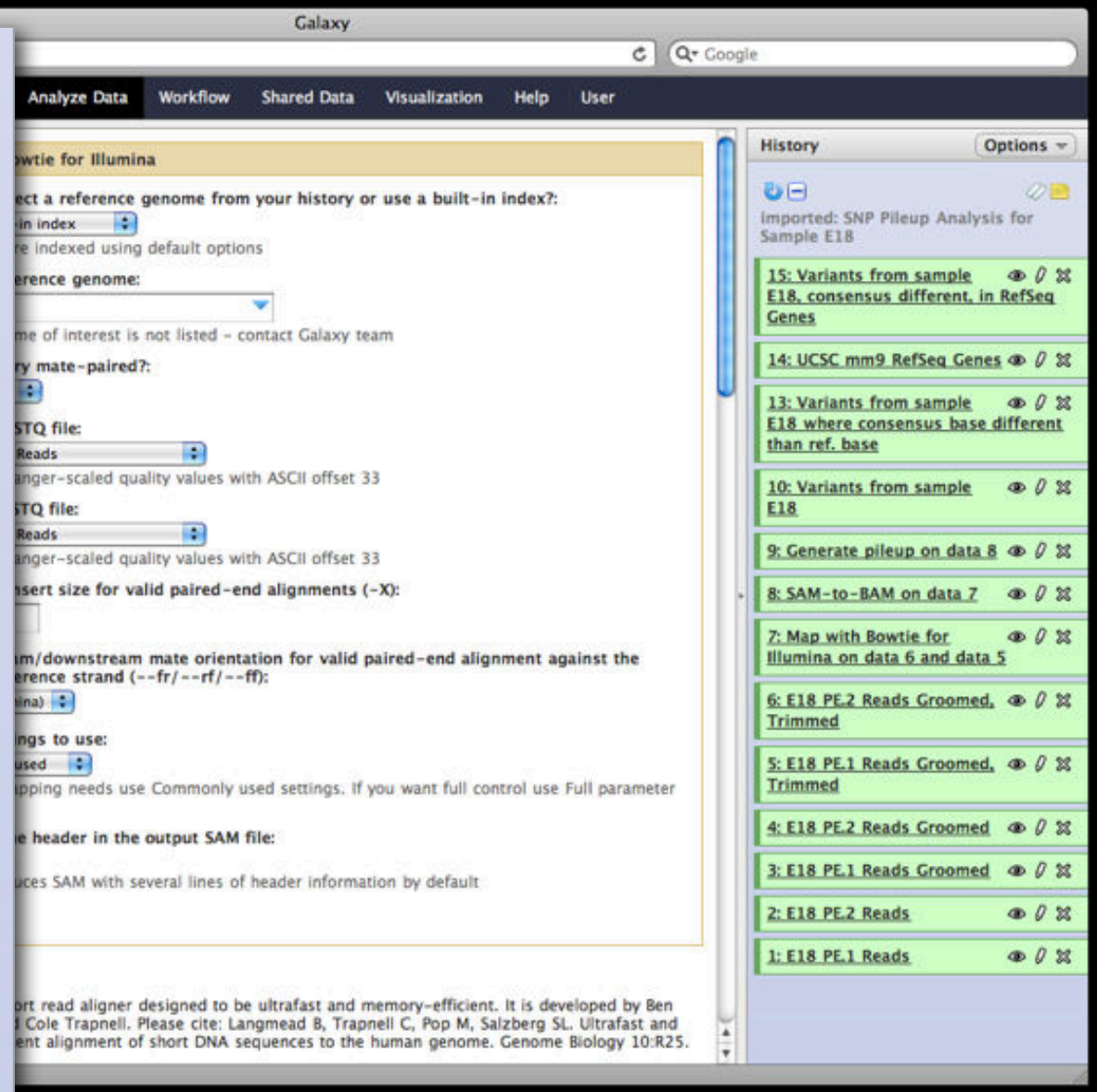
xy Analysis Workspace

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an e
- E
- Fi si
- Fi u

Operate on Genomic Intervals

- Intersect the intervals of two queries
- Subtract the intervals of two queries
- Merge the overlapping intervals of a query
- Concatenate two queries into one query
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a query
- Cluster the intervals of a query
- Join the intervals of two queries side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest feature for every interval
- Profile Annotations for a set of genomic intervals



xy Analysis Workspace

Filter and Sort

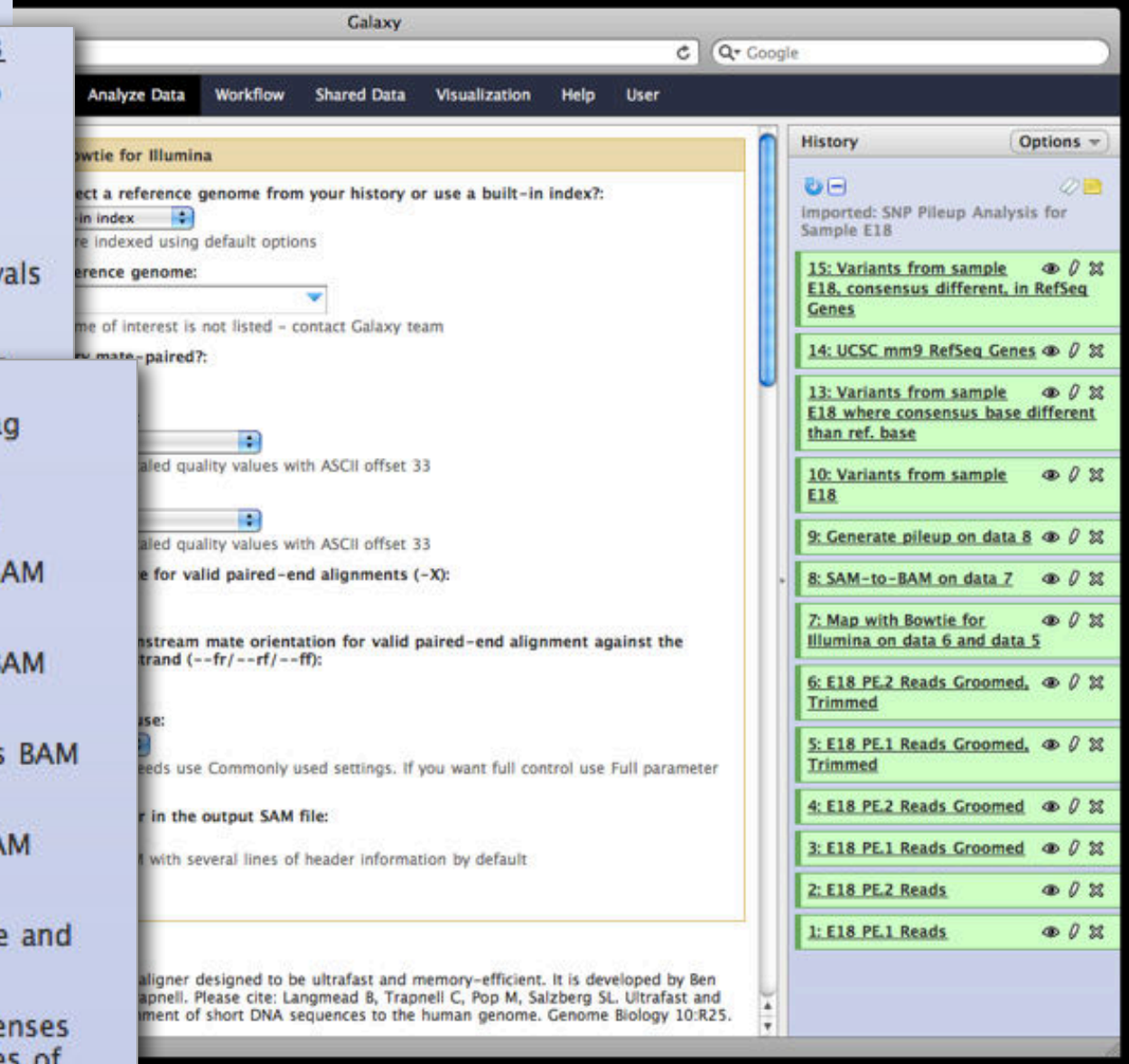
- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an e
- E
- Fi si
- Fi u

Operate on Genomic Intervals

- Intersect the intervals of two queries
- Subtract the intervals of two queries
- Merge the overlapping intervals of a query

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



Filter and Sort

- Filter data on any colum using simple expressions
- Sort data in ascending descending order
- Select lines that match e

Operate on Genom

- Intersect the inte queries
- Subtract the inte queries
- Merge the overla of a query

NGS: SAM Too

- Filter SAM o values
- Convert SAM
- SAM-to-BAM format to BA
- BAM-to-SAM format to SA
- Merge BAM files togethe
- Generate pil dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases

Filter pileup

Select dataset:

10: Variants from sample E18

which contains:

Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

3

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

No

See "Output format" below for explanation

Print total number of differences?:

No

See "Example 3" below for explanation

Print quality and base string?:

Yes

See "Example 4" below for explanation

Execute

aligner designed to be ultrafast and memory-efficient. It is developed by Ben apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and ment of short DNA sequences to the human genome. Genome Biology 10:R25.

Options

1: E18 PE.1 Reads

Filter and Sort

- Filter data on any colum using simple expressions
- Sort data in ascending descending order
- Select lines that match

Operate on Genom

- Intersect the inte queries
- Subtract the inte queries
- Merge the overla of a query

NGS: SAM Too

- Filter SAM o values
- Convert SAM
- SAM-to-BAM format to BA
- BAM-to-SAM format to SA
- Merge BAM files togethe
- Generate pil dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases

Filter pileup

Select dataset:
10: Variants from sample E18

which contains:
Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:
20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:
3

Pileup lines with coverage lower than this value will be skipped

Only report variants?:
Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:
No

See "Output format" below for explanation

Print total number of differences?:
No

See "Example 3" below for explanation

Print quality and base string?:
Yes

See "Example 4" below for explanation

Execute

aligner designed to be ultrafast and memory-efficient. It is develo
apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. U
ment of short DNA sequences to the human genome. Genome Bio

History Options

Variant Analysis for Sample E18

- 15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
- 14: UCSC mm9 RefSeq Genes
- 13: Filter to get Variants from sample E18 where consensus base different than ref. base
- 10: Filter pileup to get Variants from sample E18
- 9: Generate pileup on data 8
- 8: SAM-to-BAM on data 7
- 7: Map with Bowtie for Illumina on data 6 and data 5
- 6: E18 PE.2 Reads Groomed, Trimmed
- 5: E18 PE.1 Reads Groomed, Trimmed

Filter and Sort

- Filter data on any co simple expressions
- Sort data in ascendir descending order
- Select lines that mat

Operate on Geno

- Intersect the in queries
- Subtract the int queries
- Merge the over of a query

NGS: SAM To

- Filter SAM values
- Convert SA
- SAM-to-BA format to
- BAM-to-SA format to
- Merge BAM files togeth
- Generate p dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases

This dataset is large and only the first megabyte is shown below.
Show all | Save

chr10	6882036	6882037	A	A	107	0	60	32	.$..,.....C.....
chr10	14243075		14243076		G	G	96	0	60	35	t.
chr10	14243079		14243080		C	C	106	0	60	35	..
chr10	14465082		14465083		T	K	173	176	60	35	GG
chr10	14465083		14465084		G	K	144	144	60	35	..
chr10	14465084		14465085		T	T	117	0	60	38	..
chr10	14465085		14465086		G	G	70	0	60	38	..
chr10	14465257		14465258		C	C	79	0	60	42	..
chr10	14465258		14465259		A	A	137	0	60	46	..
chr10	14465263		14465264		A	A	136	0	60	61	..
chr10	14465366		14465367		A	A	101	0	60	38	g$
chr10	14465371		14465372		G	G	137	0	60	50	.$
chr10	14465410		14465411		G	G	184	0	60	69	.$
chr10	14465447		14465448		T	T	186	0	60	65	.$
chr10	14465456		14465457		G	G	193	0	60	70	..
chr10	14465465		14465466		T	T	177	0	60	63	.$
chr10	14465485		14465486		C	T	129	129	60	34	t$
chr10	14465569		14465570		T	T	219	0	60	84	.
chr10	14465581		14465582		G	G	240	0	60	84	,$
chr10	14465586		14465587		C	C	248	0	60	82	.$
chr10	14465621		14465622		C	C	134	0	60	49	..
chr10	14465658		14465659		C	C	134	0	60	49	,,
chr10	14465660		14465661		T	T	153	0	60	55	,
chr10	14465691		14465692		G	G	128	0	60	42	.$
chr10	14465778		14465779		C	C	89	0	60	34	,$
chr10	14465791		14465792		G	G	104	0	60	33	,$
chr10	14465881		14465882		G	G	110	0	60	41	..
chr10	17445088		17445089		A	A	103	0	60	34	..
chr10	17445271		17445272		A	A	55	0	60	34	..
chr10	17731269		17731270		T	T	113	0	60	42	,$
chr10	19928287		19928288		G	A	135	135	60	36	AA
chr10	19928468		19928469		C	T	132	132	60	35	T$
chr10	19928488		19928489		A	A	119	0	60	44	..
chr10	19928494		19928495		C	T	138	138	60	37	TT
chr10	19928527		19928528		A	A	134	0	60	45	,
chr10	19928538		19928539		G	G	144	0	60	52	,$
chr10	19928543		19928544		A	G	147	147	60	40	Gg
chr10	19928741		19928742		T	T	80	0	60	30	,
chr10	20799826		20799827		G	G	117	0	60	37	,$
chr10	28750217		28750218		C	T	138	138	60	37	TT
chr10	28750397		28750398		A	C	154	211	60	64	C$
chr10	28750401		28750402		A	A	128	0	60	47	,$
chr10	28750423		28750424		C	T	113	113	60	35	T$
chr10	28750438		28750439		A	A	95	0	60	36	.$
chr10	28750446		28750447		A	G	165	165	60	46	G$
chr10	28750487		28750488		A	A	80	0	60	31	,
chr10	28750512		28750513		G	G	220	0	60	72	.$
chr10	28750548		28750549		G	C	255	255	60	97	C$
chr10	28750574		28750575		T	T	237	0	60	83	.$
chr10	28750577		28750578		T	T	234	0	60	82	,$
chr10	28750578		28750579		T	T	242	0	60	76	,$
chr10	28750593		28750594		G	G	220	0	60	75	.$
chr10	28750640		28750641		T	C	165	165	60	46	Cc
chr10	28750746		28750747		G	A	202	202	60	58	AA
chr10	28750766		28750767		A	G	205	205	60	59	G$
chr10	28750769		28750770		T	C	175	175	60	49	cc

aligner designed to be ultrafast and memory-efficient. It is develo
apnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. U
ment of short DNA sequences to the human genome. Genome Bio

Analysis

Options

nalysis for Sample E18

sect to get Variants
nple E18, consensus different,
Genes

mm9 RefSeq Genes

to get Variants from
18 where consensus base
than ref. base

pileup to get
from sample E18

ate pileup on data 8

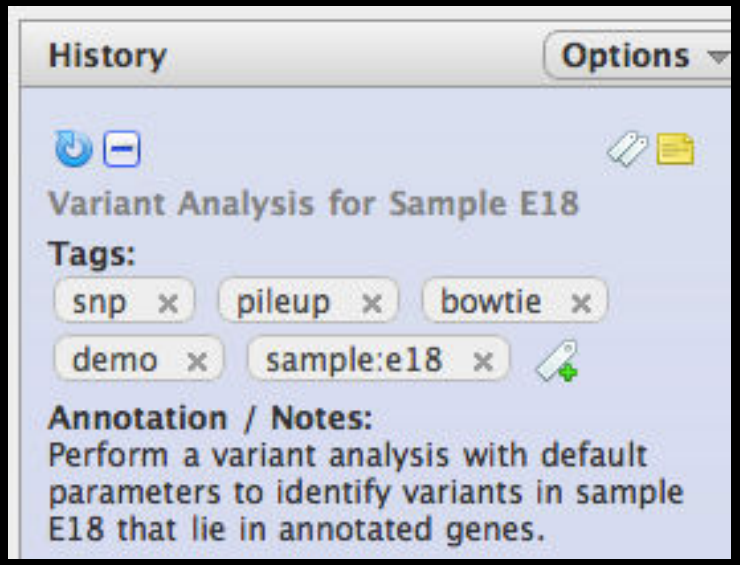
o-BAM on data 7

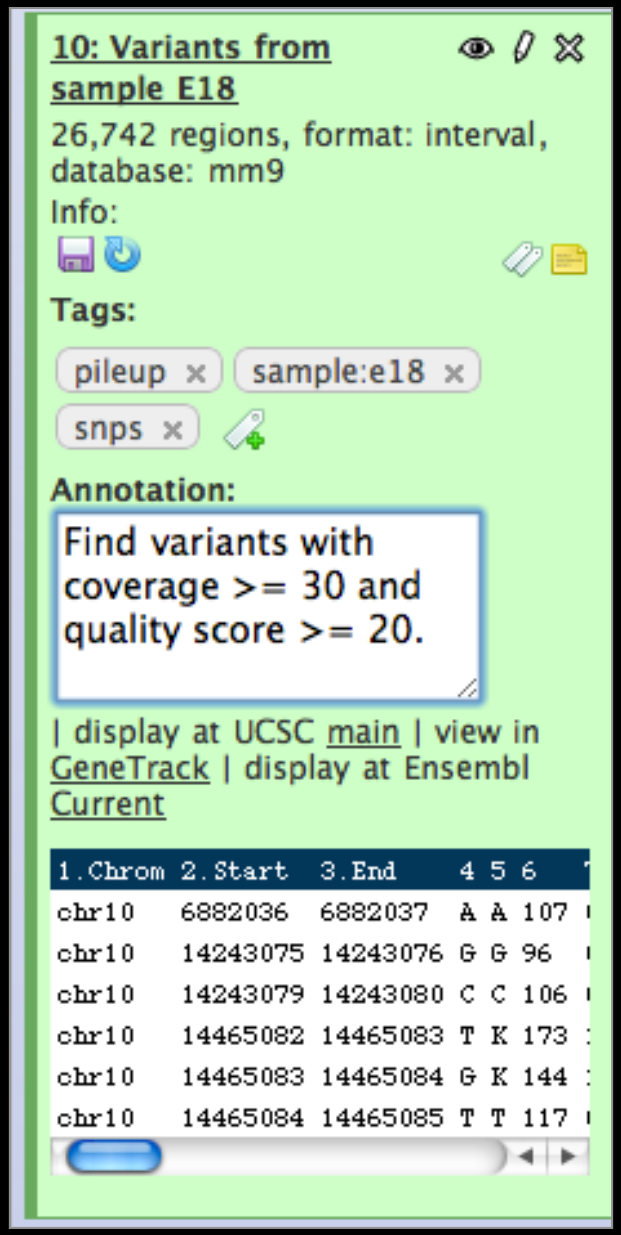
ith Bowtie for
on data 6 and data 5

6: E18 PE.2 Reads Groomed, Trimmed

5: E18 PE.1 Reads Groomed, Trimmed

User Metadata





Datasources

Upload file from your computer

- FTP support for large datasets

UCSC table browser

BioMart

interMine / modMine

EuPathDB server

EncodeDB at NHGRI

EpiGRAPH server

Tool Suites

Text Manipulation
Format Converters
Filtering and Sorting
Join, Subtract, Group
Sequence Tools
Multi-species Alignment Tools
Genomic Interval Operations
Summary Statistics
Graphing / Plotting

Regional Variation
EMBOSS
Evolution / Phylogeny
RNA-seq
ChIP-seq
GATK
Picard
RGenetics
...and more

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 MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window

Evolution

Metagenomic analyses

Human Genome Variation

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

ILLUMINA

- Map with Bowtie for Illumina
- Map with BWA for Illumina

ROCHE-454

- Lastz map short reads against reference sequence
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

- Map with Bowtie for SOLiD

NGS: SAM Tools

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Workflows

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

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

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Workflows

NGS: SAM Tools

NGS: Indel Analysis

- Filter Indels for SAM
- Extract indels from SAM
- Indel Analysis

NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

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

Dozens of tools for different HTS applications packaged with Galaxy

VCF Tools

- Intersect Generate the intersection of two VCF files
- Annotate a VCF file (dbSNP, hapmap)
- Filter a VCF file
- Extract reads from a specified region

NGS: Picard (beta)

QC/METRICS FOR SAM/BAM

- BAM Index Statistics
- Sam/bam Alignment Summary Metrics
- Sam/bam GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data
- Sam/bam Hybrid Selection Metrics For (eg exome) targeted data

BAM/SAM CLEANING

- Add or Replace Groups
- Reorder SAM
- Replace Sam Header
- Paired Read Mate Fixer for paired data
- Mark Duplicate reads

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

NGS: GATK Tools Alpha

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

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

Data Library "Bushman"

Library Actions

These are the data underlying the analyses reported in the paper "Complete Khoisan and Bantu genomes from southern Africa" by S. C. Schuster et al., published in the journal Nature, February 18, 2010. Each data set can be downloaded and/or imported into a Galaxy history. Data will be updated as the project progresses.

Name	Information	Uploaded By	Date	File Size
All SNPs in personal genomes	Summary table of SNPs in all individuals	greg@bx.psu.edu	2010-01-28	676.8 Mb
Alu insertions in KB1		greg@bx.psu.edu	2010-02-10	14.9 Kb
Alu insertions in NB1		greg@bx.psu.edu	2010-02-10	6.5 Kb
KB1_microsatellites.txt		greg@bx.psu.edu	2010-02-15	3.5 Mb
NB1_microsatellites.txt		greg@bx.psu.edu	2010-02-15	828.5 Kb
amino acid differences with functional predictions		greg@bx.psu.edu	2010-02-05	1.1 Mb
gene copy number in HPA and other personal genomes		greg@bx.psu.edu	2010-02-15	2.1 Mb
indels in ABT		greg@bx.psu.edu	2010-02-03	105.3 Kb
indels in KB1		greg@bx.psu.edu	2010-02-03	14.2 Mb
indels in MD8		greg@bx.psu.edu	2010-02-03	109.8 Kb
indels in NB1		greg@bx.psu.edu	2010-02-03	[illegible] Kb
indels in TK1		greg@bx.psu.edu	2010-02-03	123.2 Kb
novel SNPs in ABT		greg@bx.psu.edu	2010-02-09	9.4 Mb
novel SNPs in KB1		greg@bx.psu.edu	2010-02-09	16.9 Mb
novel SNPs in MD8		greg@bx.psu.edu	2010-02-09	594.1 Kb
novel SNPs in NB1		greg@bx.psu.edu	2010-02-09	4.1 Mb
novel SNPs in TK1		greg@bx.psu.edu	2010-02-09	722.6 Kb
sequenced exon-containing intervals		greg@bx.psu.edu	2010-02-03	3.1 Mb

For selected items: Import into your current history Go

http://usegalaxy.org/bushman

Managing Libraries

Loading Data

- Upload a single file
- Import datasets from a Galaxy history
- Upload a directory of files
- Directly from Sequencer using Sample Tracking System

Accessing Data

- Data contents on disk are not copied
- Dataset security: public, Role-based access control (RBAC)

Annotating Library Data: Library Templates

- Build user fillable forms
- Associate at Library, Folder or Dataset level

Overview

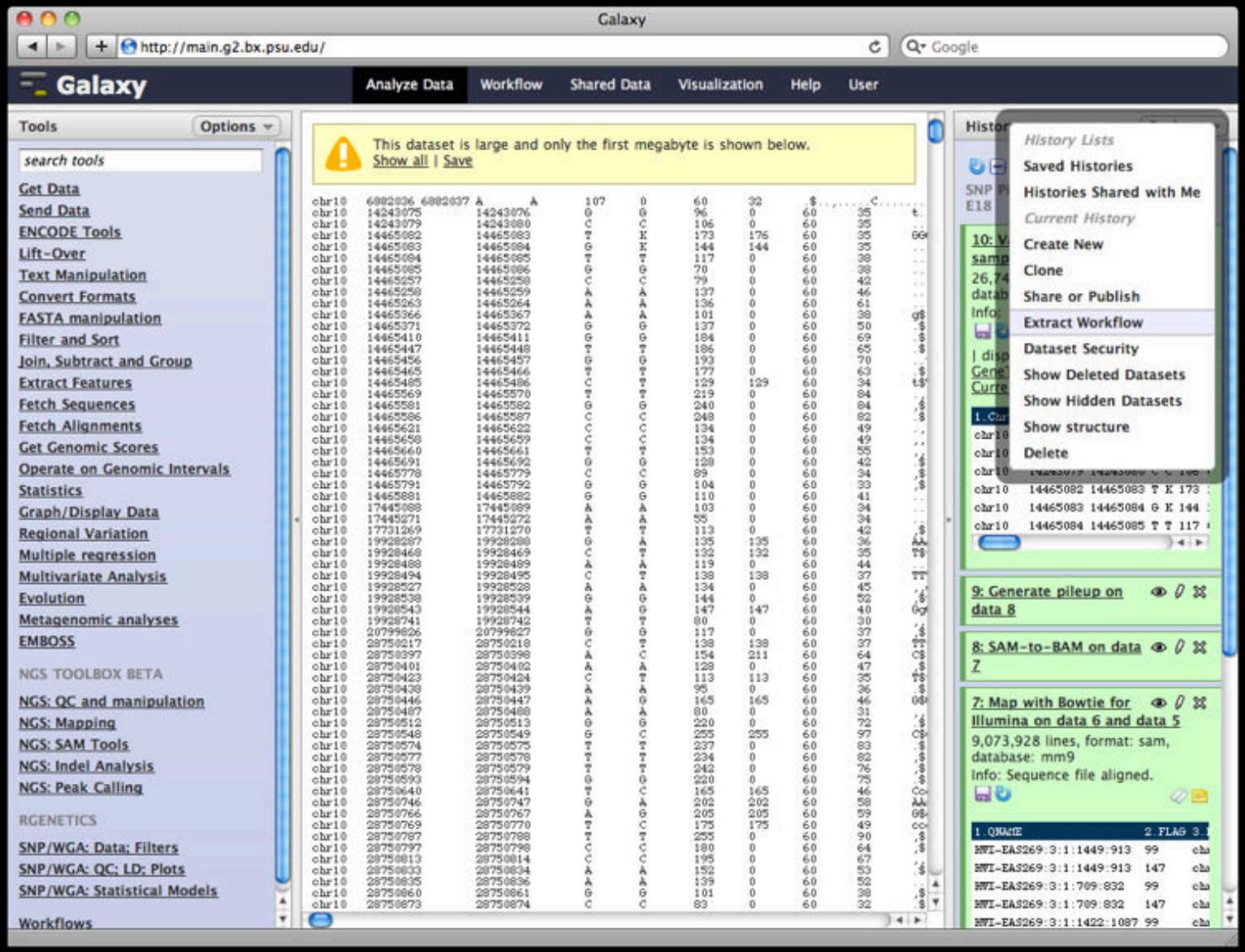
What is Galaxy?

