

An Introduction to Galaxy

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

Where you can use and build Galaxy

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

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

Session Overview

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

Galaxy Analysis Workspace

The screenshot displays the Galaxy Analysis Workspace interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The left sidebar lists various tools and categories, including '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', and 'Workflows'.

The main panel shows the 'Map with Bowtie for Illumina' tool configuration. The tool is set to use a built-in index (mm9) and is configured for paired-end reads. The forward and reverse FASTQ files are both set to '1: E18 PE.1 Reads'. The maximum insert size for valid paired-end alignments is set to 1000. The upstream/downstream mate orientation is set to 'FR (for Illumina)'. The Bowtie settings are set to 'Commonly used'. The 'Suppress the header in the output SAM file' checkbox is checked. The 'Execute' button is visible at the bottom of the tool configuration panel.

The right sidebar shows the 'History' panel, listing the workflow steps in reverse chronological order:

- 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

Below the tool configuration, a section titled 'What it does' provides a brief description of the Bowtie tool: '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.'

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

[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](#)

RGENTICS
[SNP/WGA: Data; Filters](#)
[SNP/WGA: QC; LD; Plots](#)
[SNP/WGA: Statistical Models](#)

[Workflows](#)

Galaxy Analysis Workspace

The screenshot displays the Galaxy web interface. At the top, there's a navigation bar with links: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. Below this, the 'Map with Bowtie for Illumina' tool is selected. The tool's configuration panel includes fields for 'Select a reference genome from your history or use a built-in index?', 'Select a reference genome:' (set to 'mm9'), 'Maximum insert size for valid paired-end alignments (-X):' (set to 1000), and 'Bowtie settings to use:' (set to 'Commonly used'). There's also a checkbox for 'Suppress the header in the output SAM file:' which is checked. An 'Execute' button is at the bottom of the configuration panel. To the right, a 'History' panel shows a list of previous jobs, including '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', and '1: E18 PE.1 Reads'.

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

[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](#)

RGENTICS
[SNP/WGA: Data; Filters](#)
[SNP/WGA: QC; LD; Plots](#)
[SNP/WGA: Statistical Models](#)

[Workflows](#)

Galaxy Analysis Workspace

The screenshot displays the Galaxy web interface. At the top, there's a navigation bar with links: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. Below this, the 'Map with Bowtie for Illumina' tool is selected. The tool's configuration panel includes several input fields and checkboxes:

- Reference genome:** A dropdown menu showing 'mm9'. A note below states: 'Your genome of interest is not listed - contact Galaxy team'.
- Library mate-paired?:** A checkbox labeled 'paired-end' is checked.
- Forward FASTQ file:** A dropdown menu showing '1: E18 PE.1 Reads'. A note below states: 'Must have Sanger-scaled quality values with ASCII offset 33'.
- Reverse FASTQ file:** A dropdown menu showing '1: E18 PE.1 Reads'. A note below states: 'Must have Sanger-scaled quality values with ASCII offset 33'.
- Maximum insert size for valid paired-end alignments (-X):** A text input field containing '1000'.
- The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff):** A dropdown menu showing 'FR (for Illumina)'.
- Bowtie settings to use:** A dropdown menu showing 'Commonly used'. A note below states: '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:** A checkbox is checked. A note below states: 'Bowtie produces SAM with several lines of header information by default'.

At the bottom of the tool configuration, there is an 'Execute' button and a 'What it does' section. The 'What it does' section contains the following text: '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.'

On the right side of the interface, there is a 'History' panel. It shows a list of recent jobs, each with a name, a status icon, and a link to view the job. The jobs listed are:

- 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

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

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

Galaxy Analysis Workspace

The screenshot displays the Galaxy web interface. At the top, there's a navigation bar with tabs: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. Below this, the main content area shows the configuration for the 'Map with Bowtie for Illumina' tool. The tool's description states: 'If you select a reference genome from your history or use a built-in index?'. The configuration includes a dropdown for 'Select a reference genome:' set to 'mm9', a checkbox for 'Paired-end?' which is checked, and input fields for 'Forward FASTQ file:' and 'Reverse FASTQ file:' both set to '1: E18 PE.1 Reads'. Other settings include 'Maximum insert size for valid paired-end alignments (-X):' set to '1000' and 'Bowtie settings to use:' set to 'Commonly used'. A checkbox for 'Suppress the header in the output SAM file:' is also checked. An 'Execute' button is at the bottom of the configuration panel. To the right, a 'History' panel lists previous jobs, including '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', and '1: E18 PE.1 Reads'. Each history entry has icons for viewing, deleting, and refreshing.