What you can do in Galaxy

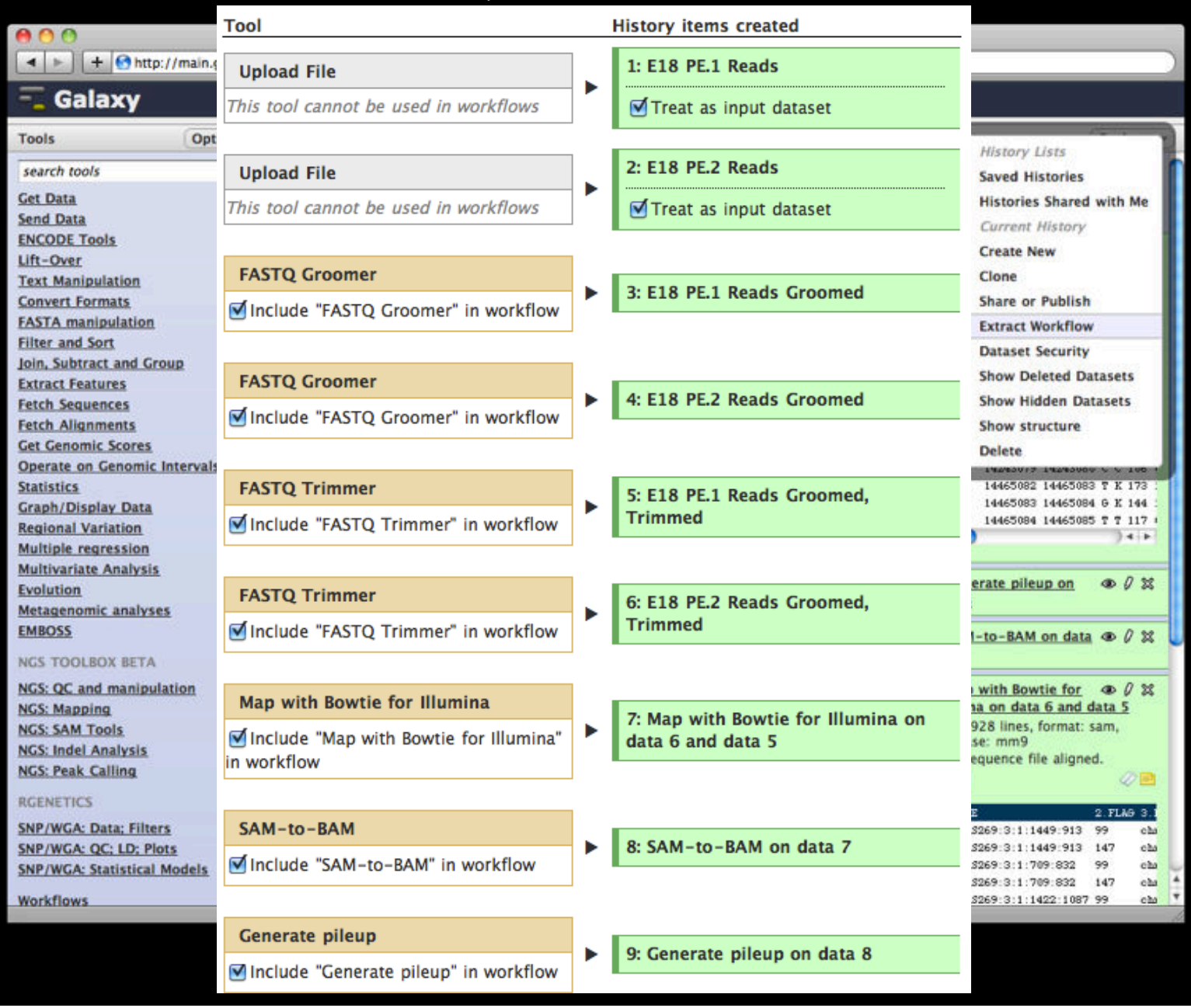
- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

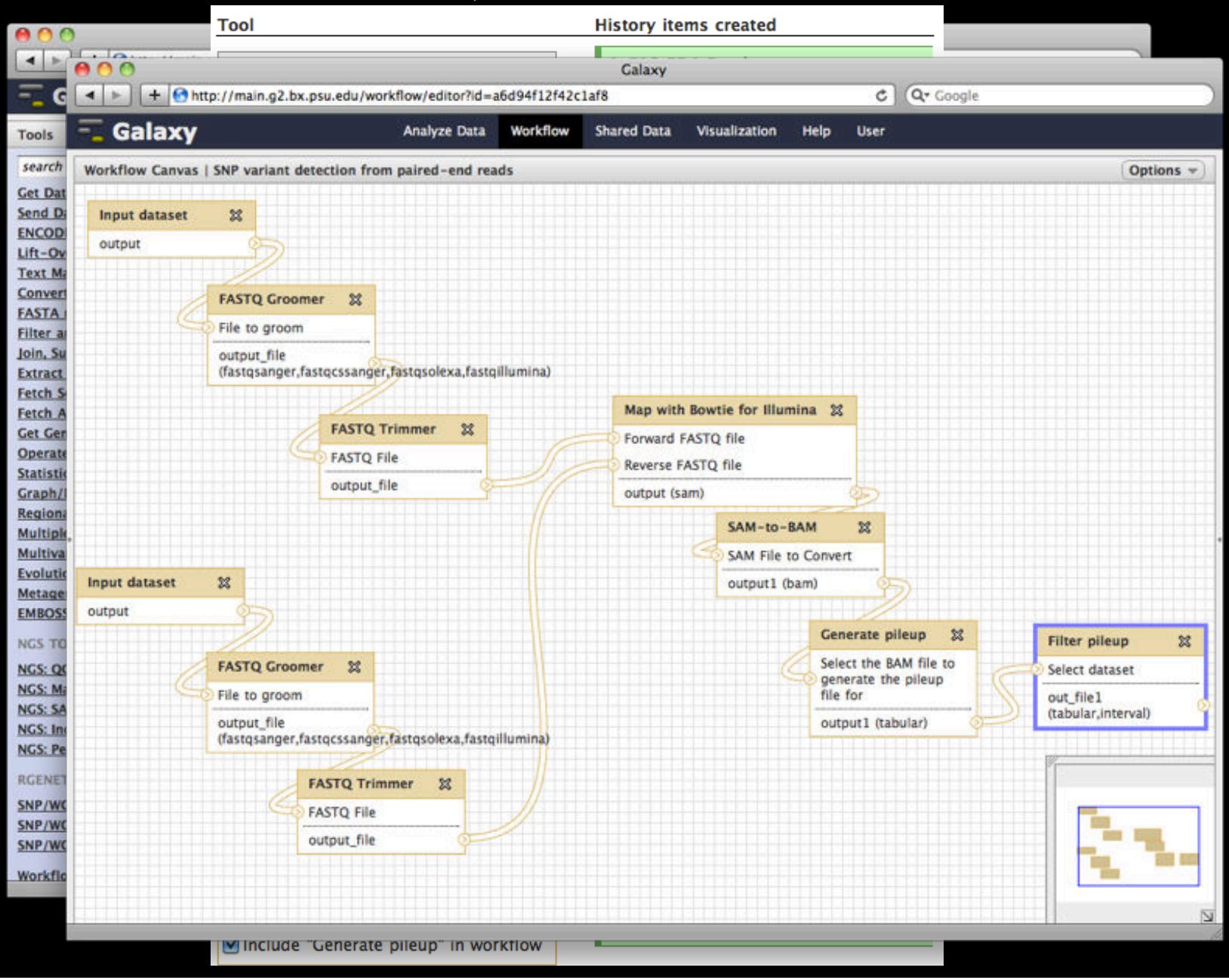
Galaxy Workflows



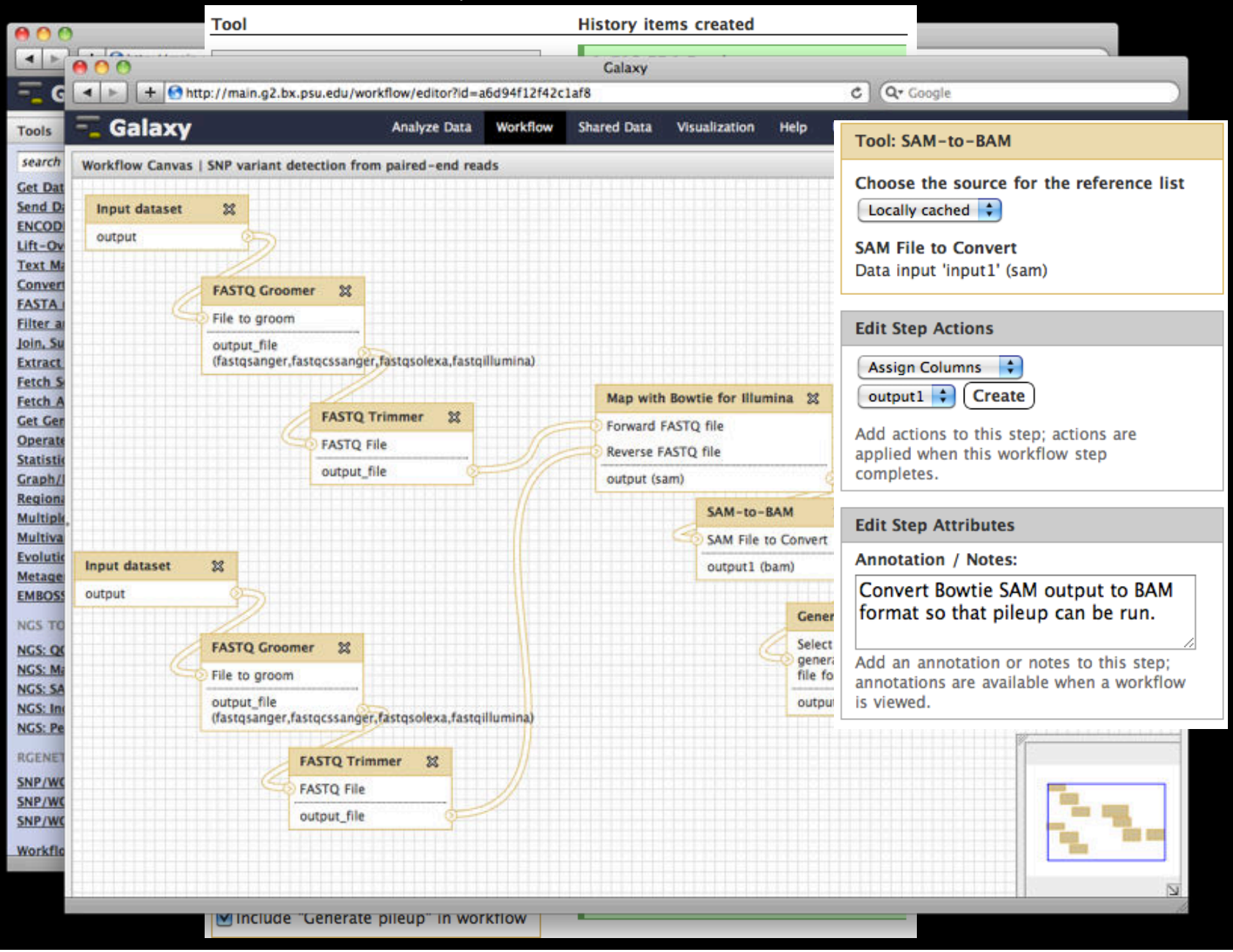
Galaxy Workflows



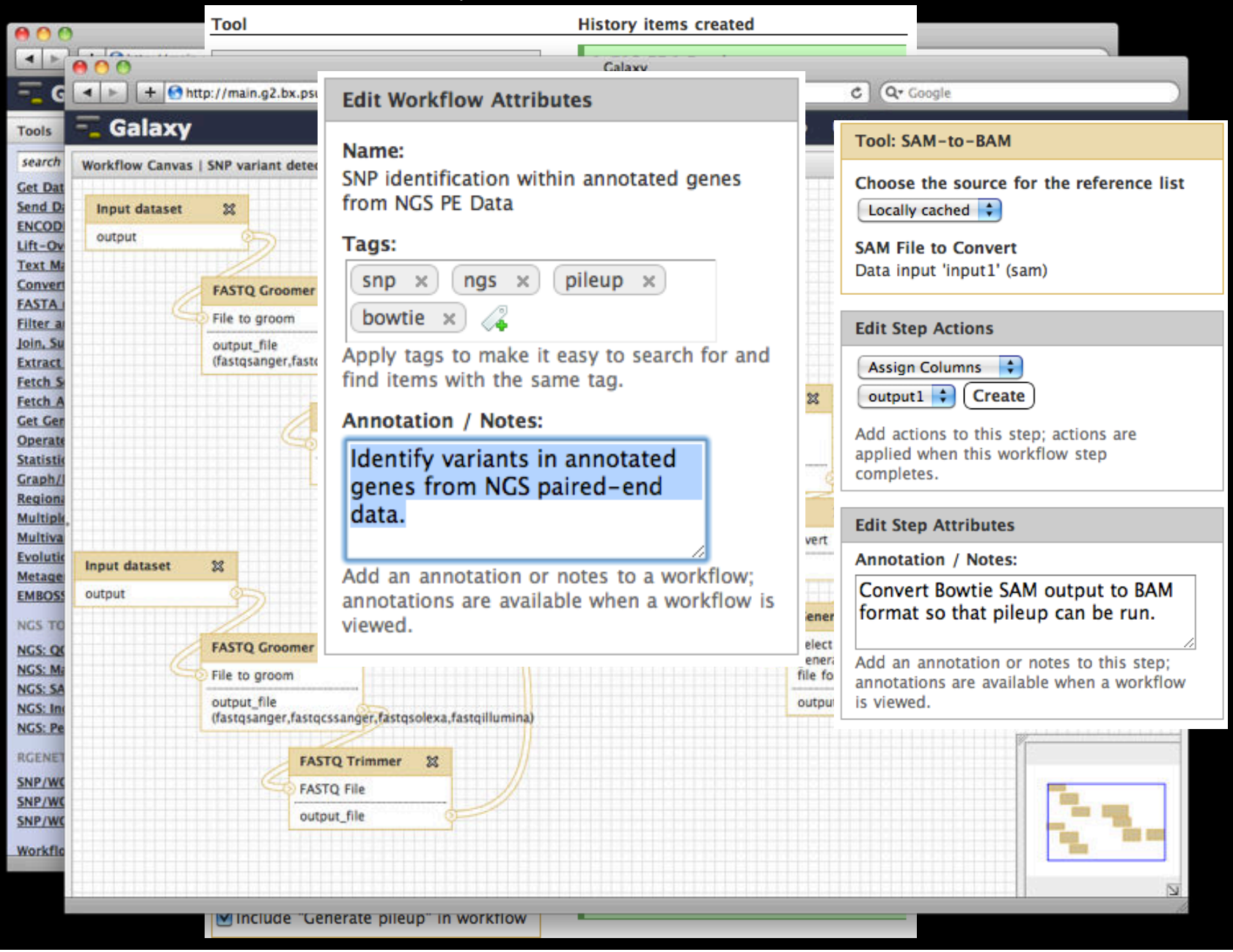
Galaxy Workflows

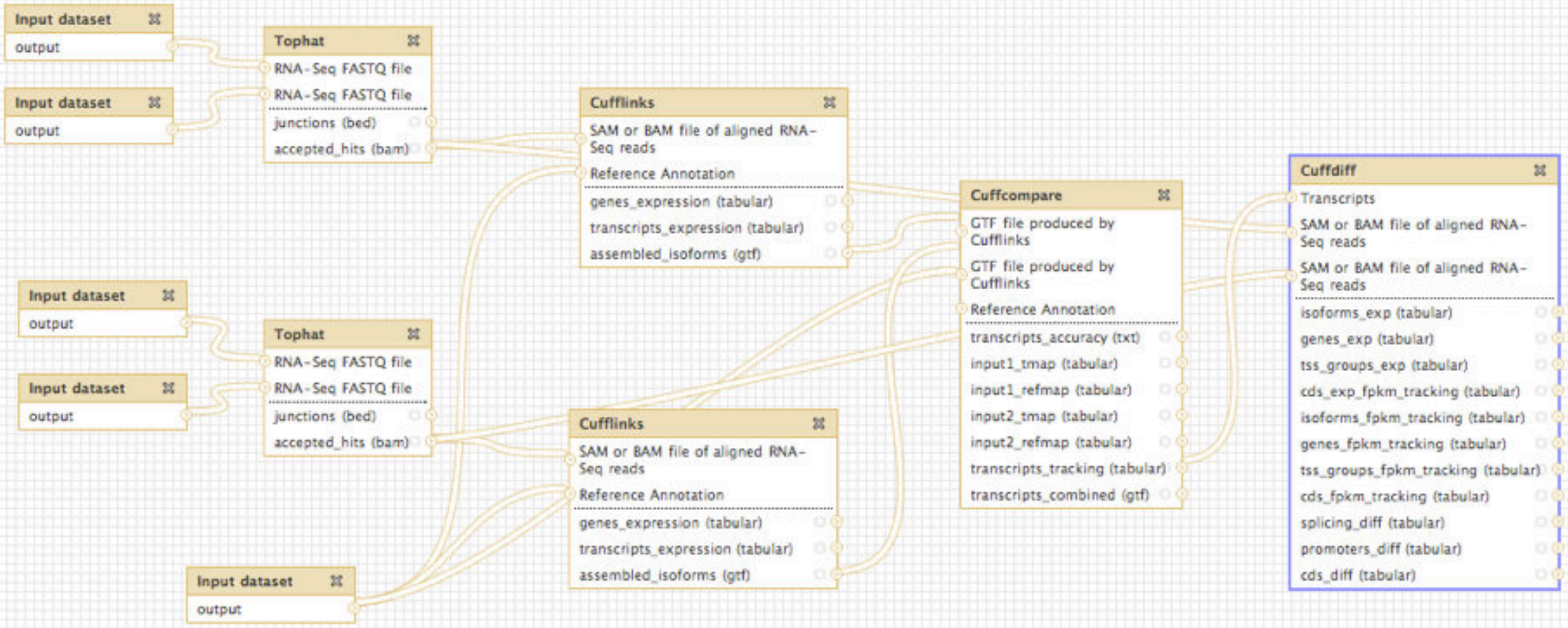


Galaxy Workflows

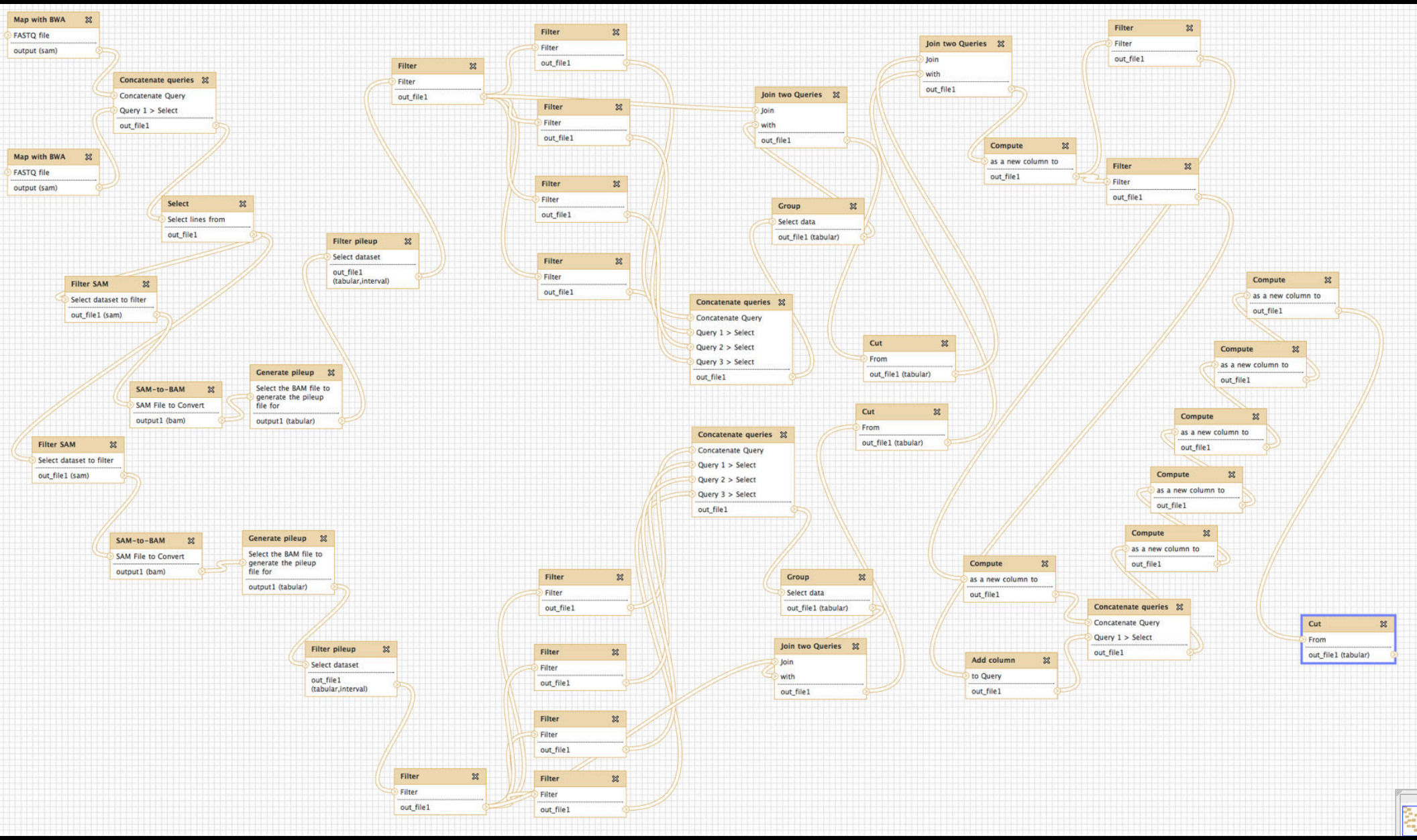


Galaxy Workflows





Example: Workflow for differential expression analysis of RNA-seq using Tophat/Cufflinks tools



Example: Diagnosing low-frequency heterosplasmic sites in two tissues from the same individual

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

Visualize

Send data results to external genome browsers

Trackster: Galaxy's genome browser

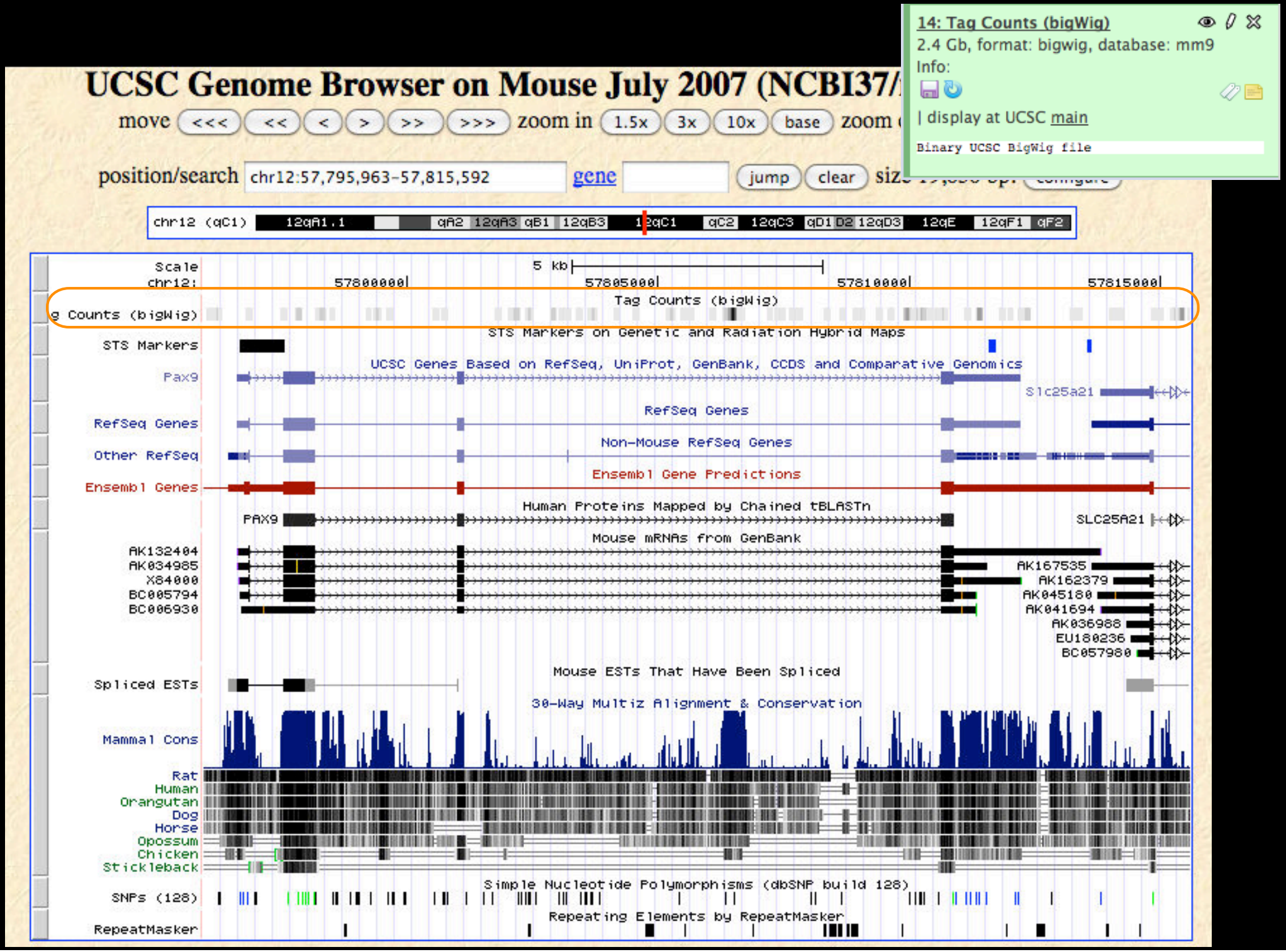
External Genome Browsers

UCSC

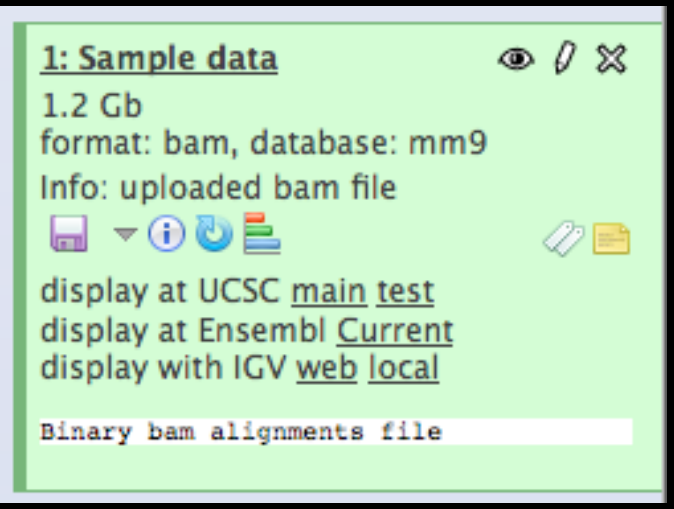
Ensembl

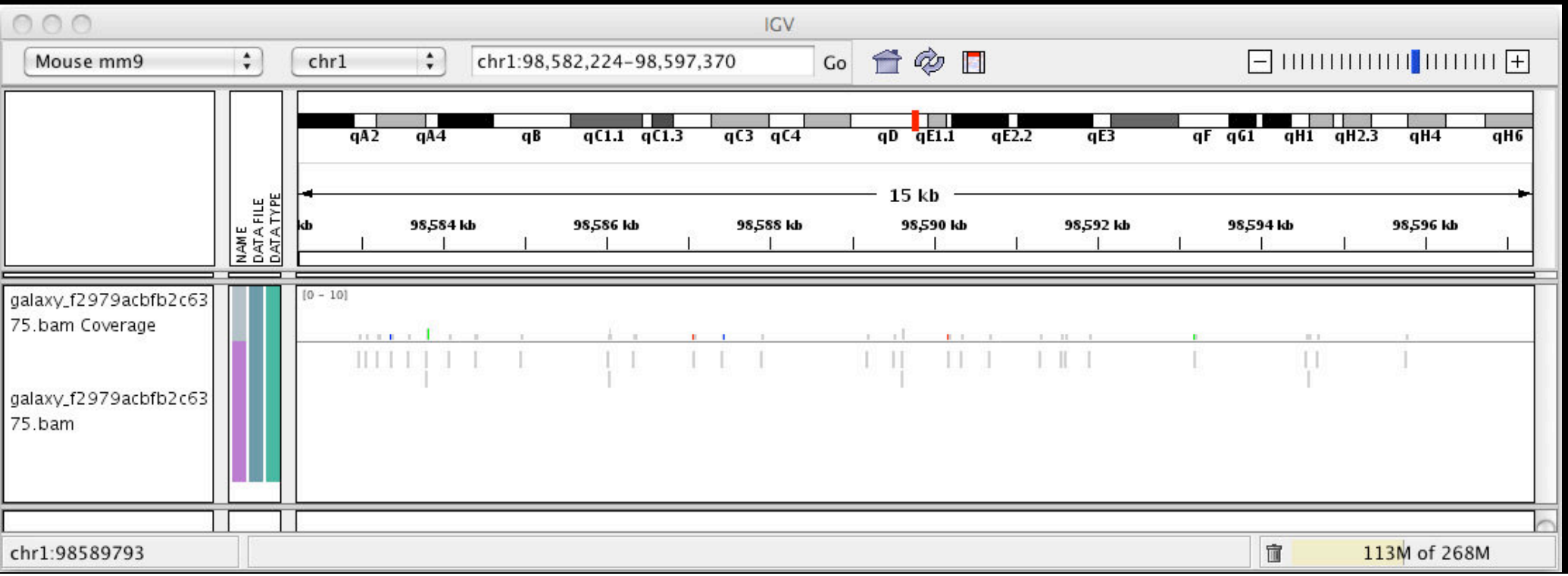
GBrowse

IGV


14: Tag Counts (bigWig)
2.4 Gb, format: bigwig, database: mm9
Info:
| display at UCSC main
Binary UCSC BigWig file
UCSC Genome Browser on Mouse July 2007 (NCBI37/
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom
position/search chr12:57,795,963-57,815,592 gene jump clear
chr12 (qC1) 12qA1.1 qA2 12qA3 qB1 12qB3 12qC1 qC2 12qC3 qD1 D2 12qD3 12qE 12qF1 qF2
Scale
chr12:
5 kb
57800000
57805000
57810000
57815000
Tag Counts (bigWig)
STS Markers
STS Markers on Genetic and Radiation Hybrid Maps
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Pax9
Slc25a21
RefSeq Genes
Other RefSeq
Non-Mouse RefSeq Genes
Ensembl Genes
Ensembl Gene Predictions
Human Proteins Mapped by Chained tBLASTn
PAX9
SLC25A21
Mouse mRNAs from GenBank
AK132404
AK034985
X84000
BC005794
BC006930
AK167535
AK162379
AK045180
AK041694
AK036988
EU180236
BC057980
Spliced ESTs
Mouse ESTs That Have Been Spliced
30-Way Multiz Alignment & Conservation
Mammal Cons
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback
SNPs (128)
Simple Nucleotide Polymorphisms (dbSNP build 128)
RepeatMasker
Repeating Elements by RepeatMasker

Integrative Genomics Viewer (IGV)





Galaxy

- tool integration framework
- heavy focus on usability
- sharing, publication framework

Genome Browser

- physical depiction of data
- visually identify correlations
- find interesting regions, features

Trackster

Trackster

View your data from within Galaxy

- No data transfers to external site
- Use it locally, even without internet access

Supports common filetypes

- BAM, BED, GFF/GTF, WIG

Unique features

- custom genomes
- highly interactive

Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Galaxy

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

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 1,290 - 4,168,475

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,000

UCSC Main on Human: knownGene (chr19) — Auto (Squish)

UCSC Main on Human: all_est (chr19) — Auto (coverage histogram)

11431

UCSC Main on Human: phyloP46wayPrimates (chr19) — Histogram

1

-1

h1-hESC Tophat Mapped Reads — Auto (coverage histogram)

8732

h1-hESC Cufflinks assembled transcripts — Auto (Squish)

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,000

Display a menu

Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Galaxy

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

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 625,719 - 682,581

630,000 640,000 650,000 660,000 670,000 680,000

UCSC Main on Human: knownGene (chr19) Auto (Squish)

UCSC Main on Human: all_est (chr19) Dense

UCSC Main on Human: phyloP46wayPrimates (chr19) Histogram

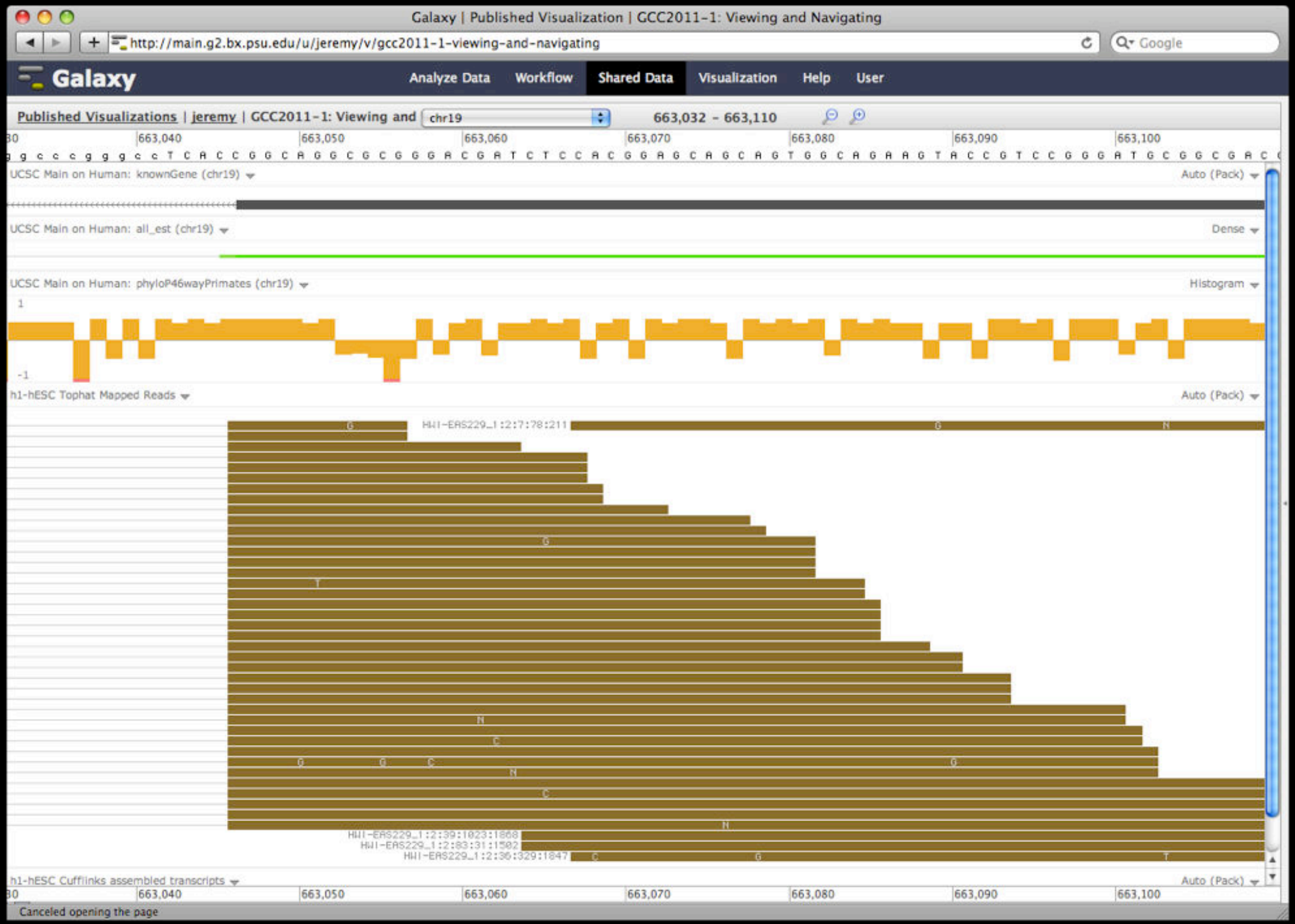
1

-1

h1-hESC Tophat Mapped Reads Auto (Squish)

630,000 640,000 650,000 660,000 670,000 680,000

Display a menu


Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
663,032 - 663,110
663,040
663,050
663,060
663,070
663,080
663,090
663,100
UCSC Main on Human: knownGene (chr19)
Auto (Pack)
UCSC Main on Human: all_est (chr19)
Dense
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
h1-hESC Tophat Mapped Reads
Auto (Pack)
h1-hESC Cufflinks assembled transcripts
Auto (Pack)
Canceled opening the page

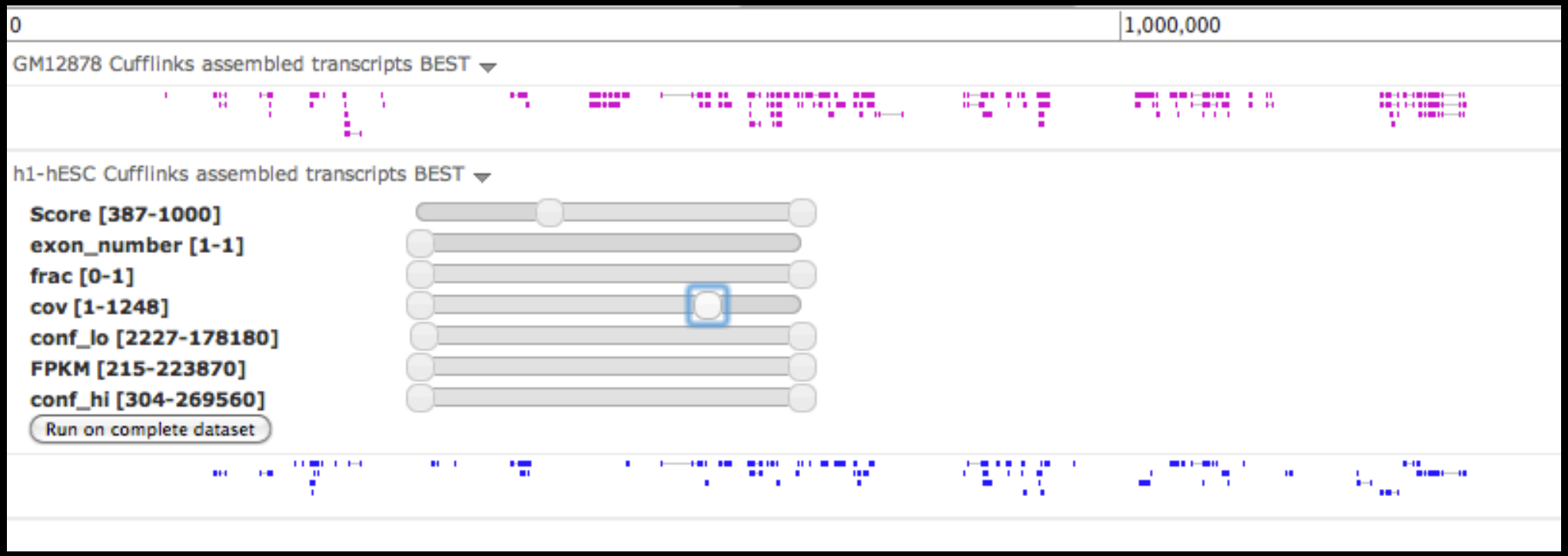
But really, why *another* genome browser

From static browsing to visual analysis

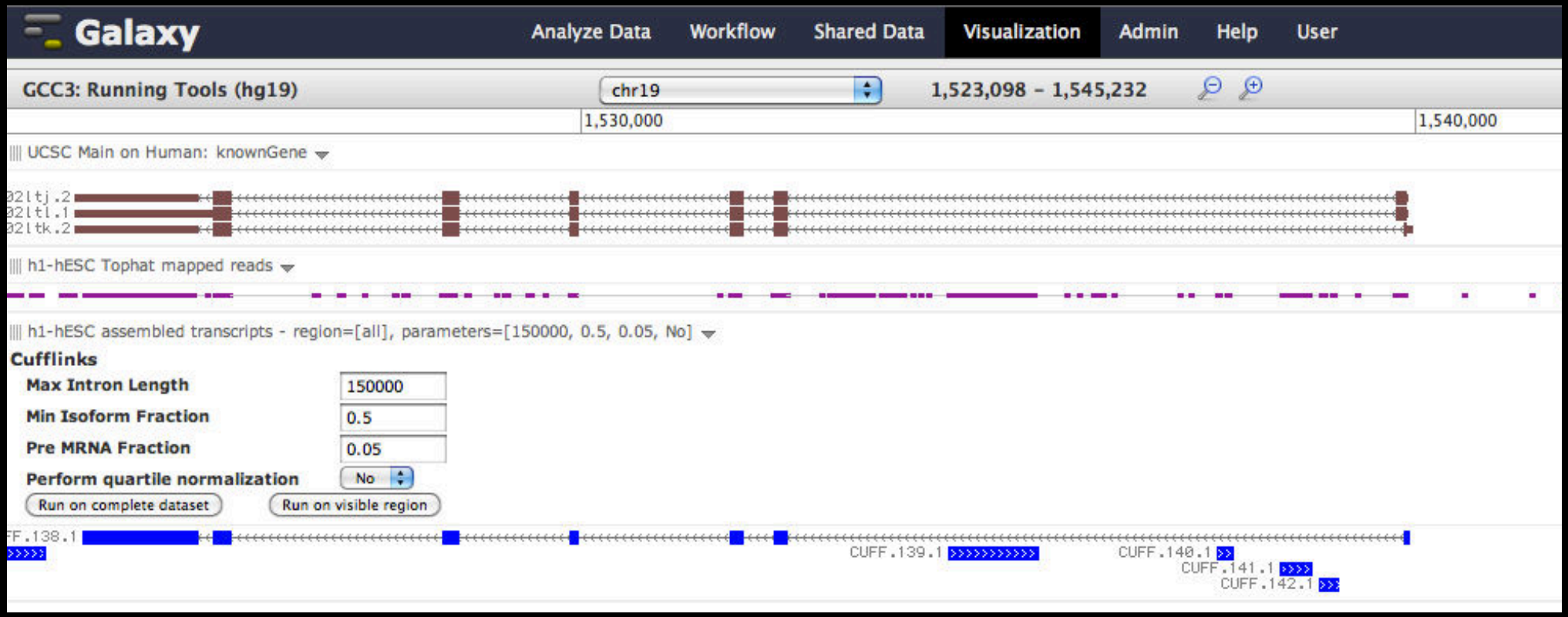
Visual feedback and experimentation needed for complex tools with many parameters

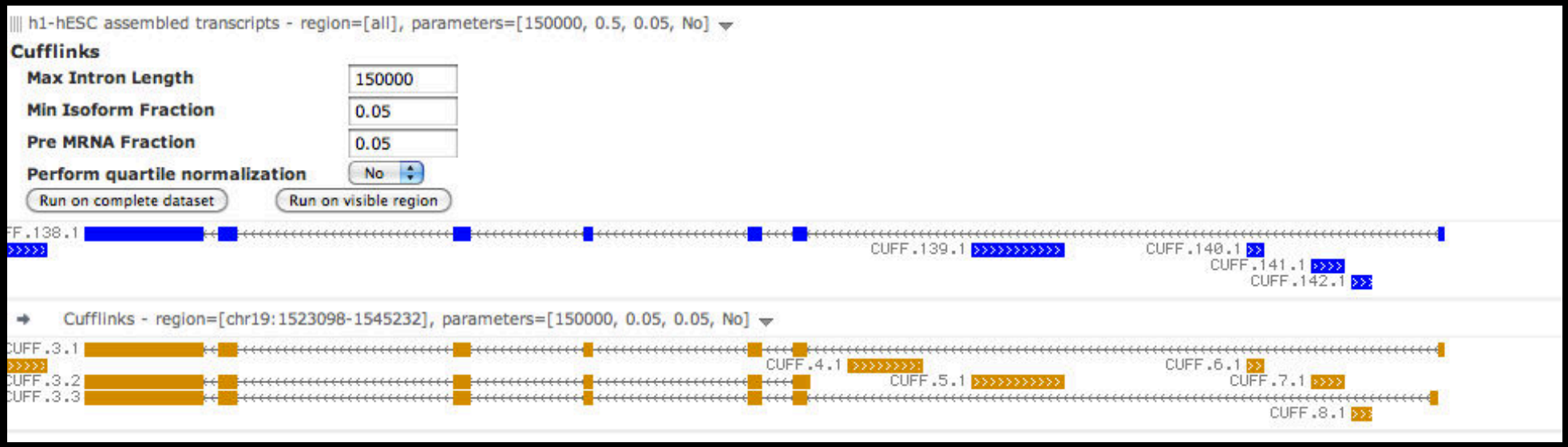
Leverage Galaxy strengths: a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number

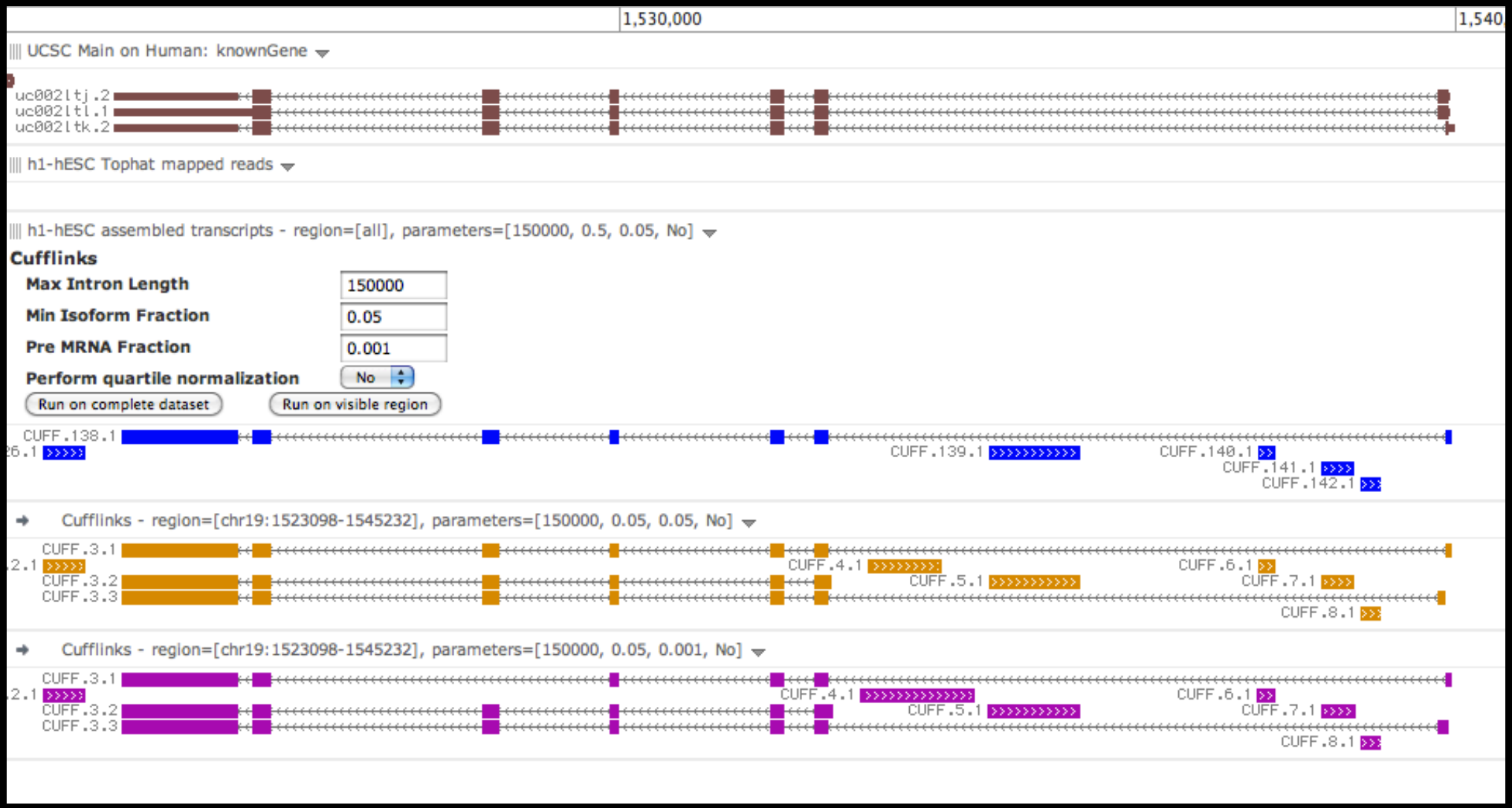
Dynamic Filtering



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

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history is currently restricted so that only you and the users listed below can access it. You can:

Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and import the history.

Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history **accessible via link and published**.

Anyone can view and import this history by visiting the following URL:

http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18

This history is publicly listed and searchable in Galaxy's Published Histories section.

You can:

Unpublish History

Removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Disable Access to History via Link and Unpublish

Disables history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Galaxy | Published History | Variant Analysis for Sample E18

http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18

Galaxy

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

Published Histories | jgoecks | Variant Analysis for Sample E18

Galaxy History ' Variant Analysis for Sample E18'

Import history

Annotation: Perform a pileup analysis with default parameters to identify variants in sample E18.

Dataset	Annotation
1: E18 PE.1 Reads	Forward reads from sample E18.
2: E18 PE.2 Reads	Reverse reads from sample E18.
3: E18 PE.1 Reads Groomed	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
4: E18 PE.2 Reads Groomed	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
5: E18 PE.1 Reads Groomed, Trimmed	Trim reads from 3' end to remove low-quality nts.
6: E18 PE.2 Reads Groomed, Trimmed	Trim reads from 3' to remove low-quality nts.
7: Map with Bowtie for Illumina on data 6 and data 5	Map paired-end reads with default parameters.
8: SAM-to-BAM on data 7	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8	Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18	Find variants with coverage >= 30.
13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes	UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	Variants with consensus different that occur in RefSeq genes.

About this History

Author

jgoecks

Related Histories

All published histories
Published histories by jgoecks

Rating

Community
(1 rating, 4.0 average)

Yours

Tags

Community:
snp | pileup | bowtie | demo | sample

Yours:
snp × | pileup × | bowtie × | demo × | sample:e18 ×

Galaxy | Published Workflow | SNP variant detection from paired-end reads

http://main.g2.bx.psu.edu/u/jgoecks/w/snp-variant-detection-from-paired-end-reads

Google

Galaxy

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

Published Workflows | jgoecks | SNP variant detection from paired-end reads

Step 6: FASTQ Trimmer

Trim reads to remove low-quality bases.

FASTQ File
Output dataset 'output_file' from step 4

Define Base Offsets as
Absolute Values

Offset from 5' end
0

Offset from 3' end
9

Keep reads with zero length
False

Step 7: Map with Bowtie for Illumina

Map reads using default parameter values.