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

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

Galaxy Analysis Workspace

The screenshot displays the Galaxy Analysis Workspace interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area shows the 'Map with Bowtie for Illumina' tool configuration. The tool is set to use a built-in index (mm9) and default options. The 'Forward FASTQ file' and 'Reverse FASTQ file' are both set to 'E18 PE.1 Reads'. The 'Maximum insert size for valid paired-end alignments (-X)' is set to 1000. The 'Bowtie settings to use' are set to 'Commonly used'. The 'Add the header in the output SAM file' checkbox is checked. The 'What it does' section explains that Bowtie is a short read aligner designed to be ultrafast and memory-efficient, developed by Ben Langmead and Cole Trapnell.

History panel (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
- 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

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

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

GENETICS

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

Workflows

Galaxy Analysis Workflows

Galaxy

Analyze Data Workflow Shared Data Visualization Help

Map with Bowtie for Illumina

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

Is a built-in index: ☐

Reads were indexed using default options

Select a reference genome:

mm9

Your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?: ☐

Forward FASTQ file:

1: 1000000000 reads

Reads were indexed using default options

Reverse FASTQ file:

1: 1000000000 reads

Reads were indexed using default options

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

1000

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

fr (forward) ☐

Which settings to use:

Commonly used ☐

For full control use Full control

Append the header in the output SAM file:

Yes ☐

Galaxy produces SAM with several lines of header information by default

What it does

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

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 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

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

What it does

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

Helping Users with Galaxy Analysis Workspace

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

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
- [FASTQ Masker](#) by quality score

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

RGENETICS

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

Where you can use and build Galaxy

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

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 (GCP) and other personal genome ▼		greg@bx.psu.edu	2010-02-15	2.1 Mb
☐ indels in AB1 ▼		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 MD6 ▼		greg@bx.psu.edu	2010-02-03	109.8 Kb
☐ indels in NB1 ▼		greg@bx.psu.edu	2010-02-03	519.1 Kb
☐ indels in TK1 ▼		greg@bx.psu.edu	2010-02-03	123.2 Kb
☐ novel SNPs in AB1 ▼		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 MD6 ▼		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:

<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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

Galaxy Workflows

The screenshot displays the Galaxy web interface at <http://main.g2.bx.psu.edu/>. The main content area shows a workflow execution page with a large table of genomic data. A yellow warning banner at the top of the table states: "This dataset is large and only the first megabyte is shown below. [Show all](#) | [Save](#)".

The table contains genomic data with columns for chromosome (chr10), position, and various metrics. The data is organized into rows, with the first row showing a position of 6882036 and a value of 107. The table is truncated, showing only the first megabyte of data.

The right sidebar displays the "History" section, listing various datasets and workflows. The "Current History" section shows a list of datasets, including "10: V", "26,74", "datab", and "Info". The "Workflow" section shows a list of workflows, including "9: Generate pileup on", "8: SAM-to-BAM on data", and "7: Map with Bowtie for".

The left sidebar contains a "Tools" section with a search bar and a list of tool categories, including "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", and "SNP/WGA: Statistical Models".