Will you select a reference genome from your history or use a built-in index?
Use a built-in index

Select a reference genome
/galaxy/data/apiMel3/bowtie_index/apiMel3

Is this library mate-paired?
Paired-end

Forward FASTQ file
Output dataset 'output_file' from step 6

Reverse FASTQ file
Output dataset 'output_file' from step 5

Maximum insert size for valid paired-end alignments (-X)
1000

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff)
FR (for Illumina)

Bowtie settings to use
Commonly used

Suppress the header in the output SAM file
True

Step 8: SAM-to-BAM

Convert Bowtie SAM output to BAM format so that pileup can be run.

Choose the source for the reference list
Locally cached

About this Workflow

Author

jgoecks

Related Workflows

All published workflows
Published workflows by jgoecks

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community:
snp bowtie

Yours:
snp bowtie

Galaxy | Published Histories

http://main.g2.bx.psu.edu/history/list_published

Galaxy

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

Published Histories

search | Advanced Search

Name	Annotation	Owner	Community Rating ↑	Community Tags	Last Updated
Galaxy vs MEGAN	Comparison of Galaxy vs. MEGAN pipeline.	aun1	★★★★★	metagenomics, megan, galaxy	Mar 19, 2010
metagenomic analysis		aun1	★★★★★	metagenomics, galaxy	Mar 19, 2010
SM_1186088	Datasets correspond to our paper published in Science by Peleg et al. entitled : Altered histone acetylation is associated with age-dependent memory impairment. Experiment layout: This history contains 4 datasets in the form of BED files of uniquely mapped reads produced after chip-seq for histone modifications H4K12ac and H3K9ac in mouse hippocampus of 3 months (young) and 16 months (old) mice after fear conditioning. For detailed information please refer to supplementary materials and methods of the respective work by peleg et al.	fischerlab	★★★★★		Apr 19, 2010
Variant Analysis for Sample E18	Perform a pileup analysis with default parameters to identify variants in sample E18.	jgoecks	★★★★☆	snp, pileup, bowtie, demo, sample	2 minutes ago
get longest exon		henri	☆☆☆☆☆	chr22, longest, marc, exon, human, workshop	Sep 02, 2010
FASTA to Tabular Test		JJ	☆☆☆☆☆		Aug 26, 2010
EKLF		yzc109	☆☆☆☆☆		Aug 24, 2010

Open "http://main.g2.bx.psu.edu/history/list_published?sort=rating&f-tags=All" in a new tab

Sharing Trackster Visualizations

"A picture is worth a 1000 words."

A fully-interactive visualization is worth many more words

Galaxy | Published Visualiz

main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Galaxy

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

Published Visualizations | jeremy | GCC2011-1: Viewin chr19 1,290 - 4,168,475

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,00

UCSC Main on Human: knownGene (chr19) — Auto (Squish)

UCSC Main on Human: all_est (chr19) — Auto (coverage histogram)

11431

UCSC Main on Human: phyloP46wayPrimates (chr19) — Histogram

1

-1

h1-hESC Tophat Mapped Reads — Auto (coverage histogram)

8732

h1-hESC Cufflinks assembled transcripts — Auto (Squish)

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,00

About this Visualization

Author

jeremy

Related Visualizations

All published visualizations
Published visualizations by jeremy

Rating

Community
(1 rating, 5.0 average)

Yours

Tags

Community:
cool

Yours:

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

Galaxy Pages

A web-based, interactive medium for presenting all aspects of an analysis: data, methods, and results

Galaxy Pages

Galaxy | Published Page | Variant Analysis for sample E18

http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18

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

Published Pages | jgoecks | Variant Analysis for sample E18

Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team

Results

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Galaxy History | Variant Analysis for Sample E18
Perform a pileup analysis with default parameters to identify variants in sample E18.

Here is a workflow for performing this analysis:

Galaxy Workflow | Variant identification within annotated genes from NGS PE Data
Identify variants in annotated genes from NGS paired-end data.

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

About this Page

Author

jgoecks

Related Pages

All published pages
Published pages by jgoecks

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community: none

Yours:

Galaxy Pages

Galaxy | Published Page | Variant Analysis for sample E18

http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18

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

Published Pages | jgoecks | Variant Analysis for sample E18

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Galaxy History | Variant Analysis for Sample E18
Perform a pileup analysis with default parameters to identify variants in sample E18.

8: SAM-to-BAM on data 7	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8	Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18	Find variants with coverage >= 30.
13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes	UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	Variants with consensus different that occur in RefSeq genes.

Here is a workflow for performing this analysis:

Galaxy Workflow | Variant identification within annotated genes from NGS PE Data
Identify variants in annotated genes from NGS paired-end data.

References

About this Page

Author

jgoecks

Related Pages

All published pages
Published pages by jgoecks

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community: none

Yours:

Galaxy Pages

Galaxy | Dataset | Filter to get Variants from sample E18 where consensus base different than ref. base

http://main.g2.bx.psu.edu/u/jgoecks/d/fdcdbcfd1207fbd2

Galaxy

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

Dataset | Filter to get Variants from sample E18 where consensus base different than ref. base

Galaxy Dataset ' Filter to get Variants from sample E18 where consensus base different than ref. base'

Annotation: Filter pileup to find variants where the consensus base is different than the reference base.

This dataset is large and only the first megabyte is shown below. | Show all

chr10	14465082	14465083	T	K	173	176	60	35	GGGGG.GGGGGGGG.GG.....
chr10	14465083	14465084	G	K	144	144	60	35	.....T.....T..TTTTT
chr10	14465485	14465486	C	T	129	129	60	34	t$TTTTTTTTTTTTTTTTTTTT
chr10	19928287	19928288	G	A	135	135	60	36	AAAAAAAAAAAAAAAAAAAAAA
chr10	19928468	19928469	C	T	132	132	60	35	T$tttttTTttttttTttTTT
chr10	19928494	19928495	C	T	138	138	60	37	TTTTttttttttttTTttTttT
chr10	19928543	19928544	A	G	147	147	60	40	GgGgGgggGGGGGGGGgGGGGG
chr10	28750217	28750218	C	T	138	138	60	37	TTttttttTTTtTttTtTTTTTT
chr10	28750397	28750398	A	C	154	211	60	64	C$.$CCCCCCCCCCcC
chr10	28750423	28750424	C	T	113	113	60	35	T$ttTTTttttTTATTTTTtTT
chr10	28750446	28750447	A	G	165	165	60	46	G$GGGggGGggGGggGGGGGgg
chr10	28750548	28750549	G	C	255	255	60	97	CCCCC$CCCCCCCCCccC
chr10	28750640	28750641	T	C	165	165	60	46	CccccccccCCCCCCCCCCCCC
chr10	28750746	28750747	G	A	202	202	60	58	AAaaaaaaaaaaaAaaaaaaaa
chr10	28750766	28750767	A	G	205	205	60	59	GgGgG$gggggggggggG
chr10	28750769	28750770	T	C	175	175	60	49	ccccCccCCCCCcCccccccCc
chr10	28750924	28750925	C	T	182	217	60	64	TTttTtttTtTtTtTttTTg
chr10	28751092	28751093	a	A	147	0	60	123	,,,,,,,,,,,,gg,,gggg,
chr10	28751096	28751097	a	A	212	0	60	119	gg,,,,,,,,,,,,,,,,,,
chr10	28751114	28751115	g	A	225	235	60	85	aaaa,aaaaaaaaaaaaaaaa
chr10	28751117	28751118	T	A	191	198	60	79	aaaaaaaaaa,$
chr10	28918972	28918973	c	M	114	114	60	75	,$,$,$,$,,,,,,,,,,,,,,
chr10	28918975	28918976	a	A	177	0	60	63	,$,$,$....c,,,,,,,,,,
chr10	28918995	28918996	C	M	154	154	60	48	a$aaaaaaaaaaaaaaaaaaaa
chr10	33613489	33613490	G	A	82	114	60	30	.AAAAAAAAAAAAAAAAA
chr10	36721501	36721502	G	K	129	129	60	43	TTTT..,,,,,,,,,,,,,,,,,
chr10	36721507	36721508	C	Y	51	51	60	54	.$.,,,,,,,,,,,,,,,,,,,,
chr10	36721695	36721696	T	A	120	120	60	31	aaaaaaaaaaa$
chr10	36805412	36805413	A	G	126	126	60	33	GGGGGGGGGGGGGGGGGGGGGG
chr10	36805605	36805606	A	G	120	120	60	31	gggggggggggggggggggggg
chr10	36853176	36853177	C	Y	138	138	60	35	.....T
chr10	36853265	36853266	A	R	17	17	60	37	.....
chr10	36854675	36854676	T	K	99	99	60	48	.....
chr10	36854678	36854679	A	M	94	94	60	48	.....
chr10	36855346	36855347	a	R	159	159	60	50	,,,,,,,,,,,,,,,,,,...GGG
chr10	36855350	36855351	a	R	156	156	60	56	,,,,,,,,,,,,,,,,,,,,GGG
chr10	36855356	36855357	a	A	134	0	60	49	,,,,g,,,,,,,,,,,,,,,,,
chr10	36855366	36855367	g	G	52	0	60	40	.....
chr10	36855370	36855371	g	G	25	0	60	32	.$.$.$.$.$.$.....
chr10	36855665	36855666	A	M	11	11	60	39	ccccc$,,,,,,,,,,,
chr10	39144591	39144592	G	C	126	126	60	33	CCCCCCCCCCCCCCCCCCCCCC
chr10	42722938	42722939	C	Y	63	63	60	34	ttttt$,,,,,,,,,,,,

About this Dataset

Author

jgoecks

Tags

Community: none

Yours:

Galaxy Pages

Galaxy | Published Page | Variant Analysis for sample E18

http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18

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

Published Pages | jgoecks | Variant Analysis for sample E18

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Import history

Galaxy History | Variant Analysis for Sample E18
Perform a pileup analysis with default parameters to identify variants in sample E18.

8: SAM-to-BAM on data 7	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8	Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18	Find variants with coverage >= 30.
13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes	UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	Variants with consensus different that occur in RefSeq genes.

Here is a workflow for performing this analysis:

Galaxy Workflow | Variant identification within annotated genes from NGS PE Data
Identify variants in annotated genes from NGS paired-end data.

References

Open "http://main.g2.bx.psu.edu/history/imp?id=e0b8bd5d661b10c2" in a new tab

About this Page

Author

jgoecks

Related Pages

All published pages
Published pages by jgoecks

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community: none

Yours:

Galaxy Pages

Galaxy

http://main.g2.bx.psu.edu/

Google

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

Tools | Options

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling

RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

Filter pileup

Select dataset:
9: Generate pileup on data 8

which contains:
Pileup with ten columns (with consensus)
See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:
20
No variants with quality below this value will be reported

Do not report positions with coverage lower than:
30
Pileup lines with coverage lower than this value will be skipped

Only report variants?:
Yes
See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:
Yes
See "Output format" below for explanation

Print total number of differences?:
Yes
See "Example 3" below for explanation

Print quality and base string?:
Yes
See "Example 4" below for explanation

Execute

What it does

Allows one to find sequence variants and/or sites covered by a specified number of reads with bases above a set quality threshold. The tool works on six and ten column pileup formats produced with *samtools pileup* command. However, it also allows you to specify columns in the input file manually. The tool assumes the following:

- the quality scores follow phred33 convention, where input qualities are ASCII characters equal to the Phred quality plus 33.
- the pileup dataset was produced by the *samtools pileup* command (although you

History | Options

15: Variants from sample E18, consensus different, in RefSeq Genes

14: UCSC mm9 RefSeq Genes

13: Variants from sample E18 where consensus base different than ref. base

10: Variants from sample E18
26,742 regions, format: interval,

Run this job again

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

1.Chrom	2.Start	3.End	4	5	6
chr10	6882036	6882037	A	A	107
chr10	14243075	14243076	G	G	96
chr10	14243079	14243080	C	C	106
chr10	14465082	14465083	T	K	173
chr10	14465083	14465084	G	K	144
chr10	14465084	14465085	T	T	117

9: Generate pileup on data 8

8: SAM-to-BAM on data 7

7: Map with Bowtie for Illumina on data 6 and data 5

6: E18 PE.2 Reads

Open "http://main.g2.bx.psu.edu/tool_runner/rerun?id=1703758" in a new tab

Galaxy Pages

Galaxy | Published Page | Variant Analysis for sample E18

http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18

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

Published Pages | jgoecks | Variant Analysis for sample E18

Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team

Results

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Galaxy History | Variant Analysis for Sample E18
Perform a pileup analysis with default parameters to identify variants in sample E18.

Here is a workflow for performing this analysis:

Import workflow

Galaxy Workflow | Variant identification within annotated genes from NGS PE Data
Identify variants in annotated genes from NGS paired-end data.

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

[3] Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 -2079 (2009).

About this Page

Author

jgoecks

Related Pages

All published pages
Published pages by jgoecks

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community: none

Yours:

Open "http://main.g2.bx.psu.edu/workflow/imp?id=58d16d45527990b7" in a new tab

Creating a Page

Galaxy

http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427

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

Page Editor | Title : Variant Analysis for sample E18 | Save | Close

Paragraph type | Insert Link to Galaxy Object | Embed Galaxy Object

Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team

Results

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Here is a workflow for performing this analysis:

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

Creating a Page

Galaxy

http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427

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

Page Editor | Title : Variant Analysis for sample E18 — Save | Close

Paragraph type | Insert Link to Galaxy Object | Embed Galaxy Object

Variant Analysis of Embr

Jeremy Goecks, Anton Nekrutenko, James Taylor,

Results

To demonstrate how Galaxy can support accessible, variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq bryonic development.

The initial analysis produced support for 27,742 pos base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the Genes. These potential variants are:

Method

In the first step of this analysis, the reads were groo t, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] f ads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was complete analysis is contained in this history:

Here is a workflow for performing this analysis:

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

Embed Histories

search | Advanced Search

	Name	Tags	Last Updated ↑
☑	Variant Analysis for Sample E18	5 Tags	15 minutes ago
☐	Pileup analysis, sample E18	4 Tags	2 days ago
☐	Unnamed history	0 Tags	Sep 07, 2010
☐	Unnamed history	0 Tags	Dec 17, 2009
☐	imported: Hsitory with ~100 items	5 Tags	Dec 10, 2009
☐	imported: Galaxy vs MEGAN	0 Tags	Dec 04, 2009
☐	imported: Galaxy vs MEGAN	2 Tags	Oct 06, 2009
☐	imported: Galaxy vs MEGAN	0 Tags	Oct 06, 2009
☐	imported: metagenomic analysis	0 Tags	Sep 30, 2009
☐	imported: Galaxy vs MEGAN	0 Tags	Sep 30, 2009

Page: 1 2 | Show all histories on one page

For 1 selected histories:

☑ Make the selected histories accessible so that they can viewed by everyone.

Embed | Cancel

Creating a Page

Galaxy

http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427

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

Page Editor | Title : Variant Analysis for sample E18 | Save | Close

Paragraph type | Insert Link to Galaxy Object | Embed Galaxy Object

Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team

Results

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Embedded Galaxy History 'Variant Analysis for Sample E18'

[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]

Here is a workflow for performing this analysis:

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 12741-12746 (2009).

Open # on this page in a new tab

Creating a Page

Galaxy

http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427

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

Page Editor | Title : Variant Analysis for sample E18 | Save | Close

Paragraph type | Insert Link to Galaxy Object | Embed Galaxy Object

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Embedded Galaxy Dataset 'Variants from sample E18, consensus different, in RefSeq Genes'

[Do not edit this block; Galaxy will fill it in with the annotated dataset when it is displayed.]

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Embedded Galaxy History 'Variant Pileup Analysis for Sample E18'

[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]

Here is a workflow for performing this analysis:

Embedded Galaxy Workflow 'SNP identification within annotated genes from NGS PE Data'

[Do not edit this block; Galaxy will fill it in with the annotated workflow when it is displayed.]

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

[3] Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 -2079 (2009).

The power of Galaxy publishing

Galaxy's publishing features facilitate access and reproducibility without any extra leg work

One click grants access to the *actual analysis* you performed to generate your original results

- Not just data access: the full pipeline
- Annotate each step
- Anyone can import your work and immediately reproduce or build on it

Overview

What is Galaxy?

What you can do in Galaxy

- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Galaxy 101 Exercise

EMORY

Enis Afgan

Dannon Baker

Dave Clements

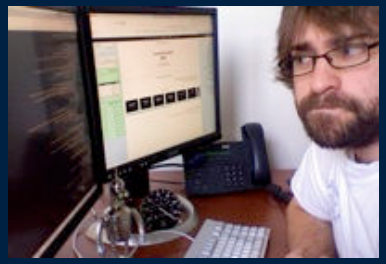
Jeremy Goecks

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

Galaxy 101
http://usegalaxy.org/galaxy101

A simple question...

- Which coding exons have highest number of single nucleotide polymorphisms?

Galaxy 101
http://usegalaxy.org/galaxy101

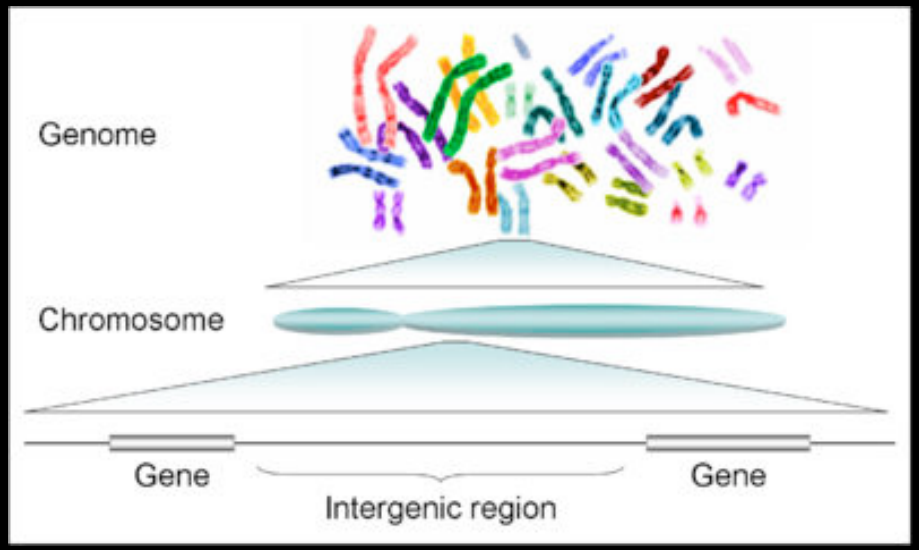
Overview

- Interactively Analyze Data
- Create reusable generic Workflow
- Share analysis Results, History, Workflow

Required Data

- Genomic Coordinates of coding exons and SNPs

Genomic Coordinates



http://library.kiwix.org:4201/A/Human_genome.html

```
>chr1
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac
```

chrom	start	end	name	score	strand
chr1	0	10	first_ten_bases	0	+

see also:
https://bitbucket.org/galaxy/galaxy-central/wiki/GopsDesc
https://bitbucket.org/galaxy/galaxy-central/wiki/zero_based_coordinates.pdf

Galaxy 101: Basic Steps
http://usegalaxy.org/galaxy101

Get Genomic data from UCSC Table Browser

Determine each SNP that overlaps with a specific coding exon

Calculate count of overlapping SNPs for each exon

Sort and select exons by greatest SNP counts

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......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
!"#$%&'()*+,-./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

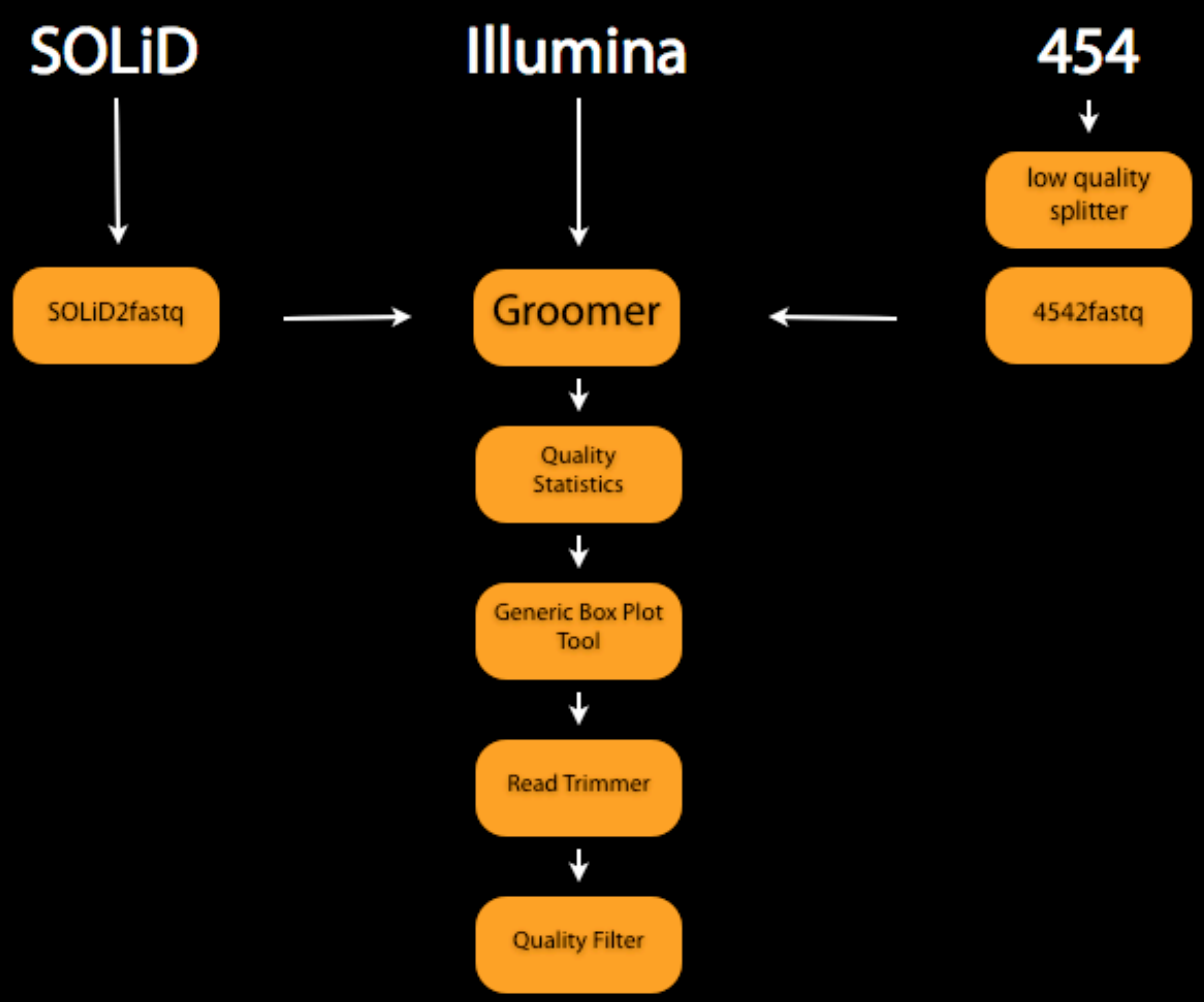
High-throughput Sequencing (HTS) Data

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

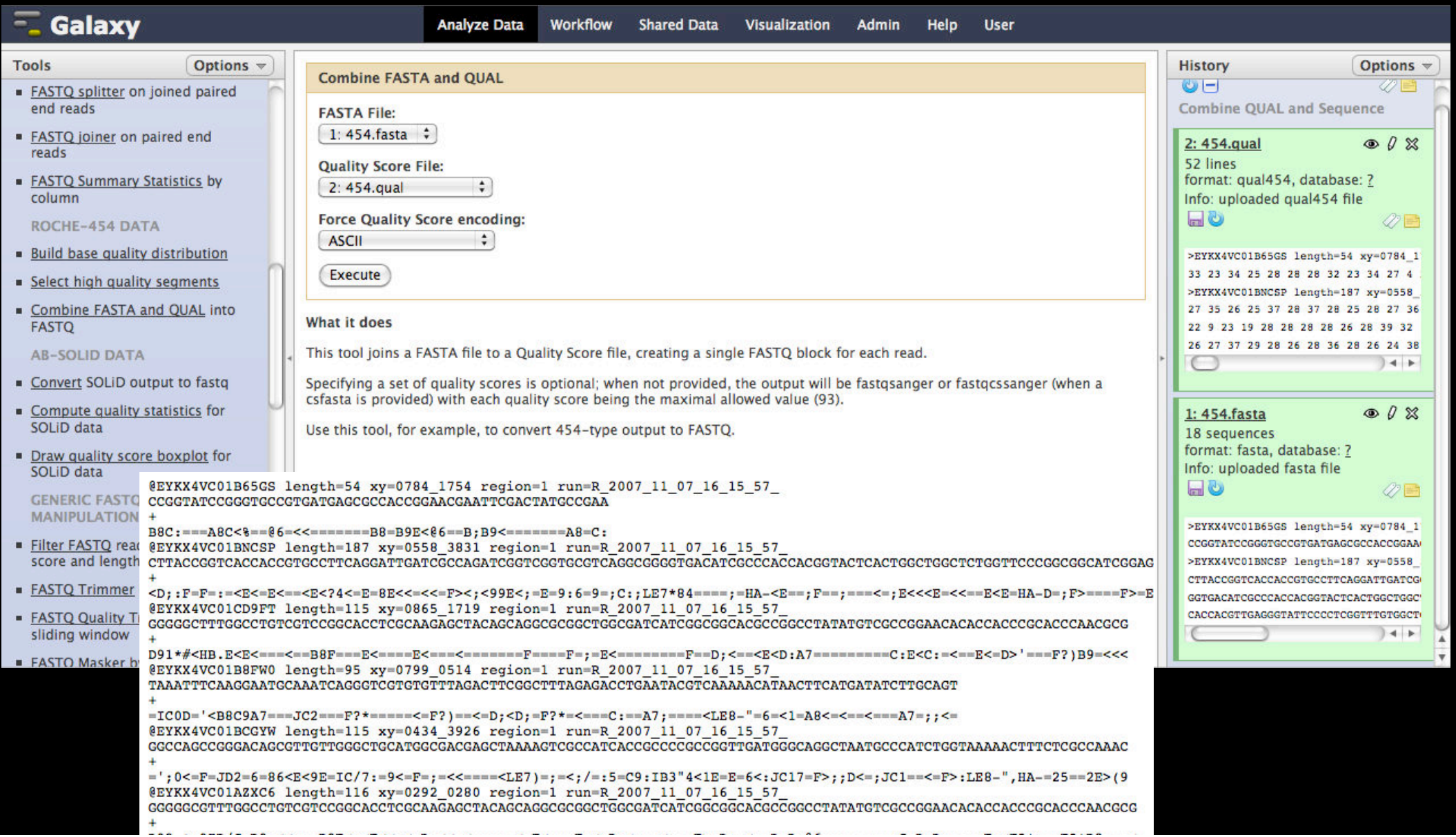
Galaxy exercises: ChIP-seq and RNA-seq

Prepare and Quality Check

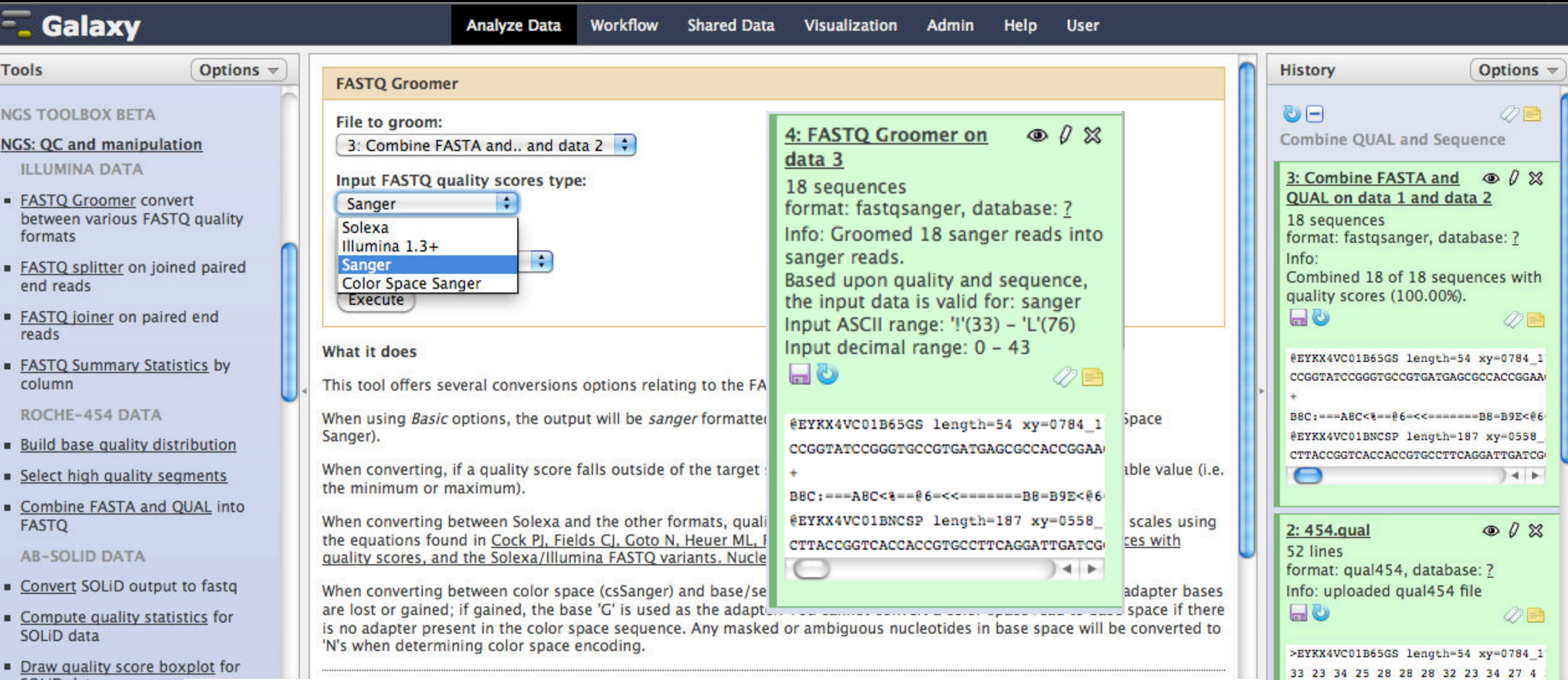


Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5.

Combining Sequences and Qualities



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

- Solexa
- Illumina 1.3+
- Sanger
- Color Space 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

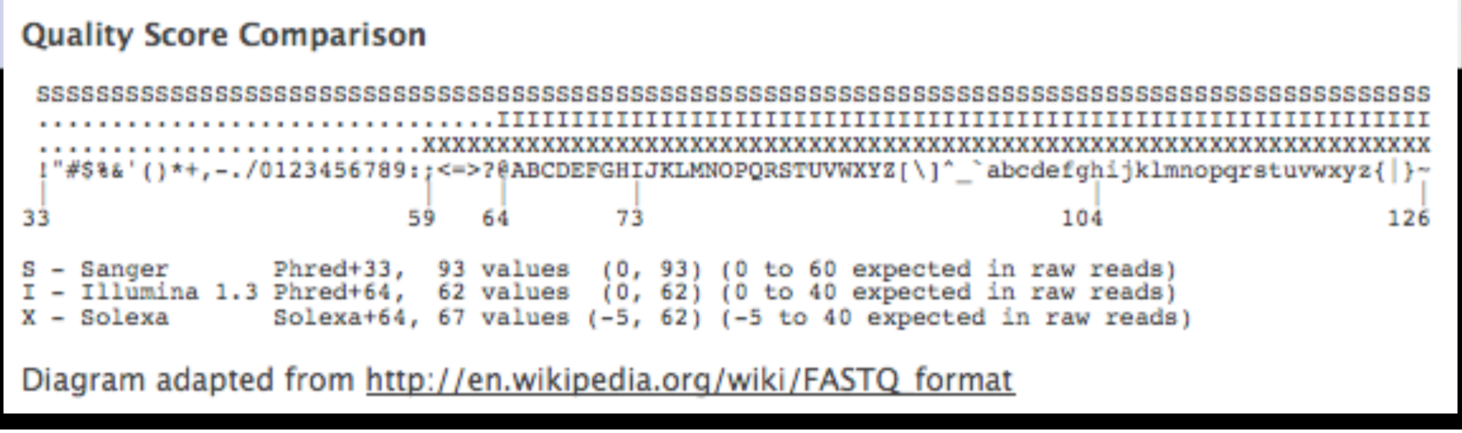
18 sequences
format: fastqsanger, database: ?
Info:
Combined 18 of 18 sequences with quality scores (100.00%).