Galaxy Workflows

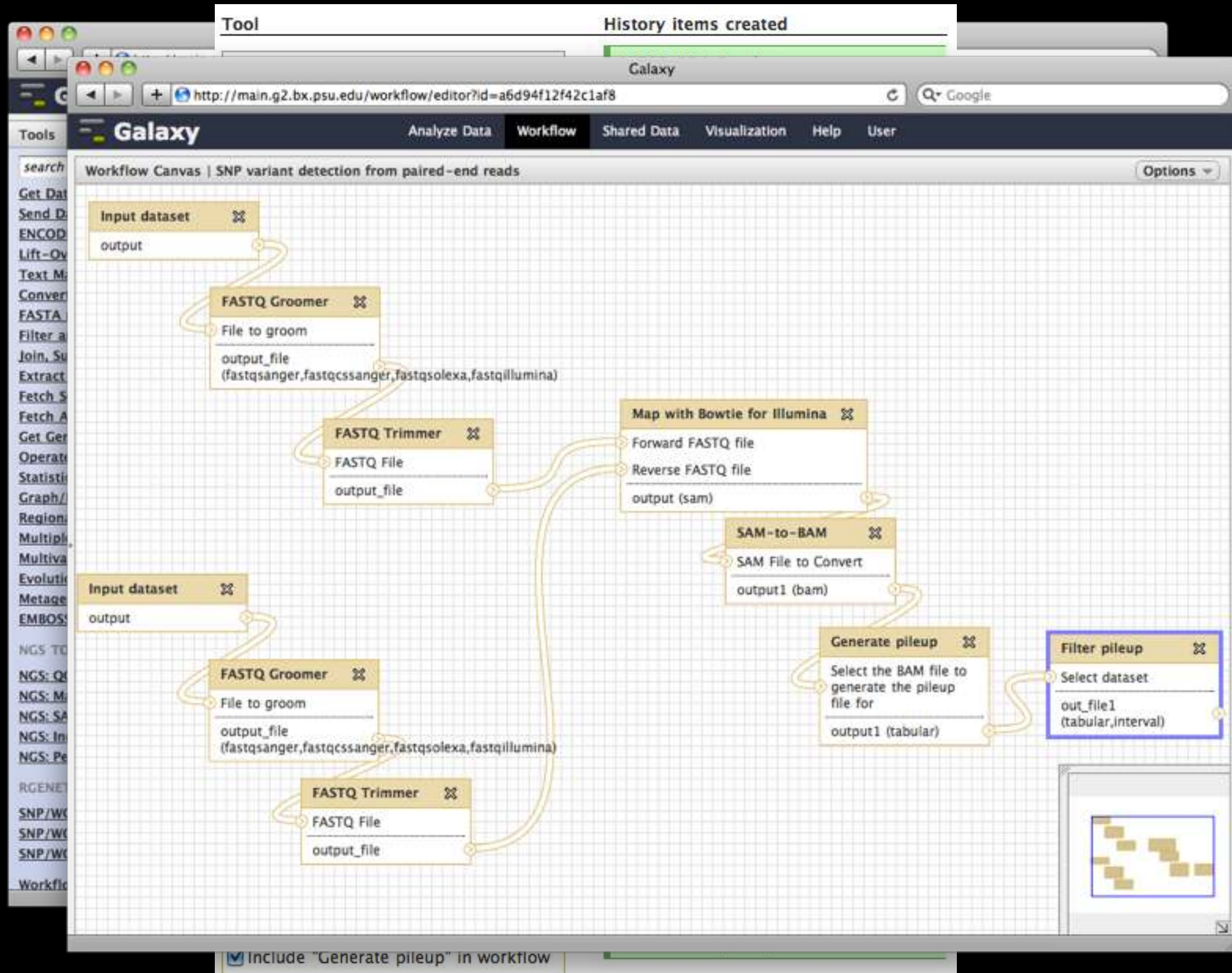
The screenshot displays the Galaxy web interface's workflow builder. The main area is divided into two columns: 'Tool' and 'History items created'.

Tool	History items created
Upload File <i>This tool cannot be used in workflows</i>	1: E18 PE.1 Reads ☒ Treat as input dataset
Upload File <i>This tool cannot be used in workflows</i>	2: E18 PE.2 Reads ☒ Treat as input dataset
FASTQ Groomer ☒ Include "FASTQ Groomer" in workflow	3: E18 PE.1 Reads Groomed
FASTQ Groomer ☒ Include "FASTQ Groomer" in workflow	4: E18 PE.2 Reads Groomed
FASTQ Trimmer ☒ Include "FASTQ Trimmer" in workflow	5: E18 PE.1 Reads Groomed, Trimmed
FASTQ Trimmer ☒ Include "FASTQ Trimmer" in workflow	6: E18 PE.2 Reads Groomed, Trimmed
Map with Bowtie for Illumina ☒ Include "Map with Bowtie for Illumina" in workflow	7: Map with Bowtie for Illumina on data 6 and data 5
SAM-to-BAM ☒ Include "SAM-to-BAM" in workflow	8: SAM-to-BAM on data 7
Generate pileup ☒ Include "Generate pileup" in workflow	9: Generate pileup on data 8

The left sidebar shows the 'Tools' menu with categories like 'Get Data', 'Text 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', and 'Workflows'.

The right sidebar shows the 'History Lists' menu with options like 'Saved Histories', 'Histories Shared with Me', 'Current History', 'Create New', 'Clone', 'Share or Publish', 'Extract Workflow', 'Dataset Security', 'Show Deleted Datasets', 'Show Hidden Datasets', 'Show structure', and 'Delete'.