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

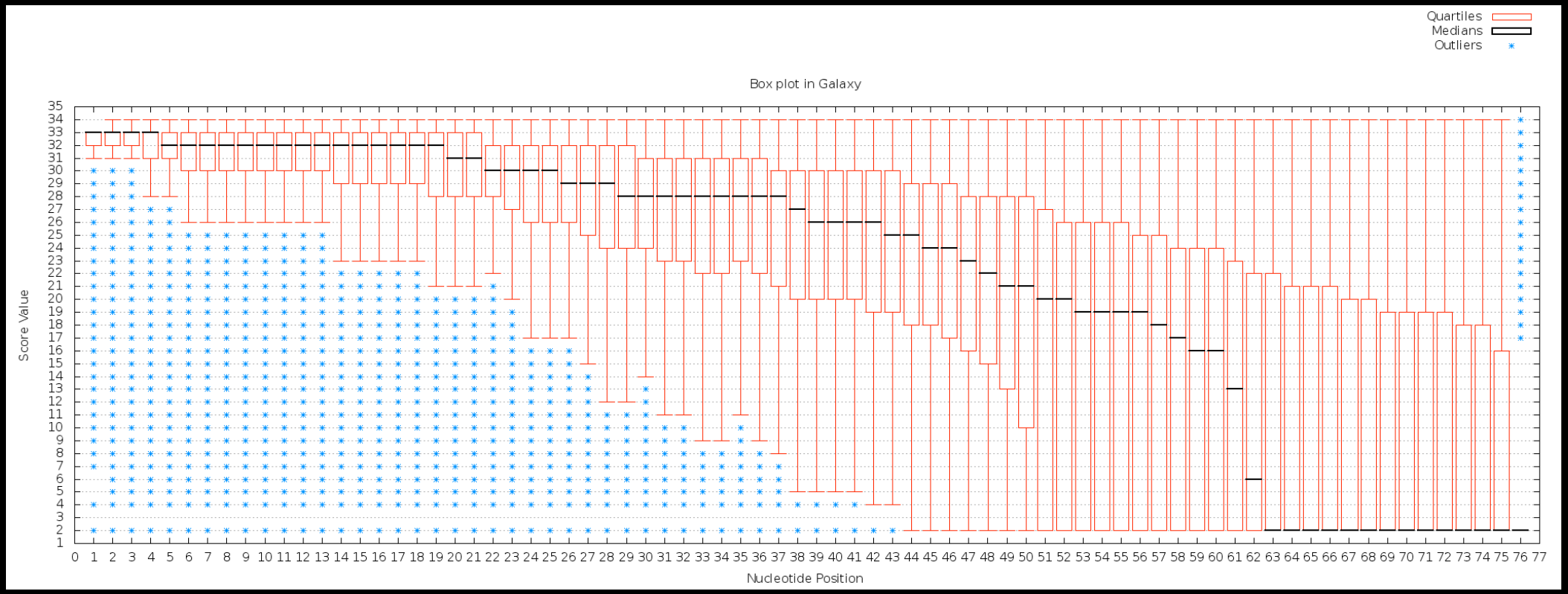
NGS: QC and manipulation

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

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 agg

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

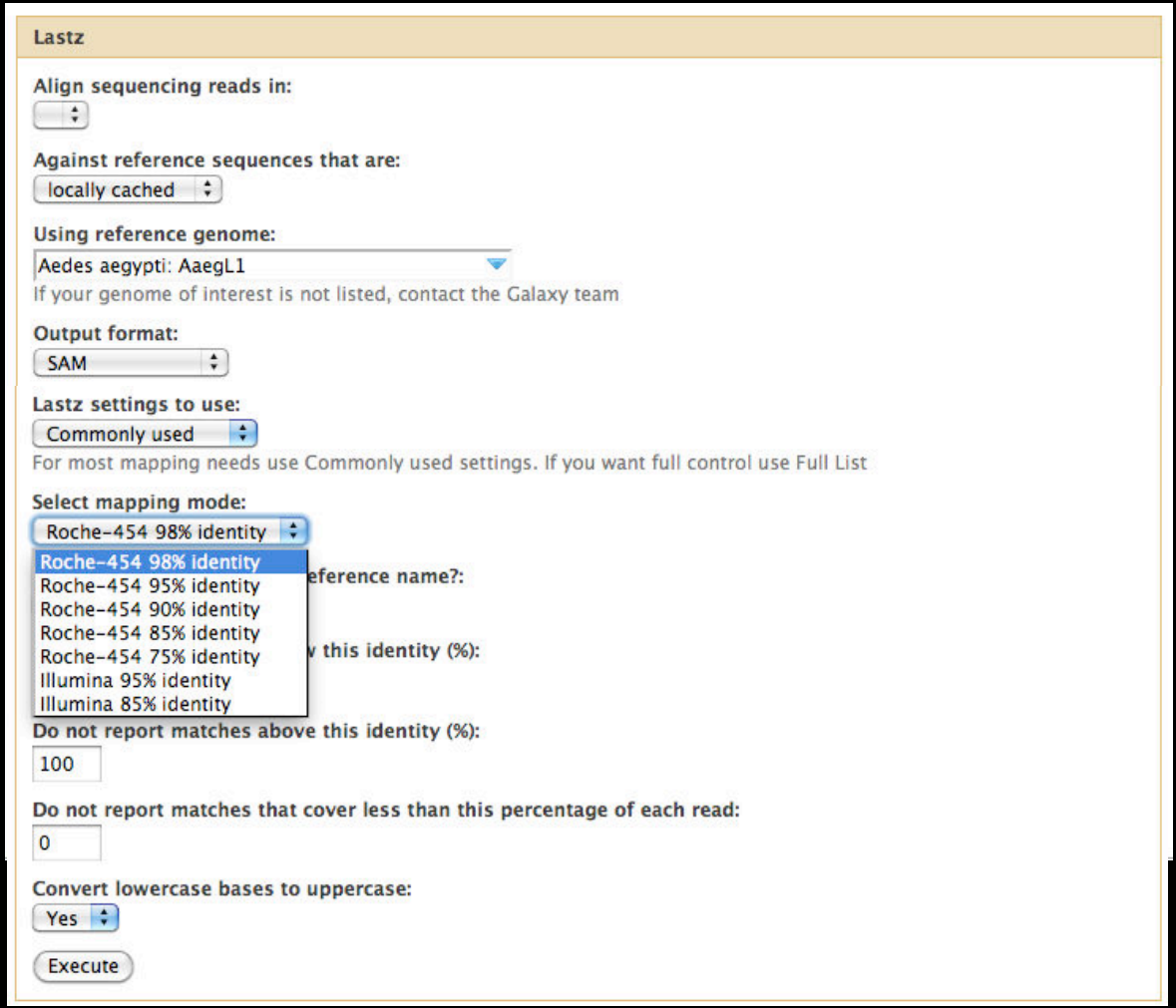
Metagenomics

- 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

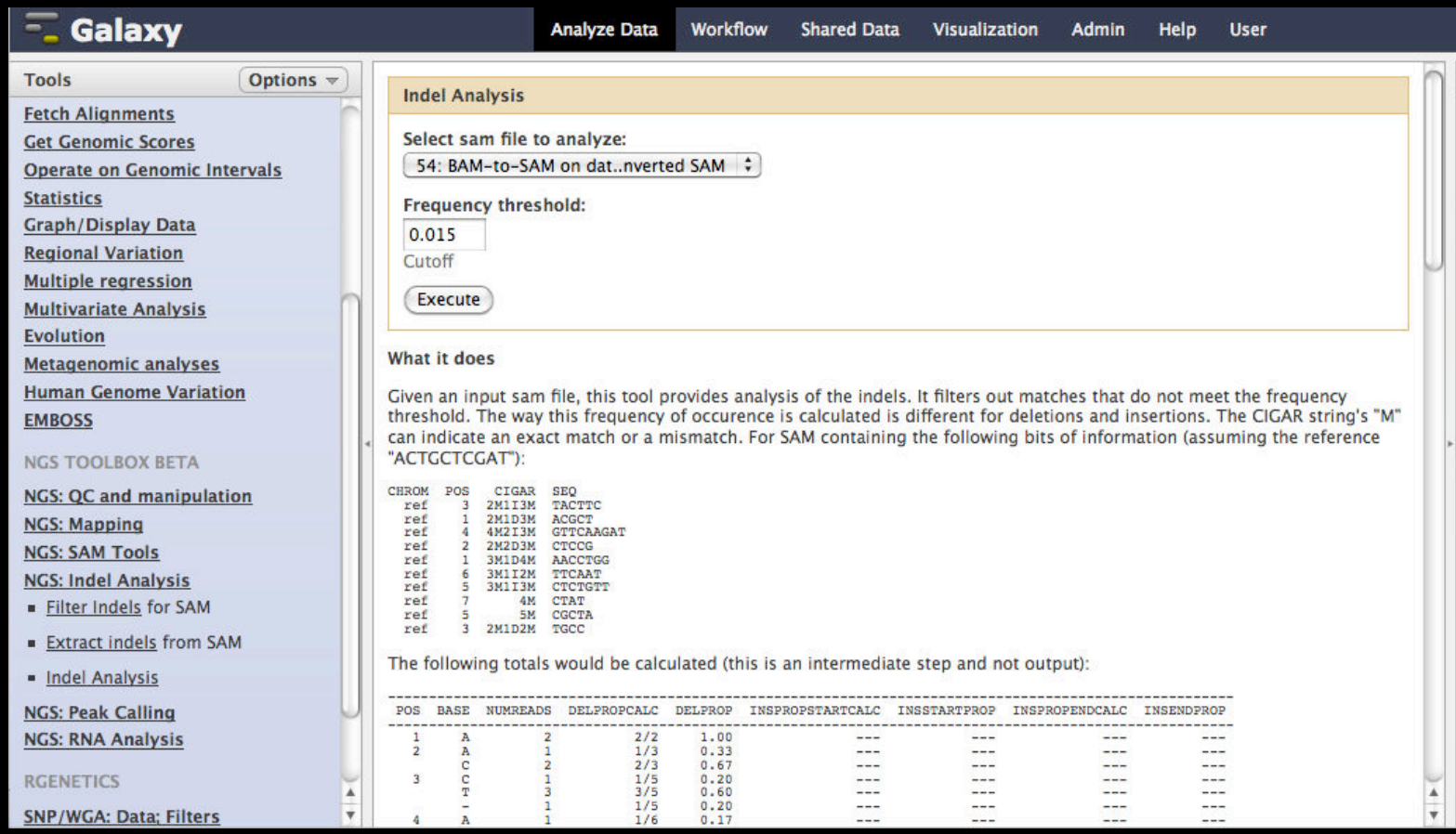
SNPs & INDELs

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/

Unified Genotyper

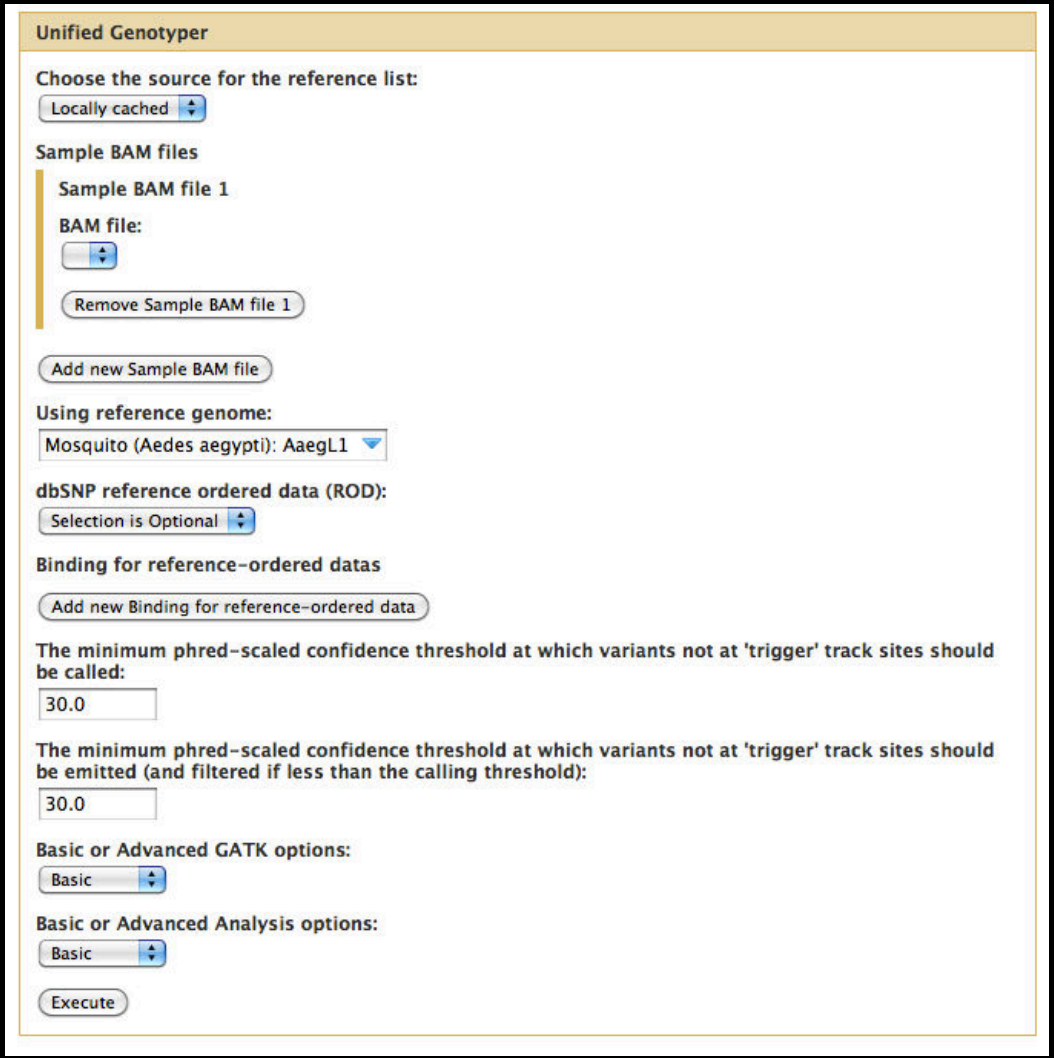
Inputs

- 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

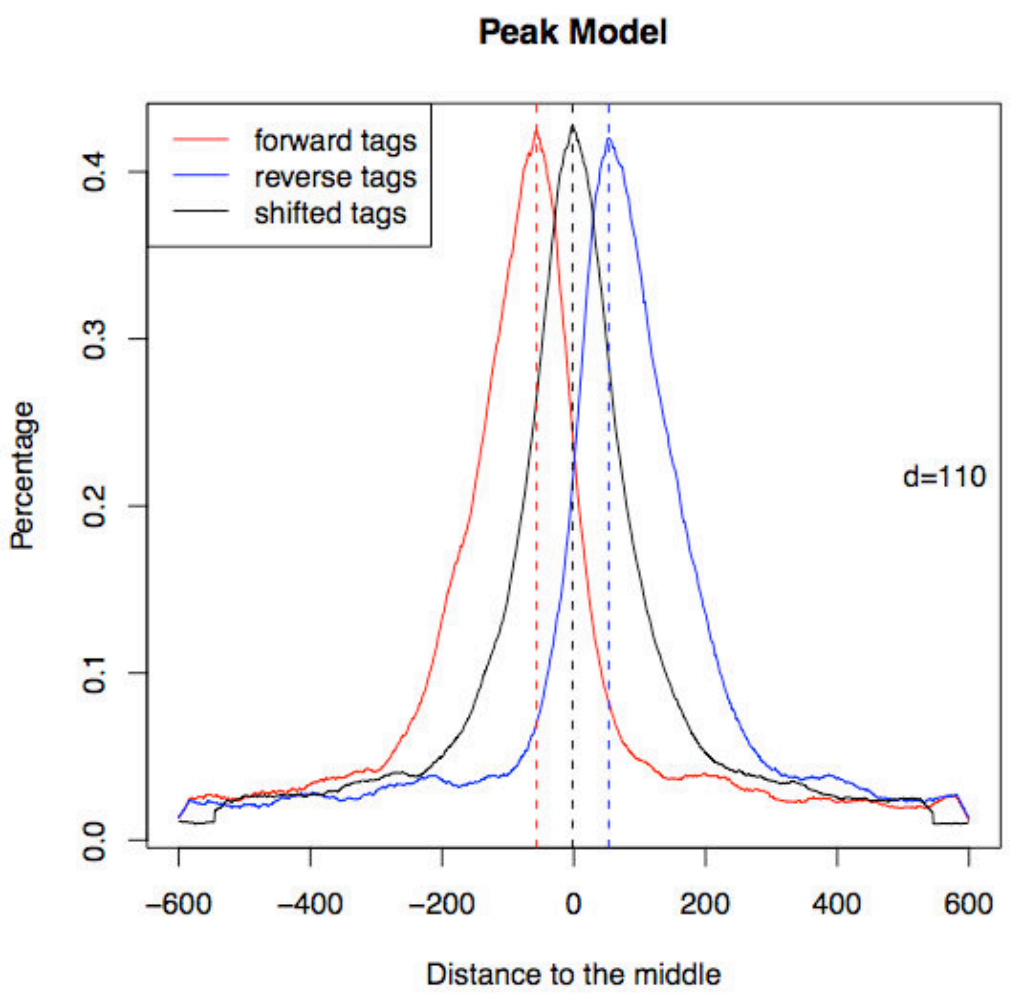
Punctate Binding --> MACS

Inputs

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

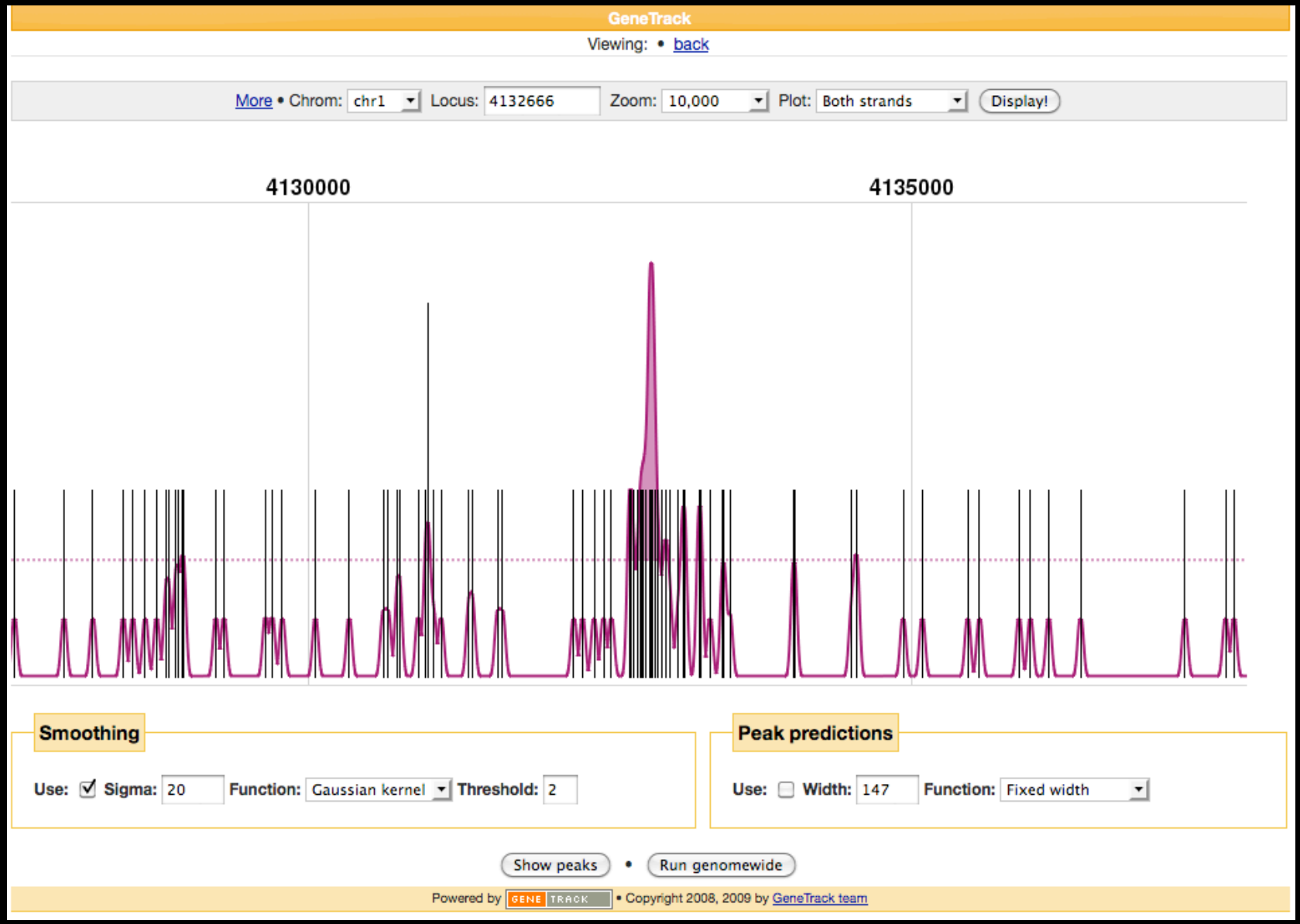
Outputs

- 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

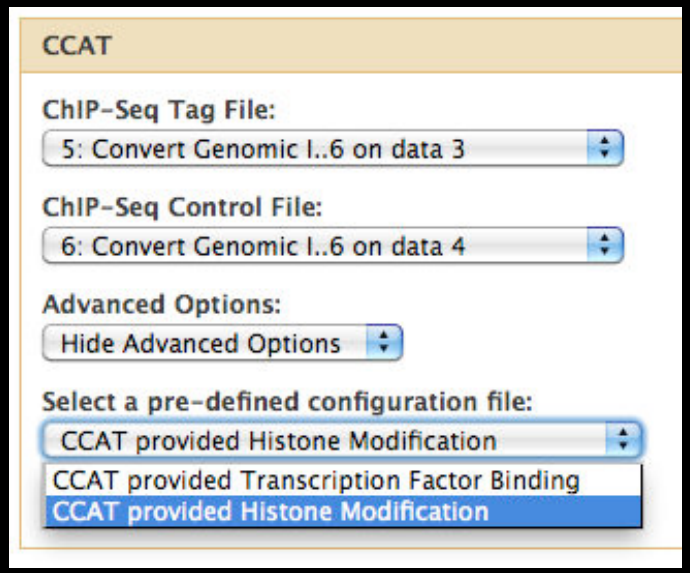
MACS --> GeneTrack

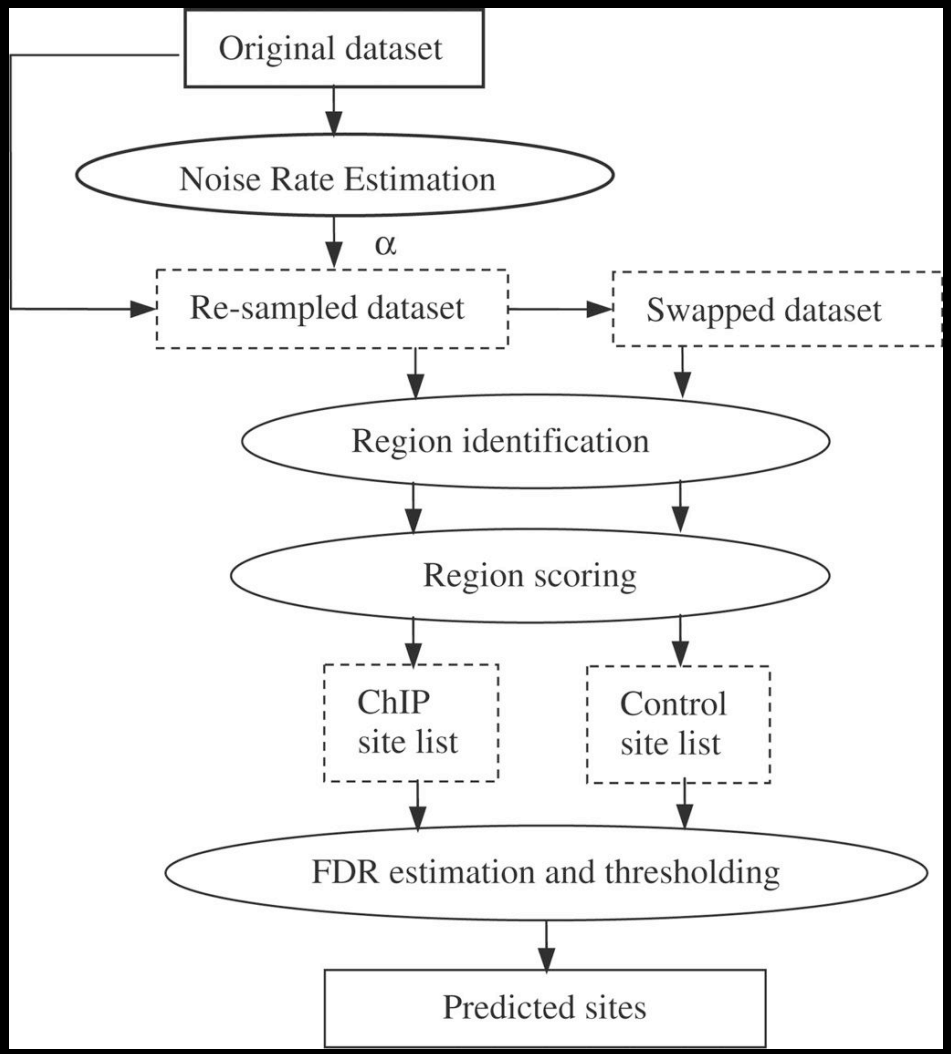


Albert I, Wachi S, Jiang C, Pugh BF. GeneTrack--a genomic data processing and visualization framework. Bioinformatics. 2008 May 15;24(10):1305-6. Epub 2008 Apr 3.

Diffuse Binding

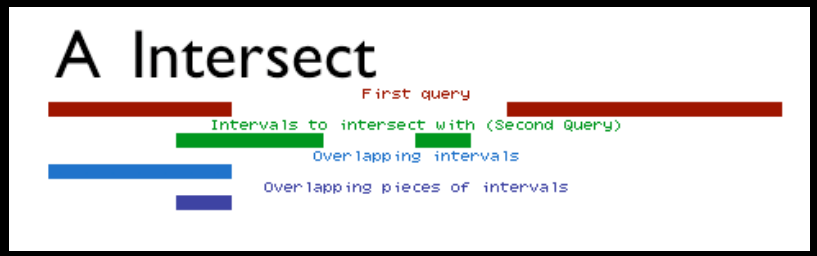
CCAT (Control-based ChIP-seq Analysis Tool)

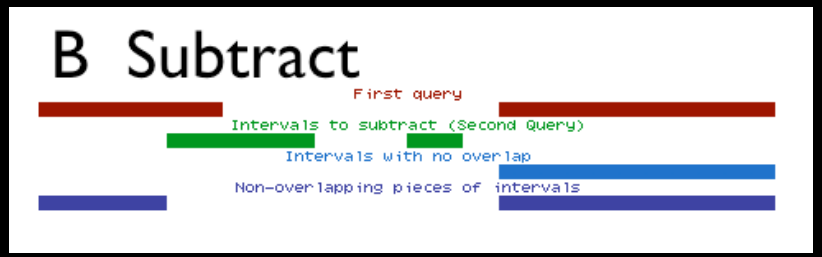


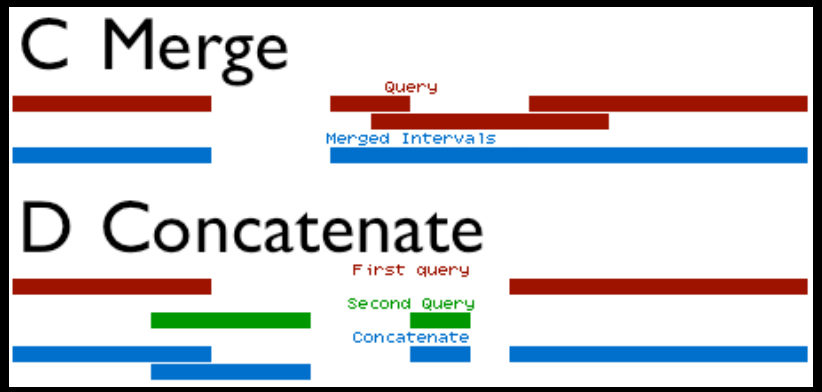


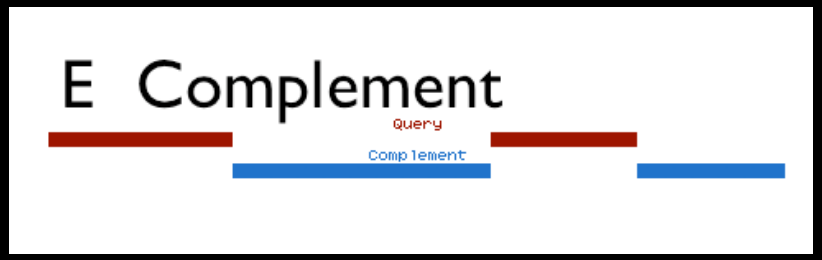
Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F, Sung WK. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics. 2010 May 1;26(9):1199-204.

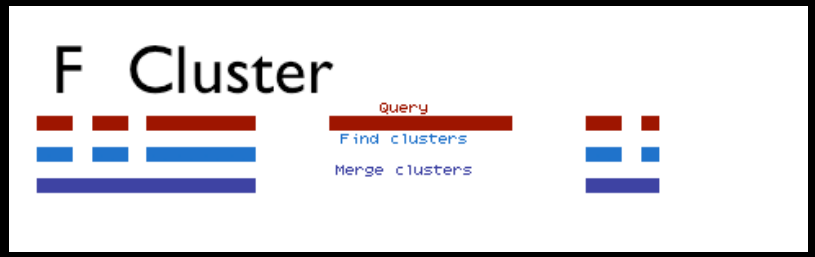
I have Peaks, now what?











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

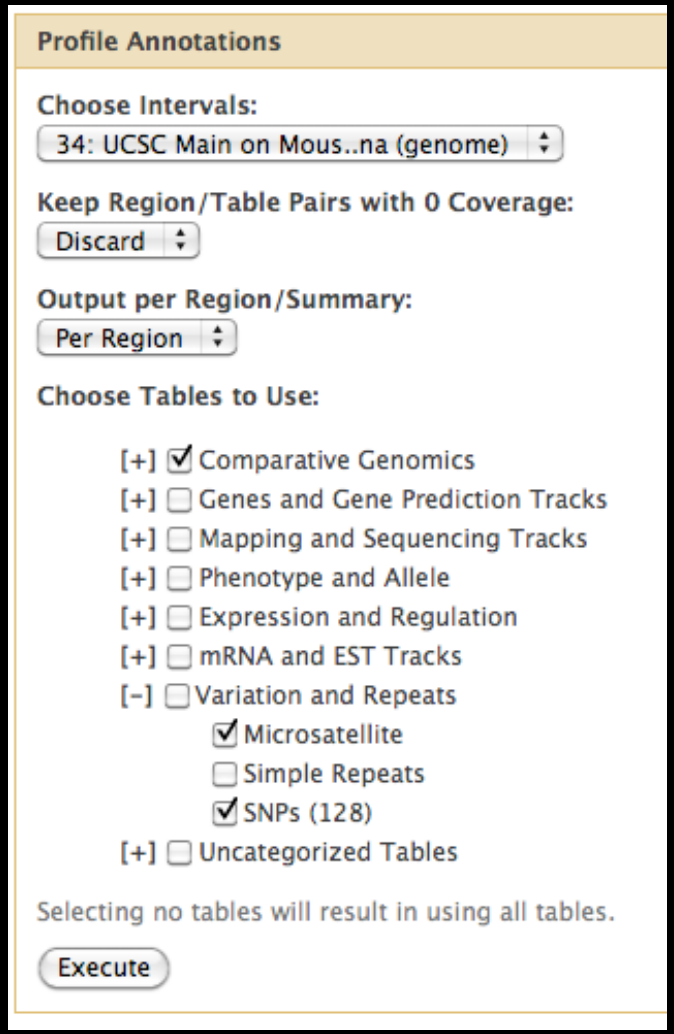
1.Chrom	2.Start	3.End	4.Name	5	6.
chr1	134212701	134230065	NM_028778	0	+
chr1	134212701	134230065	NM_001195025	0	+
chr1	33510655	33726603	NM_008922	0	-
chr1	58714963	58752833	NM_175370	0	-
chr1	25124320	25886552	NM_175642	0	-

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

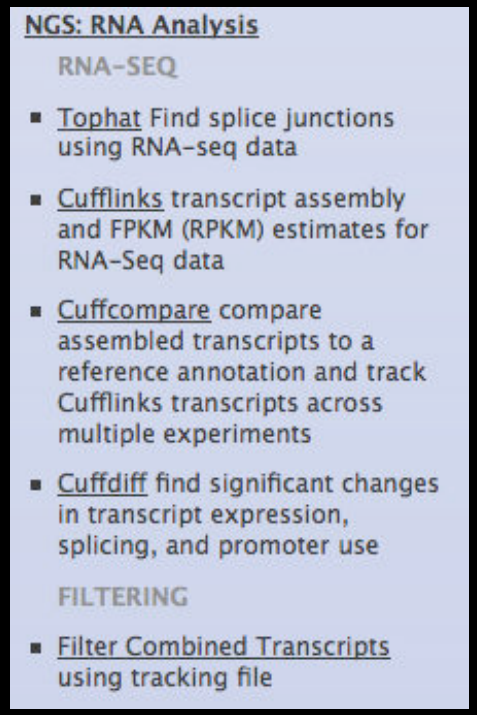
- Prepare, quality control and manipulate reads
- Read Mapping
- SNP & INDEL analysis
- Binding sites analysis and peak calling
- Transcriptome analysis

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

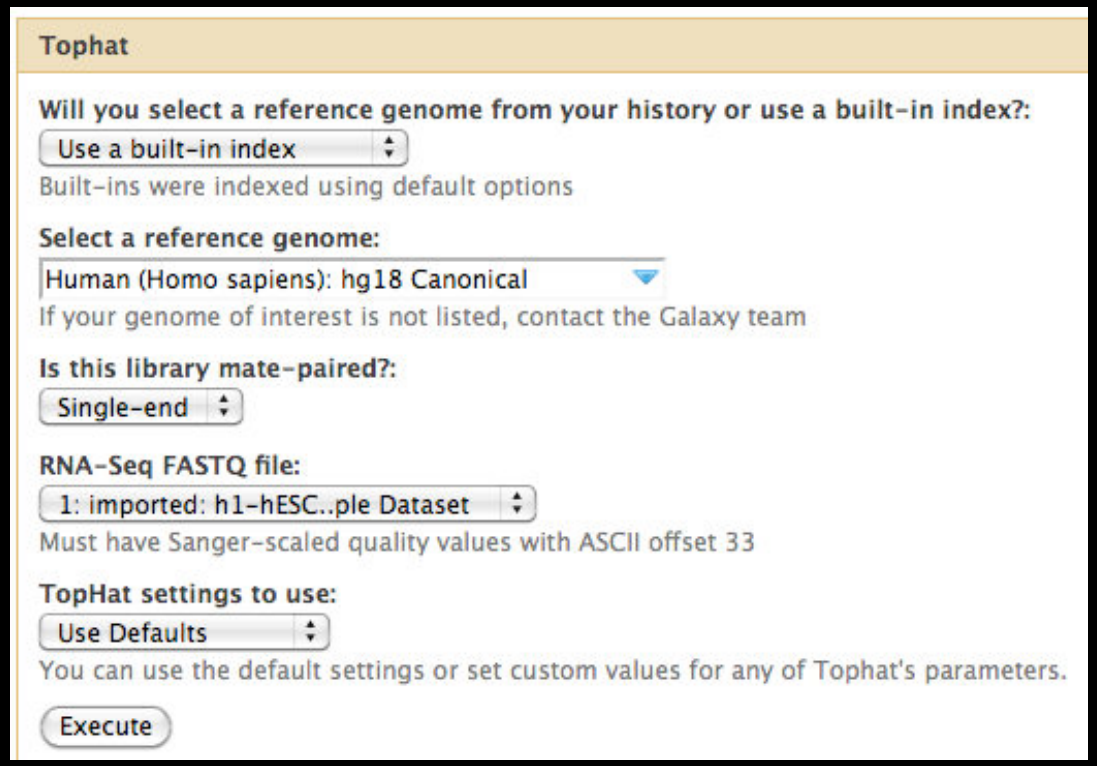
TopHat

Map RNA (FASTQ) to a reference Genome

- gapped mapper

Outputs

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



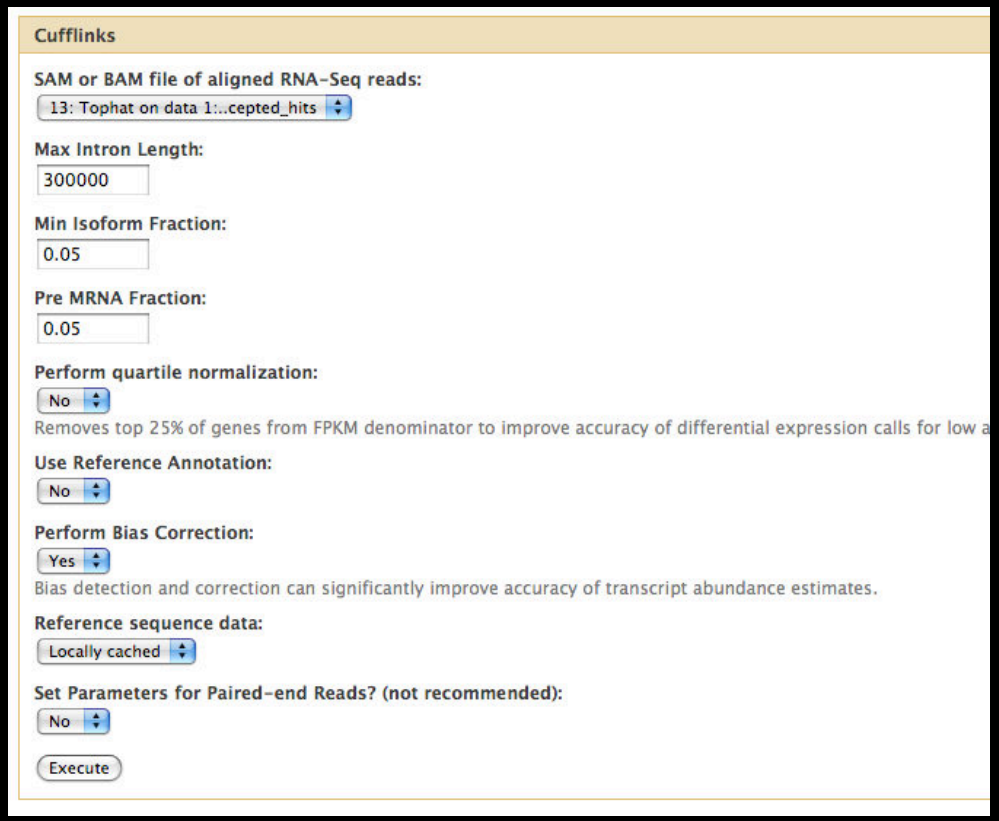
Cufflinks

Goal: transcript assembly and quantitation

Input: aligned RNA-Seq reads, usually from TopHat

Outputs

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



Cuffcompare

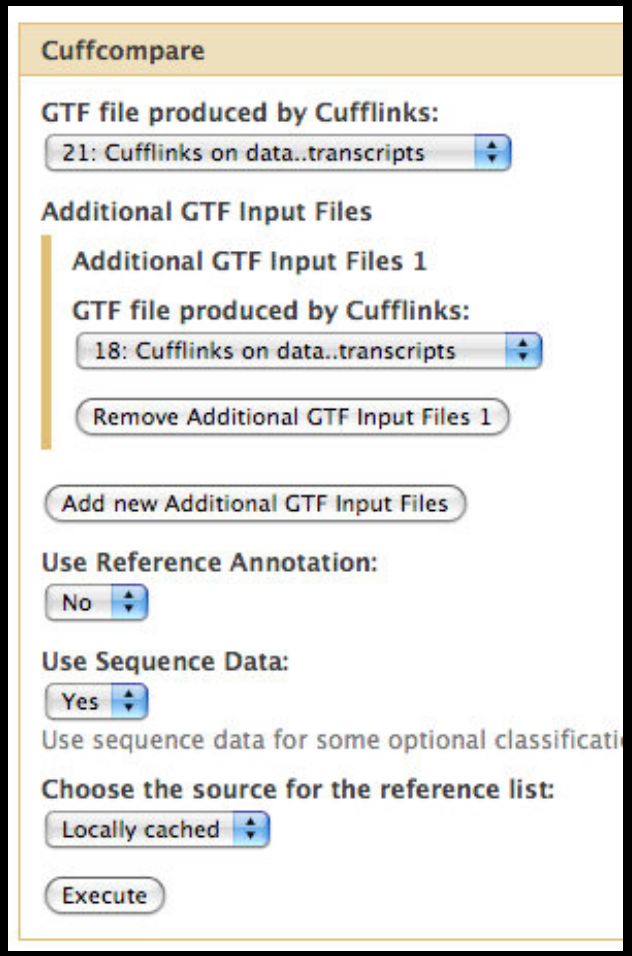
Goals

- 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

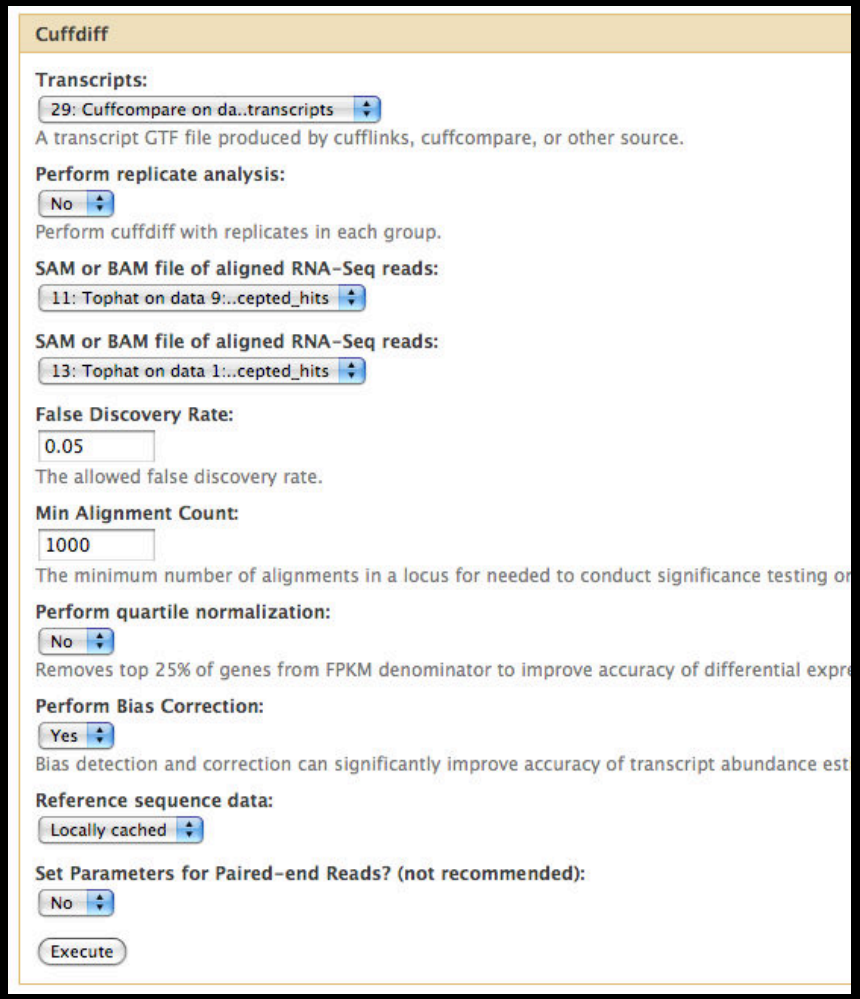
- differential expression testing
- transcript quantitation

Inputs

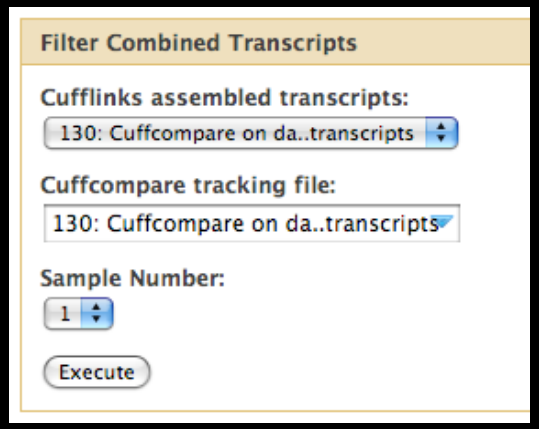
- Combined set of transcripts
- mapped reads from 2+ samples