Galaxy Workflows



Galaxy Workflows

Tool History items created

Galaxy

http://main.g2.bx.psu.edu/workflow/editor?id=a6d94f12f42c1af8

Tools

search

Get Data
Send Data
ENCODING
Lift-Over
Text Manipulation
Conversion
FASTA
Filter and Sort
Join, Split
Extract
Fetch SRA
Fetch Assembly
Get Genomes
Operate on Genomes
Statistical Analysis
Graph/Network
Regions
Multiple Sequence Alignment
Multivariate Analysis
Evolutionary Analysis
Metagenomics
EMBOSS

NGS TOOLS
NGS: QC
NGS: Mapping
NGS: Assembly
NGS: Variant Calling
NGS: Phylogenetics
RGENE
SNP/Variant
SNP/Variant
SNP/Variant
Workflow

Workflow Canvas | SNP variant detection from paired-end reads

Input dataset
output

FASTQ Groomer
File to groom
output_file
(fastqsanger,fastqcassanger,fastqsolexa,fastqillumina)

FASTQ Trimmer
FASTQ File
output_file

Map with Bowtie for Illumina
Forward FASTQ file
Reverse FASTQ file
output (sam)

SAM-to-BAM
SAM File to Convert
output1 (bam)

Generate Pileup
Select
genera
file for
output

Tool: SAM-to-BAM

Choose the source for the reference list
Locally cached

SAM File to Convert
Data input 'input1' (sam)

Edit Step Actions

Assign Columns
output1 Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:
Convert Bowtie SAM output to BAM format so that pileup can be run.

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Include "Generate pileup" in workflow

Galaxy Workflows

The screenshot displays the Galaxy web interface. In the background, a workflow canvas titled "Workflow Canvas | SNP variant detection" is visible, showing a sequence of tools: "Input dataset" (output), "FASTQ Groomer" (output_file: fastqsanger, fastq), and "FASTQ Trimmer" (output_file). The "FASTQ Trimmer" tool is currently selected, and its configuration panel is open in the foreground.

Tool: SAM-to-BAM

Choose the source for the reference list
Locally cached

SAM File to Convert
Data input 'input1' (sam)

Edit Step Actions

Assign Columns
output1 Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:
Convert Bowtie SAM output to BAM format so that pileup can be run.