Outputs

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

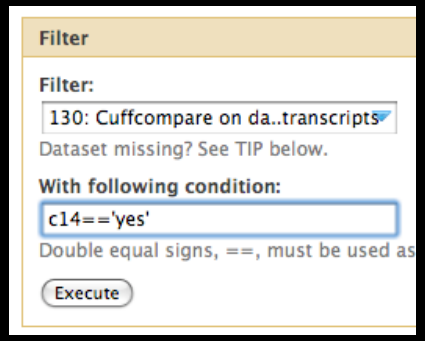


Next Steps

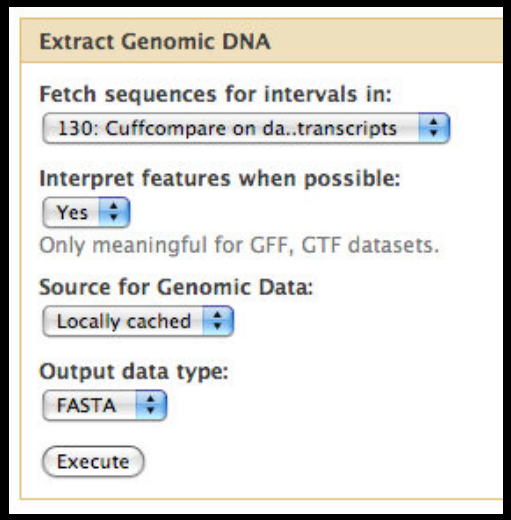


Filtering

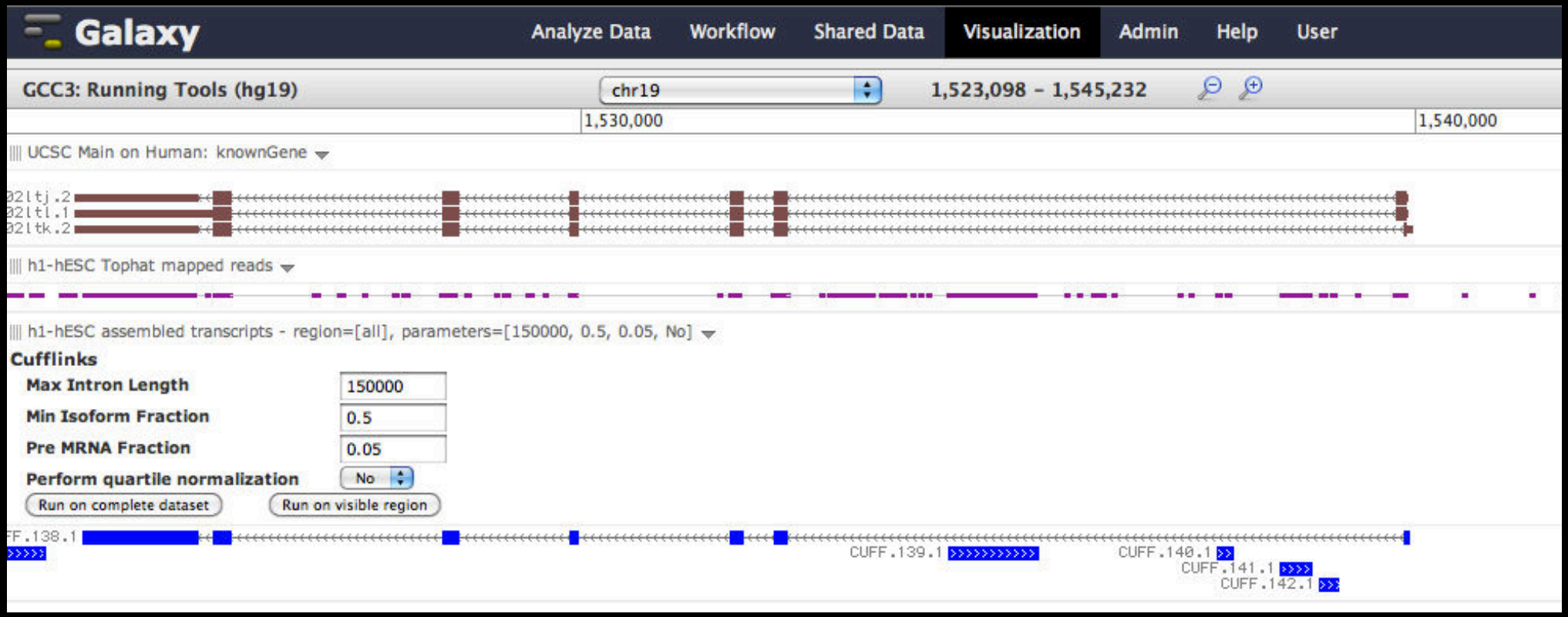
- for differentially expressed elements
- combined transcripts (e.g. for those differentially expressed between samples)

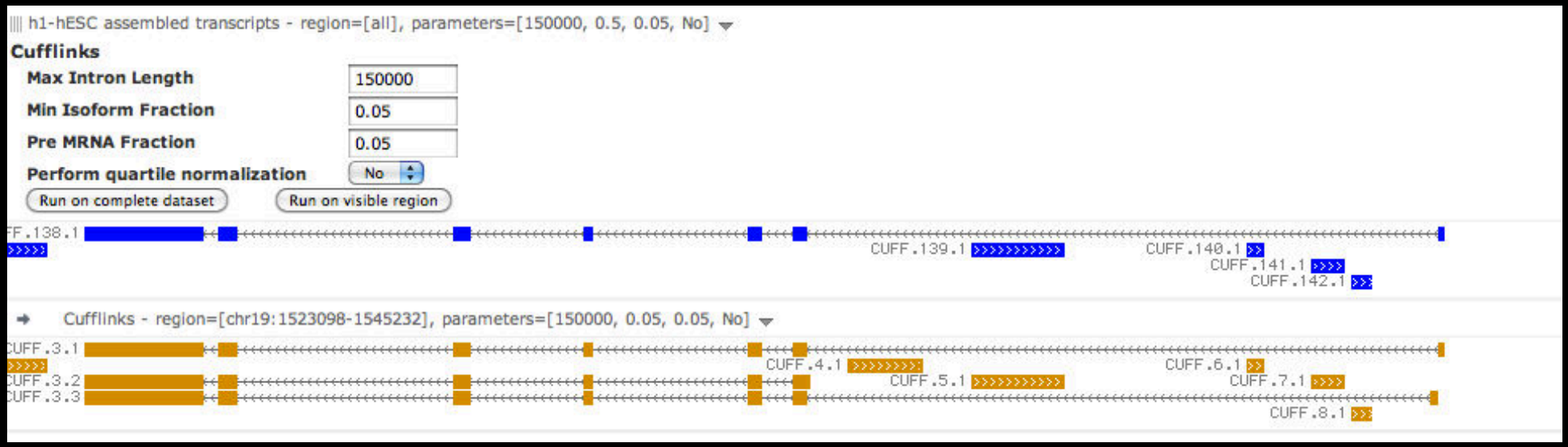


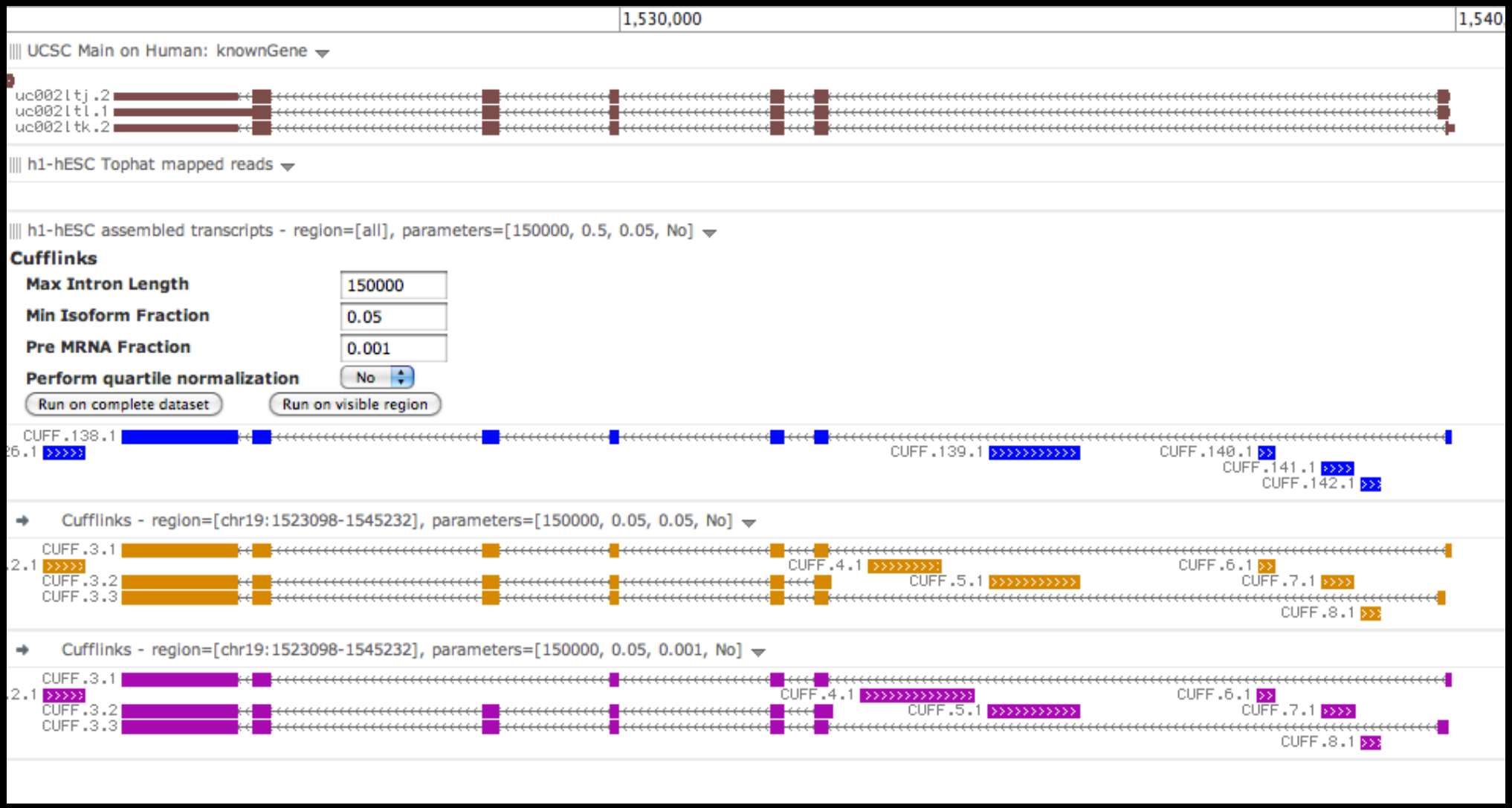
Extract transcript sequences and profile sequences for function



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

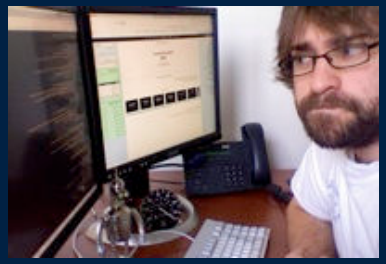
Jeremy Goecks

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

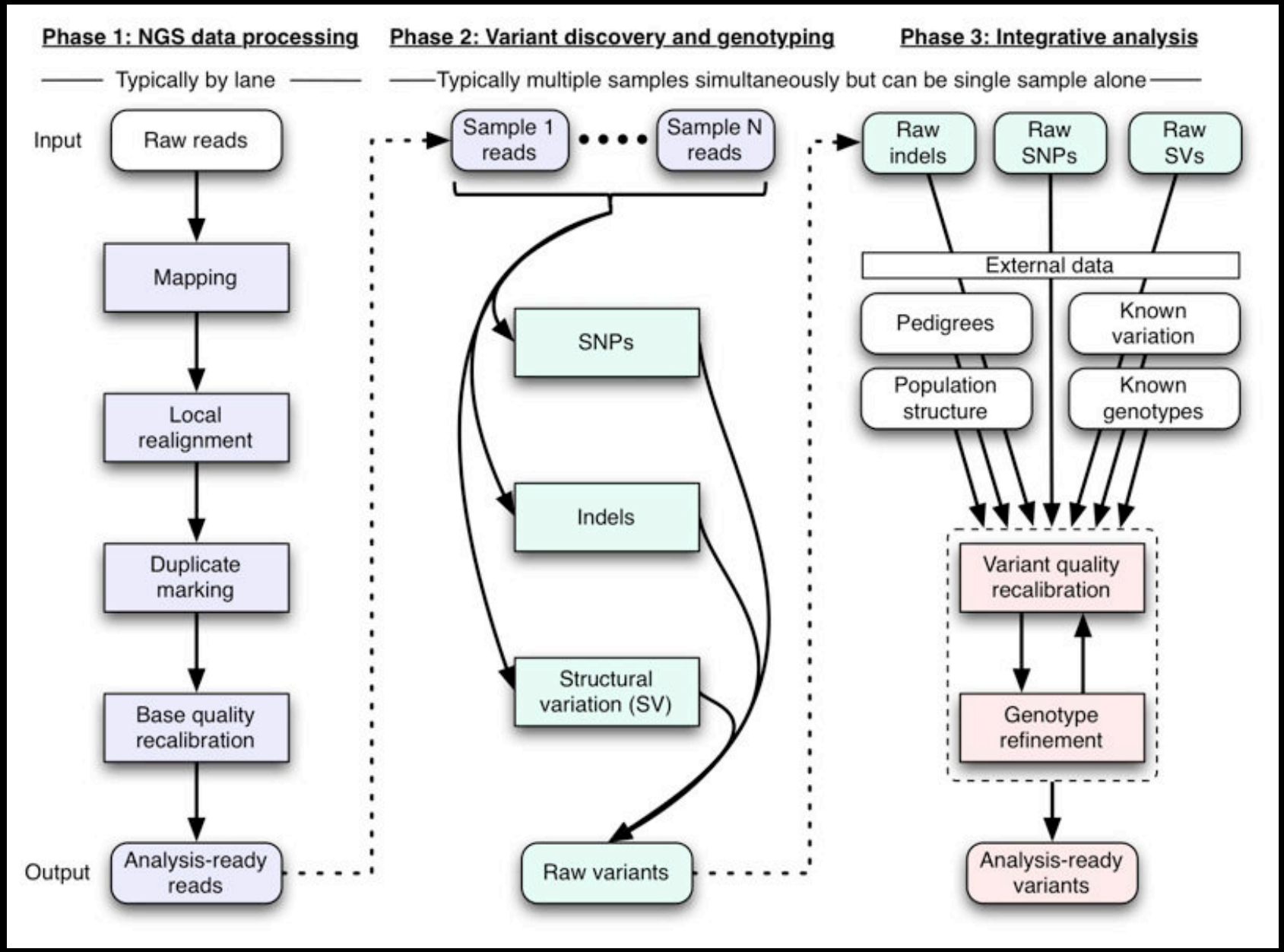
RNA-seq

- 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.

Running and Enhancing your own Galaxy

Daniel Blankenberg
The Galaxy Team
http://UseGalaxy.org

Overview

Where and How you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Exercise: Installing Galaxy and adding Tools

Overview

Where and How you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Exercise: Installing Galaxy and adding Tools

Galaxy main site (http://usegalaxy.org)

Public web site, anybody can use

~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month, every month is our busiest month ever...

Will continue to be maintained and enhanced, but with limits and quotas

Centralized solution cannot scale to meet data analysis demands

Overview

Where and How you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Exercise: Installing Galaxy and adding Tools

Local Galaxy instances (http://getgalaxy.org)

Galaxy is designed for local installation and customization

- Just download and run, completely self-contained
- Easily integrate new tools
- Easy to deploy and manage on nearly any (unix) system
- Run jobs on existing compute clusters

Especially useful for sensitive data

- can secure data and abide by regulations

Scale up on existing resources

Move intensive processing (tool execution) to other hosts

Frees up the application server to serve requests and manage jobs

Utilize existing resources

Supports any scheduler that supports DRMAA (most of them)

Running a Production Server

Use a real database server: PostgreSQL, MySQL

Run on compute cluster resources

External Authentication: LDAP, Kerberos, OpenID

Load balancing; proxy support

https://bitbucket.org/galaxy/galaxy-central/wiki/Config/ProductionServer

Lack IT knowledge or resources?

No problem, just use the Cloud

Overview

Where and How you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Exercise: Installing Galaxy and adding Tools

Cloud Computing

network accessible compute resources that can be rapidly acquired, configured, and released on demand

Infrastructure as a service

Compute resources provided and configured on demand (compute nodes, storage, network)

Public commercial: Amazon Web Services, Rackspace, ...
Build your own: Eucalyptus, Nimbus, OpenStack, ...

When to use the cloud?

Limited informatics expertise or infrastructure

Extended or particular resource needs

Cannot upload data to a shared resource

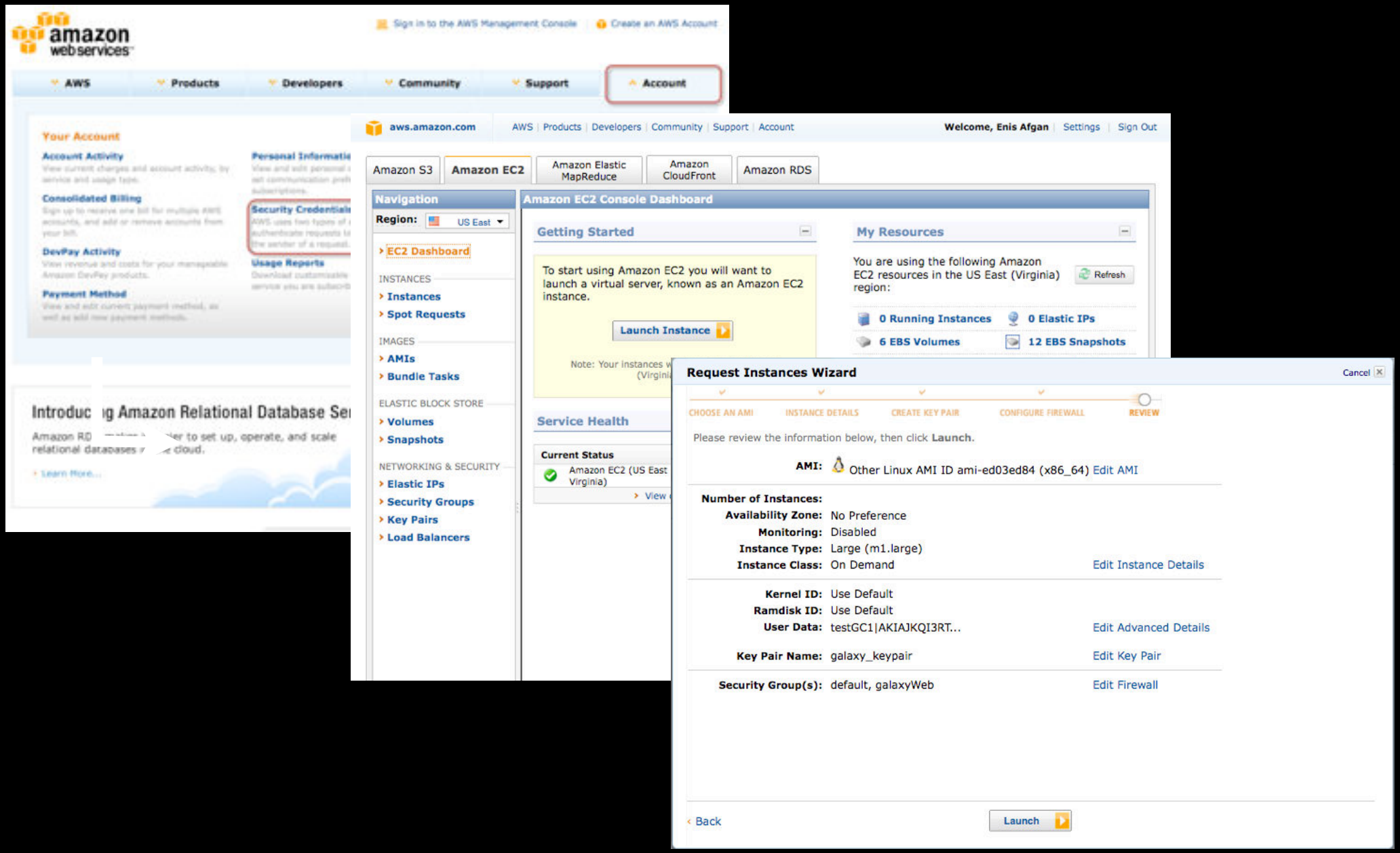
Need for customization

Have oscillating data volume

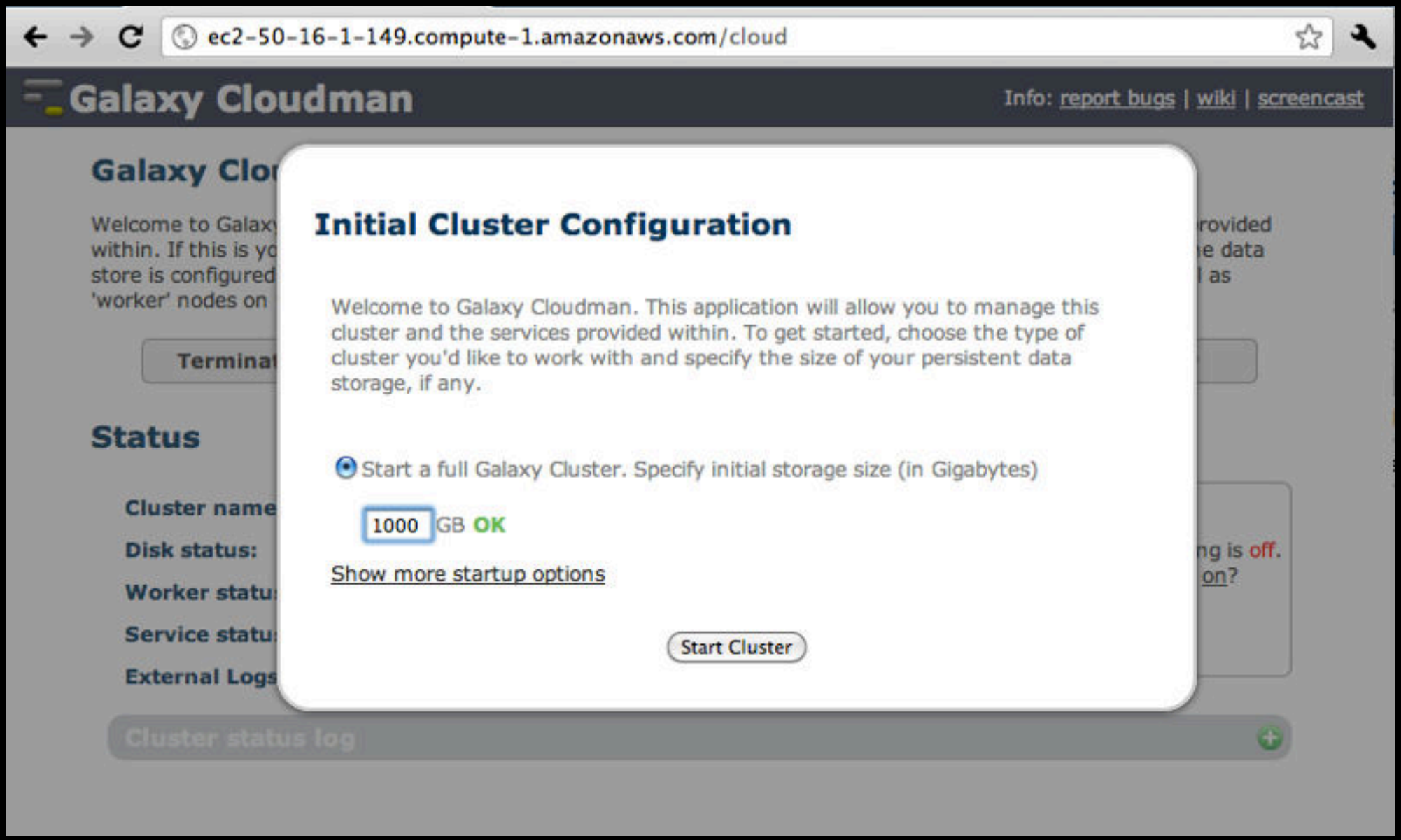
Deploying Galaxy on the AWS Cloud
http://usegalaxy.org/cloud

1. Open an AWS account (only once)
2. Use the AWS Management Console to start a master EC2 instance
3. Use the Galaxy CloudMan web interface on the master instance to manage the cluster

2. Start an EC2 Instance



3. Configure Your Cluster



Galaxy Cloud

http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud

Galaxy

Info: report bugs | wiki | screencasts

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.

Terminate cluster | Add nodes ▾ | Remove nodes | Access Galaxy

Status

Cluster name: ttt

Disk status: 0 / 0 (0%)

Worker status: **Idle**: 0 **Available**: 0 **Requested**: 0

Service status: Applications Data

Pending
Starting
Ready
Error

Cluster status log

Galaxy

http://ec2-75-101-213-19.compute-1.amazonaws.com/

Google

Galaxy

Analyze Data | Workflow | Data Libraries | Help | User

Tools

Get Data
Text Manipulation
Filter and Sort
Join, Subtract and Group
Operate on Genomic Intervals
Graph/Display Data

NGS TOOLBOX BETA

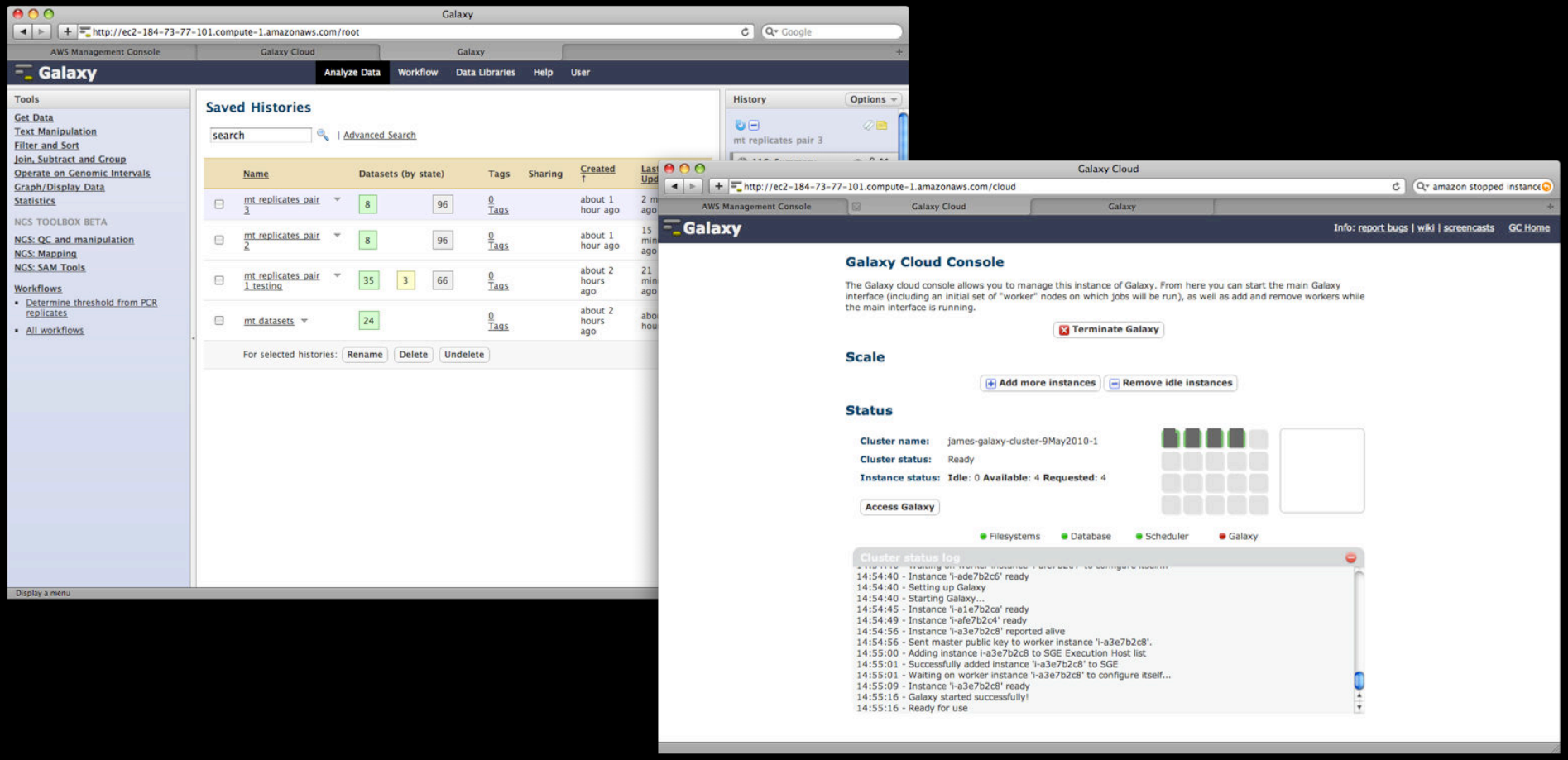
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools

Welcome to Galaxy on the Cloud

History

Options

Your history is empty. Click 'Get Data' on the left pane to start



Can use like any other Galaxy instance, with additional compute nodes acquired and released (*automatically*) in response to usage

Galaxy Cloud

http://ec2-184-73-135-47.compute-1.amazonaws.com/cloud/

AWS Management Console | Galaxy Cloud

Galaxy Cloudman

Info: report bugs | wiki | screencast

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.

Terminate cluster | Add nodes ▾ | Remove nodes | Access Galaxy

Status

Cluster name: james-cm-31march

Disk status: 181M / 100G (1%)

Worker status: **Idle**: 0 **Available**: 0 **Requested**: 0

Service status: Applications ● Data ●

External Logs: Galaxy Log

Autoscaling is off. Turn on?

Cluster status log

Galaxy Cloud

http://ec2-184-73-135-47.compute-1.amazonaws.com/cloud/

AWS Management Console | Galaxy Cloud

Galaxy Cloudman

Info: report bugs | wiki | screencast

Galaxy Cloudman Console

Welcome to Galaxy Cloudn loudMan. Your previous data store has been recon d and remove 'worker' nodes for running jobs.

Terminate clust | Access Galaxy

Status

Cluster name: jame

Disk status: 181

Worker status: Idle

Service status: Appli

External Logs: Gala

Cluster status log

Autoscaling is off. Turn on?

Currently shared instances

Share-an-instance

This form allows you to share this cluster instance, at its current state, with others. You can make the instance public or share it with specific users by providing their account information below.
You may also share the instance with yourself by specifying your own credentials, which will have the effect of saving the instance at its current state.

While setting up an instance to be shared, all currently running cluster services will be stopped. Then, a snapshot of your data volume and a folder in your cluster's bucket will be created (under 'shared/[current date and time]); this folder will contain your cluster's current configuration. The created snapshot and the folder will be given READ permissions to the users you choose (or make it public). This will enable those users to instantiate their own instances of the given cluster instance. This implies that you will only be paying for the created snapshot while users deriving a cluster from yours will incur costs for running the actual cluster. After the sharing process is complete, services on your cluster will automatically resume.

Public Shared

Share-an-instance

Display a menu

Automation

Cloud instances include all tools available in main Galaxy and more

Tool installation and configuration, image creation, etc, all completely automated and extensible

Same automation approach can be used for configuring tool dependencies for a local Galaxy

VM image with tools (not data) also available, currently at http://usegalaxy.org/vm

Overview

Where and How you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Exercise: Installing Galaxy and adding Tools

The Problem

You have written a Python script to analyze genomic data and you want to share it with command-line averse colleagues

The Galaxy Solution

Solution: Integrate the script as a new Tool into your own Galaxy server

Steps:

- Obtain and install Galaxy source code (GetGalaxy.org)
- Write an XML file describing the inputs and outputs and how to execute the script
- Instruct Galaxy to load the tool

Adding your Own

Write or download a command-line executable

Determine number and kind of

- Input and Output Datasets
- Input Parameters

Construct a descriptive tool configuration XML file

- Write a wrapper script, only if required

Tool Configuration

Tool Action - Default tool action should be adequate (Upload tool uses custom tool action)

Tool Command

Inputs

- Action - Used by datasource tools
- Parameters

Outputs

Help

Tests

A Basic Tool

```
<tool id="fa_gc_content_1" name="Compute GC content">
  <description>for each sequence in a file</description>
  <command interpreter="perl">toolExample.pl $input $output</command>
  <inputs>
    <param format="fasta" name="input" type="data" label="Source file"
  </inputs>
  <outputs>
    <data format="tabular" name="output" />
  </outputs>

  <tests>
    <test>
      <param name="input" value="fa_gc_content_input.fa"/>
      <output name="out_file1" file="fa_gc_content_output.txt"/>
    </test>
  </tests>

  <help>
This tool computes GC content from a FASTA file.
  </help>

</tool>
```

Compute GC content

Source file:

1: Uploaded FASTA File

Execute

This tool computes GC content from a FASTA file.

Tools

Get Data

MyTools

- Compute GC content for each sequence in a file

Send Data

```
<section name="MyTools" id="mTools">
    <tool file="myTools/toolExample.xml" />
 </section>
```

tool_conf.xml

Cluster

Cluster intervals of: 6: UCSC Main on Human: knownGene

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

TIP: If your query does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns

---

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be considered "clustered". **Negative** values for distance are allowed, and are useful for clustering intervals that overlap.
- **Minimum intervals per cluster** allow a threshold to be set on the minimum number of intervals to be considered a cluster. Any area with less than this minimum will not be included in the ouput.
- **Merge clusters into single intervals** outputs intervals that span the entire cluster.
- **Find cluster intervals; preserve comments and order** filters out non-cluster intervals while maintaining the original ordering and comments in the file.
- **Find cluster intervals; output grouped by clusters** filters out non-cluster intervals, but outputs the cluster intervals so that they are grouped together. Comments and original ordering in the file are lost.

Example

cluster.xml

```xml
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python2.4">
    gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
                    -d $distance -m $minregions -o $returntype
  </command>
  <inputs>
    <param format="interval" name="input1" type="data">
      <label>Cluster intervals of</label>
    </param>
    <param name="distance" size="5" type="integer" value="1" help="(bp
      <label>max distance between intervals</label>
    </param>
    <param name="minregions" size="5" type="integer" value="2">
      <label>min number of intervals per cluster</label>
    </param>
    <param name="returntype" type="select" label="Return type">
      <option value="1">Merge clusters into single intervals</option>
      <option value="2">Find cluster intervals; preserve comments and
      <option value="3">Find cluster intervals; output grouped by clus
      <option value="4">Find the smallest interval in each cluster</op
      <option value="5">Find the largest interval in each cluster</opt
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does not appear in the pulldown menu -> it is n

-----

**Screencasts!**

See Galaxy Interval Operation Screencasts_ (right click to open this l

.. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc

-----

**Syntax**

- **Maximum distance** is greatest distance in base pairs allowed betw
- **Minimum intervals per cluster** allow a threshold to be set on the
- **Merge clusters into single intervals** outputs intervals that span
- **Find cluster intervals; preserve comments and order** filters out
- **Find cluster intervals; output grouped by clusters** filters out n
```

Line: 87 Column: 8 XML Soft Tabs: 2

Input Parameter types

Basic

- Text
- Integer
- Float
- Select
  - Static
  - Dynamic
- Boolean
- Genome build
- Data column
- Data
- Hidden
- Base URL
- File
- Drill down
- Grouping
  - Conditional
  - Repeat
- Config Files

Datasets and Datatypes

All datasets are associated with a Datatype

- File format
- Type of Data: genomic intervals, sequence, alignment
- Hierarchical structure useful for inputs
- Automatic conversion possible
- Metadata

datatypes_conf.xml and lib/galaxy/datatypes

https://bitbucket.org/galaxy/galaxy-central/wiki/AddingDatatypes

Adding your Own Display Application

Define An XML configuration which describes how and where to present the data to the External Web Application

- Static
- Dynamic - display options can be loaded from a file

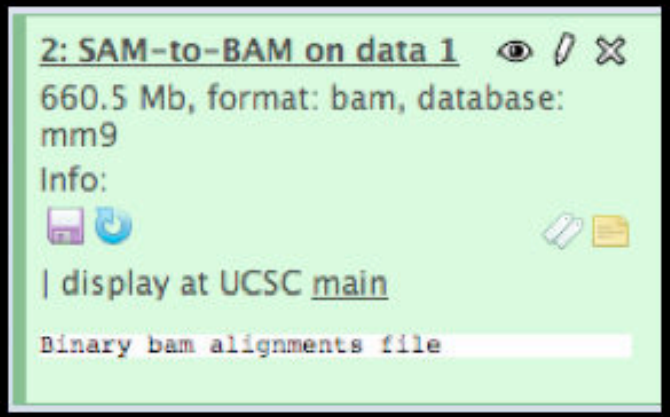
Inform Galaxy about the new display by adding to the appropriate datatype in datatypes_conf.xml

https://bitbucket.org/galaxy/galaxy-central/wiki/ExternalDisplayApplications/Tutorial

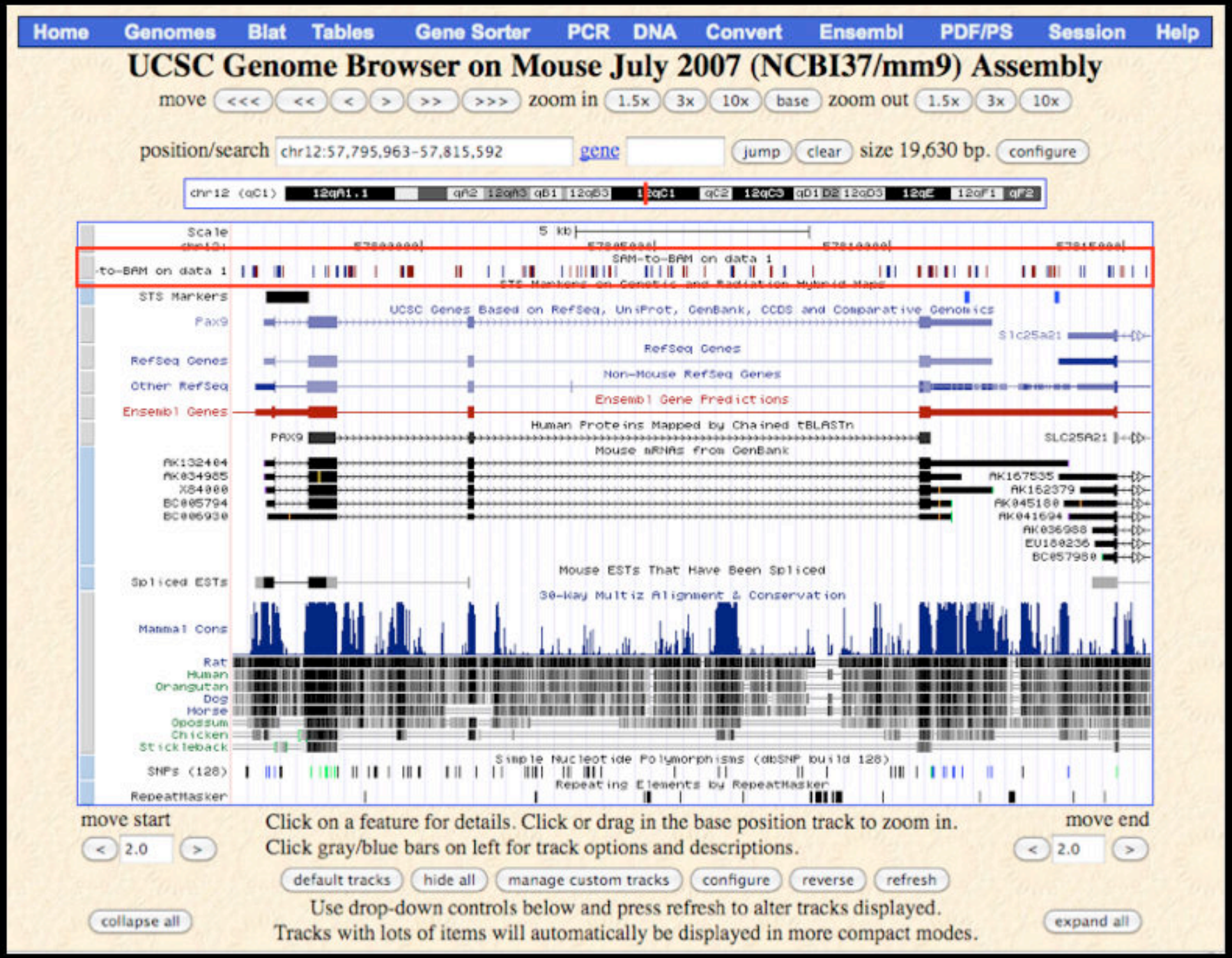
Static External Display Application

```
<display id="ucsc_bam" version="1.0.0" name="display at UCSC">
    <link id="main" name="main">
        <url>http://genome.ucsc.edu/cgi-bin/hgTracks?db=${qp($bam_file.dbkey)}&amp;hgt.customText=${qp($track.url)}</url>
        <param type="data" name="bam_file" url="galaxy.bam" strip_https="True" />
        <param type="data" name="bai_file" url="galaxy.bam.bai" metadata="bam_index" strip_https="True" />
        <param type="template" name="track" viewable="True" strip_https="True">
            track type=bam name="${bam_file.name}" bigDataUrl=${bam_file.url} db=${bam_file.dbkey}
        </param>
    </link>
</display>
```

```
<datatype extension="bam" type="galaxy.datatypes.binary:Bam"
    mimetype="application/octet-stream" display_in_upload="true">
    <display file="ucsc/bam.xml" />
</datatype>
```



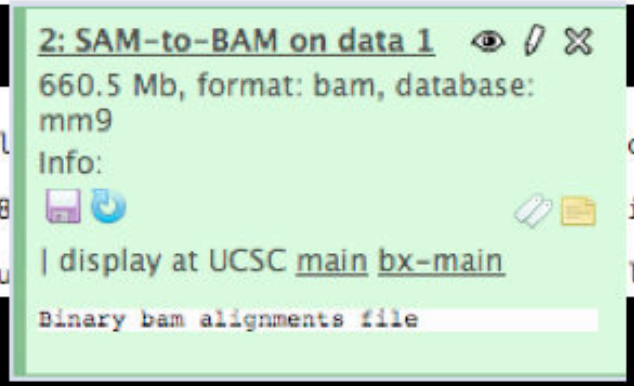
BAM at UCSC


Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl PDF/PS Session Help
UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr12:57,795,963-57,815,592 gene jump clear size 19,630 bp. configure
SAM-to-BAM on data 1
STS Markers
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Non-Mouse RefSeq Genes
Ensembl Gene Predictions
Human Proteins Mapped by Chained tBLASTn
Mouse mRNAs from GenBank
Mouse ESTs That Have Been Spliced
30-Way Multiz Alignment & Conservation
Mammal Cons
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 128)
SNPs (128)
Repeating Elements by RepeatMasker
RepeatMasker
move start
move end
Click on a feature for details. Click or drag in the base position track to zoom in.
Click gray/blue bars on left for track options and descriptions.
default tracks hide all manage custom tracks configure reverse refresh
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
collapse all
expand all

Dynamic External Display Application

```
<display id="ucsc_bam" version="1.0.0" name="display at UCSC">
    <!-- Load links from file: one line to one link -->
    <dynamic_links from_file="tool-data/shared/ucsc/ucsc_build_sites.txt" skip_startswith="#" id="0" name="0">

        <!-- Define parameters by column from file, allow splitting on builds -->
        <dynamic_param name="site_id" value="0"/>
        <dynamic_param name="ucsc_link" value="1"/>
        <dynamic_param name="builds" value="2" split="True" separator="," />

        <!-- Filter out some of the links based upon matching site_id to a Galaxy application configuration parameter and b
        <filter>${site_id in $APP.config.ucsc_display_sites}</filter>
        <filter>${dataset.dbkey in $builds}</filter>

        <!-- We define url and params as normal, but values defined in dynamic_param are available by specified name -->
        <url>${ucsc_link}db=${qp($bam_file.dbkey)}&amp;hgt.customText=${qp($track.url)}</url>
        <param type="data" name="bam_file" url="galaxy_${DATASET_HASH}.bam" strip_https="True" />
        <param type="data" name="bai_file" url="galaxy_${DATASET_HASH}.bam.bai" metadata="bam_index" strip_https="True" />
        <param type="template" name="track" viewable="True" strip_https="True">
            track type=bam name="${bam_file.name}" bigDataUrl=${bam_file.url} db=${bam_file.dbkey}
        </param>

    </dynamic_links>
</display>
```

```
#Harvested from http://genome.ucsc.edu/cgi-bin/das/dsn
main    http://genome.ucsc.edu/cgi-bin/hgTracks?    anoCar1,ce6,ce4,ce2,rn3,l
#Harvested from http://archaea.ucsc.edu/cgi-bin/das/dsn
archaea http://archaea.ucsc.edu/cgi-bin/hgTracks?   therSibi1,symbTher_IAM148
#Harvested from http://main.genome-browser.bx.psu.edu/cgi-bin/das/dsn
bx-main http://main.genome-browser.bx.psu.edu/cgi-bin/hgTracks? oviAri1,eriEu
```

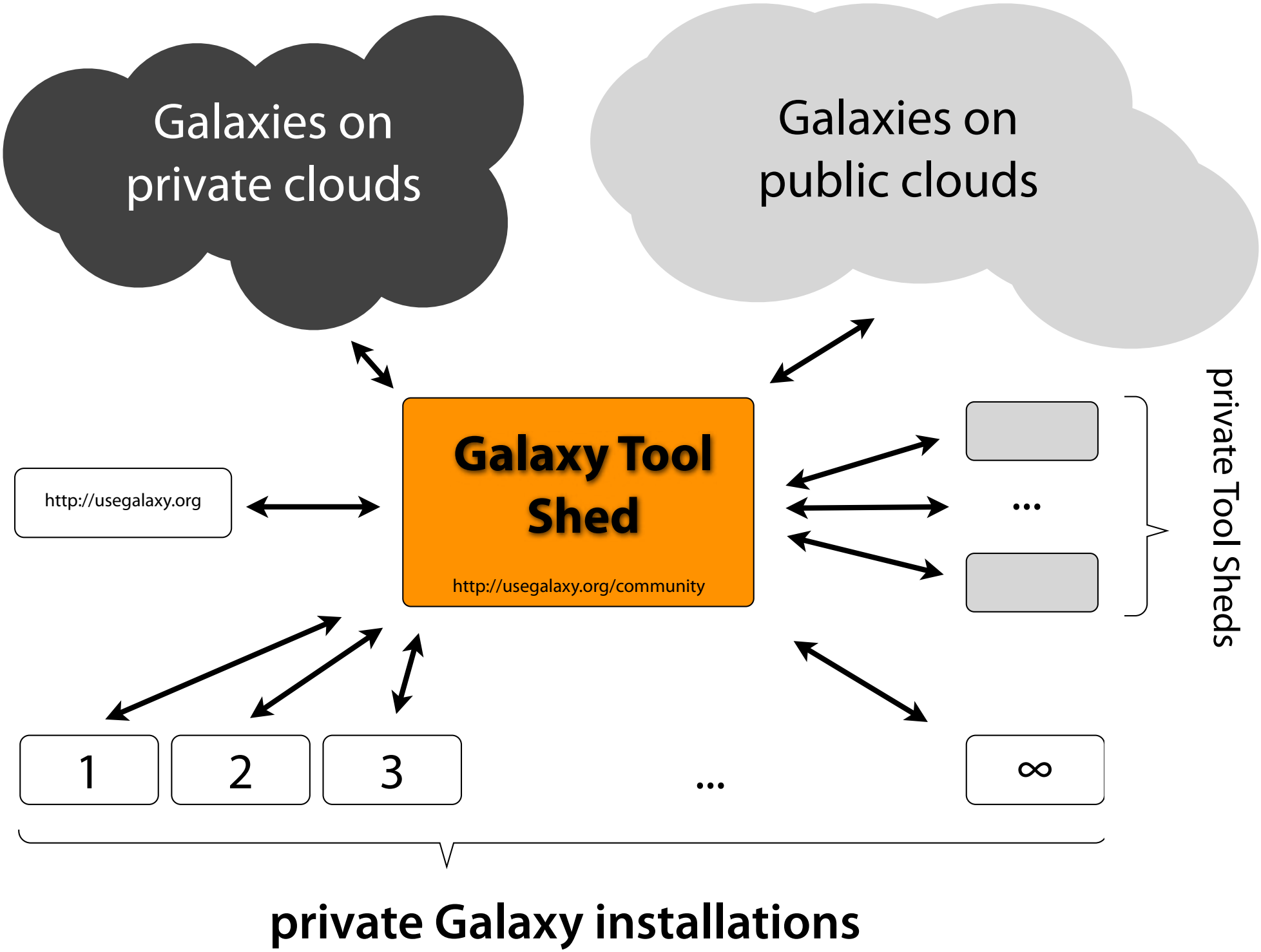


You added a tool, now what?

Share it with the community!

Galaxy Tool Shed

- Upload and Download contributed tools
- Rate and provide comments and feedback


Galaxies on private clouds
Galaxies on public clouds
Galaxy Tool Shed
http://usegalaxy.org/community
http://usegalaxy.org
...
private Tool Sheds
1
2
3
...
∞
private Galaxy installations

Get and Contribute Tools

Galaxy Tool Shed / (beta)

Tools Help User

Community

Tools

- Browse by category
- Browse all tools
- Login to upload

Categories

search Advanced Search

Name ↓	Description	Tools
Convert Formats	Tools for converting data formats	4
Data Source	Tools for retrieving data from external data sources	1
Fasta Manipulation	Tools for manipulating fasta data	5
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	5
Ontology Manipulation	Tools for manipulating ontologies	1
SAM	Tools for manipulating alignments in the SAM format	0
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	7
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	1
Statistics	Tools for generating statistics	1
Text Manipulation	Tools for manipulating data	3
Visualization	Tools for visualizing data	1

http://usegalaxy.org/community

Try it now:

http://usegalaxy.org

Develop and deploy:

http://getgalaxy.org

http://galaxyproject.org

Come do cool stuff, contact us at:

http://wiki.g2.bx.psu.edu/News/Galaxy is Hiring

Opportunities for collaboration, positions for postdocs, researchers, software engineers

Overview

Where and How you can use and build Galaxy

- public website
- local instance
- on the cloud
- tool shed/contributing tools

Exercise: Installing Galaxy and adding Tools

EMORY

Enis Afgan

Dannon Baker

Dave Clements

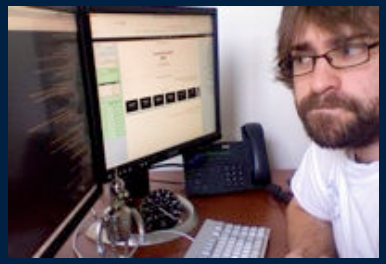
Jeremy Goecks

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

Download and Install

GetGalaxy.org

Requirements:

- Linux / Mac OS
- Python 2.5 - 2.7
- Mercurial (hg) for downloading (preferred), tar.gz available
- Internet connectivity for setup of dependencies

Follow directions: http://GetGalaxy.org

Adding a Tool

GetGalaxy.org/wiki

Requirements:

- Have or write a Command Line executable
- Determine inputs and outputs of tool
- Write XML description of tool
- Instruct Galaxy to load tool

Follow directions: http://wiki.g2.bx.psu.edu/Admin/Tools/Add Tool Tutorial

Deploying Galaxy on the AWS Cloud
http://usegalaxy.org/cloud

1. Open an AWS account (only once)
2. Use the AWS Management Console to start a master EC2 instance
3. Use the Galaxy CloudMan web interface on the master instance to manage the cluster

Second Half: Running Your Own Instance

You Need a Mac or Linux machine

If you have windows, you can use a virtual machine setup, such as virtualbox with Ubuntu

VirtualBox: https://www.virtualbox.org/

Ubuntu: http://www.ubuntu.com/

A preconfigured VM is available (less preferred for learning setup): http://usegalaxy.org/vm

WIFI: OICR Guest

Username: setup

password: oicrguest

The Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and transparent computational biomedical research

What is Galaxy?

GUI for genomics

- for complete analyses: analyze, visualize, share, publish

A free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

Open source software that makes integrating your own tools and data and customizing for your own site simple