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Tool: FASTQ Trimmer

Name:
SNP identification within annotated genes from NGS PE Data

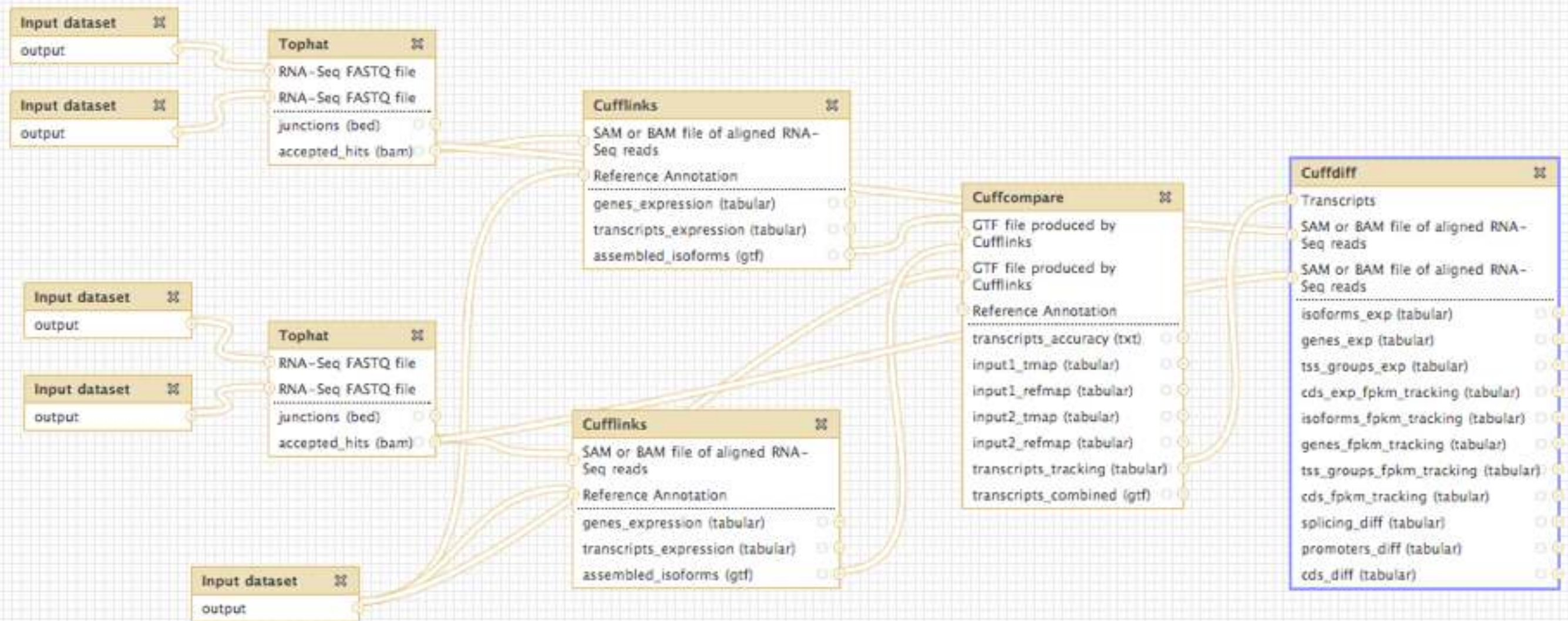
Tags:
snp x ngs x pileup x bowtie x

Apply tags to make it easy to search for and find items with the same tag.

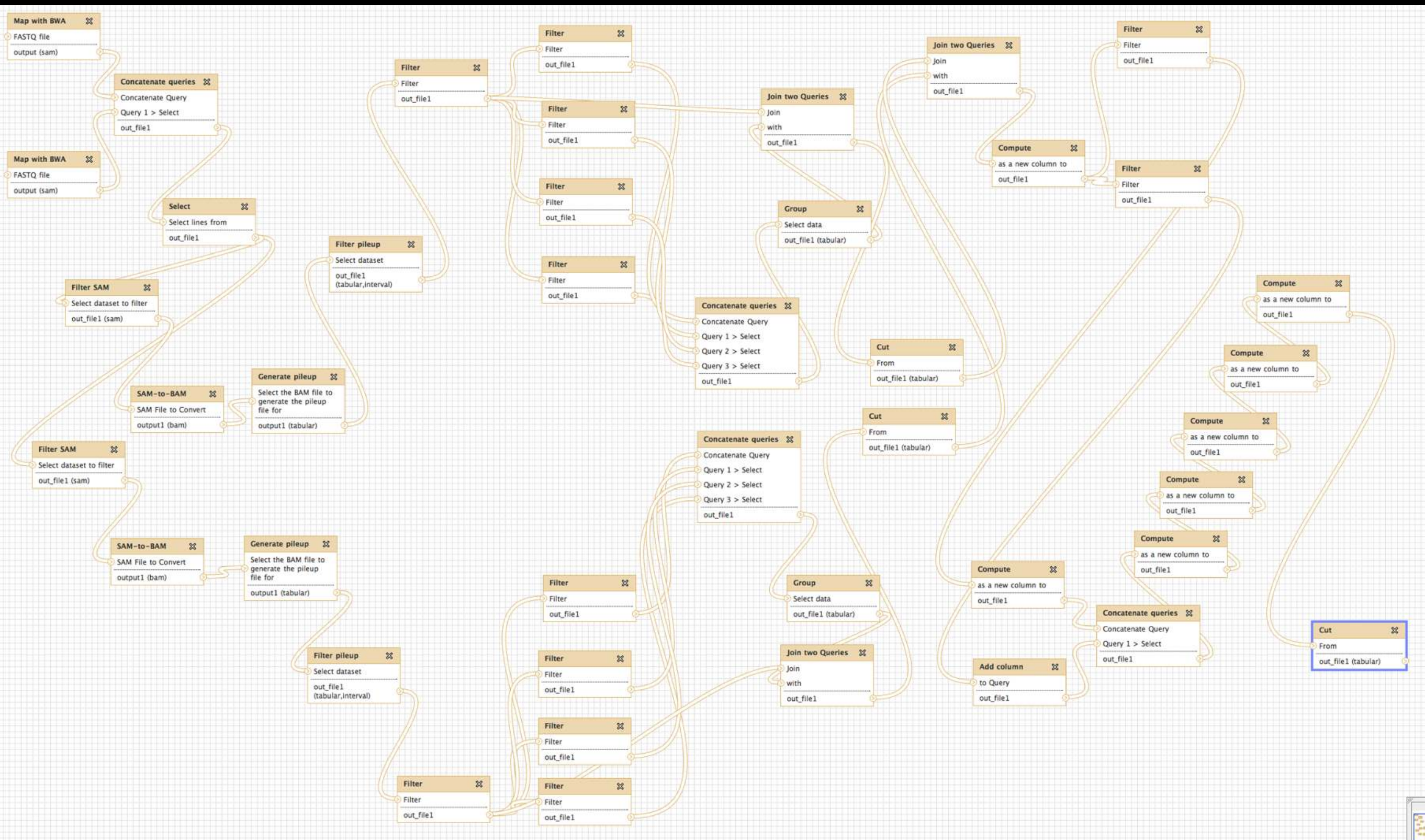
Annotation / Notes:
Identify variants in annotated genes from NGS paired-end data.

Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

☒ Include "Generate pileup" in workflow



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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

Visualize

Send data results to external genome browsers

Trackster: Galaxy's genome browser

External Genome Browsers

UCSC

Ensembl

GBrowse

IGV

UCSC Genome Browser on Mouse July 2007 (NCBI37)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out

position/search chr12:57,795,963-57,815,592

gene

jump

clear

size

chr12 (qC1) 12qA1.1 qA2 12qA3 qB1 12qB3 12qC1 qC2 12qC3 qD1 D2 12qD3 12qE 12qF1 qF2

14: Tag Counts (bigWig)

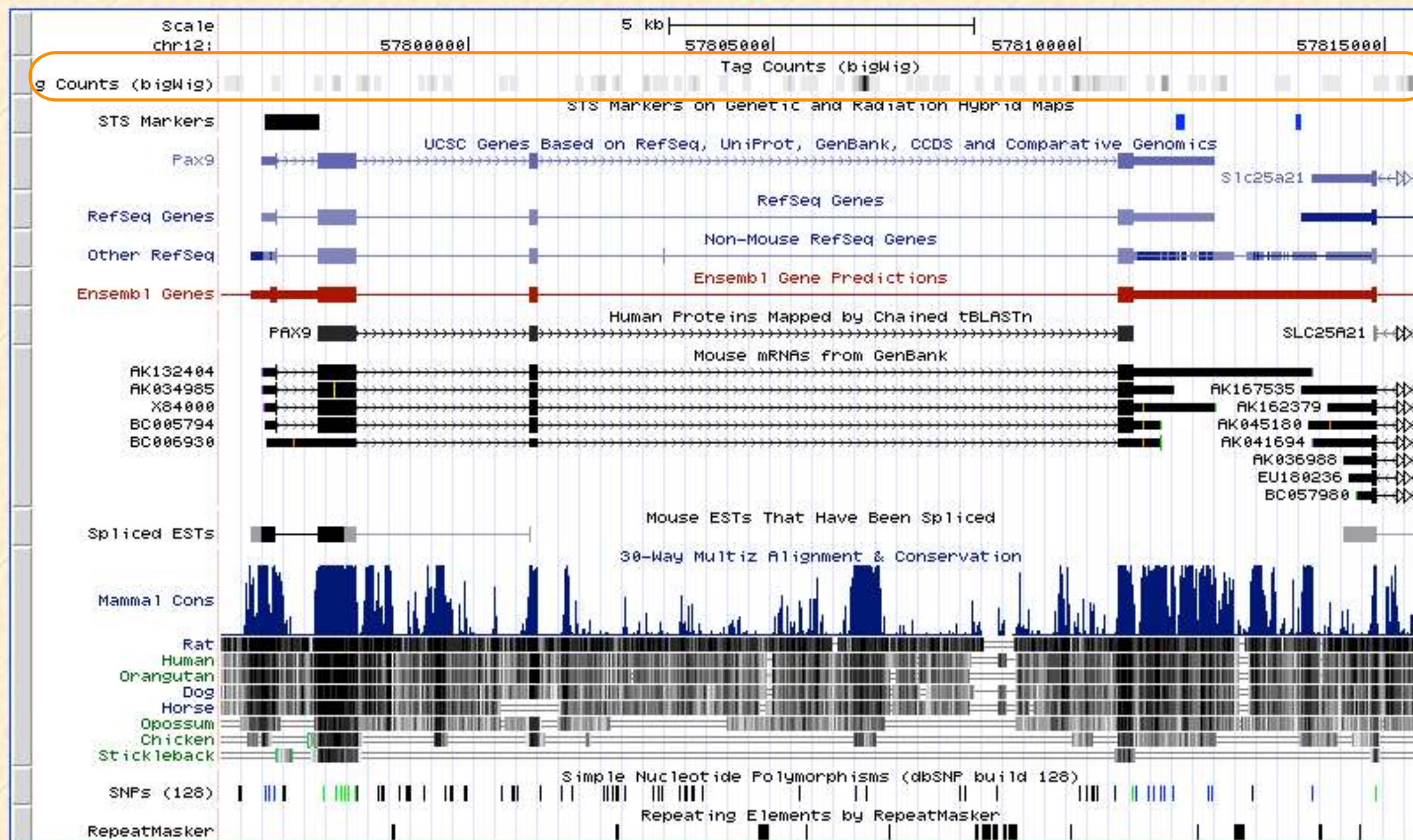
2.4 Gb, format: bigwig, database: mm9

Info:



display at UCSC main

Binary UCSC BigWig file



Integrative Genomics Viewer (IGV)

1: Sample data

1.2 Gb
format: bam, database: mm9
Info: uploaded bam file



display at UCSC [main test](#)
display at Ensembl [Current](#)
display with IGV [web local](#)

Binary bam alignments file



The application "IGV 1.5" from "www.broadinstitute.org" is requesting access to your computer.

The digital signature could not be verified.

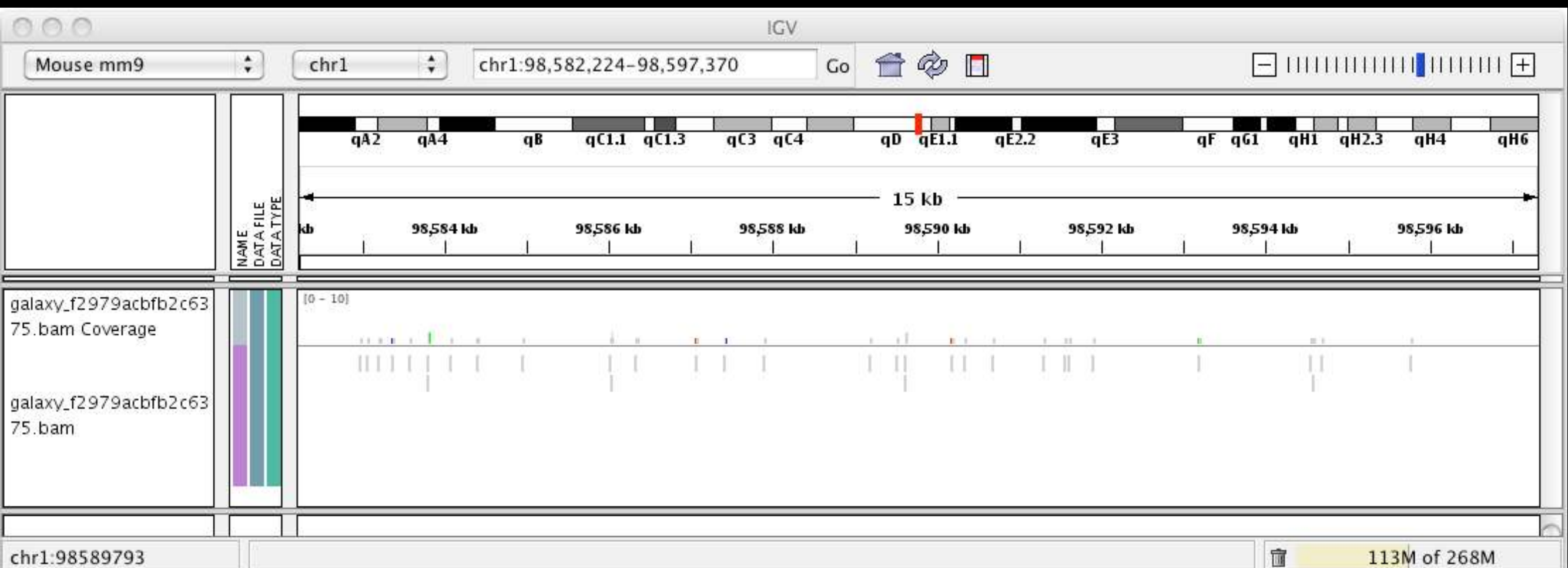
☐ Allow all applications from "www.broadinstitute.org" with this signature



Show Details...

Deny

Allow

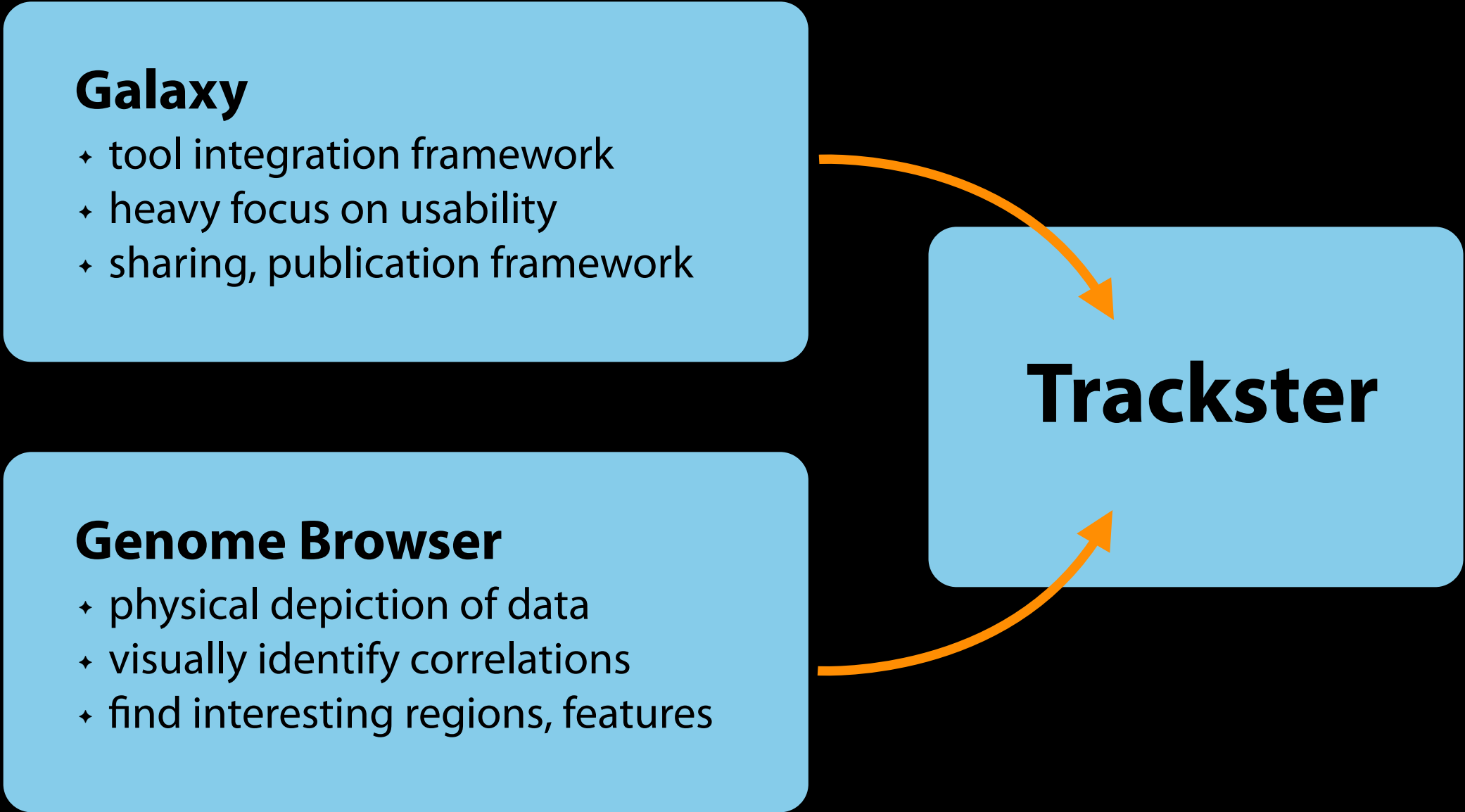


Galaxy

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

Genome Browser

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



```
graph LR; Galaxy[Galaxy] --> Trackster[Trackster]; GB[Genome Browser] --> Trackster;
```

Trackster

Trackster

View your data from within Galaxy

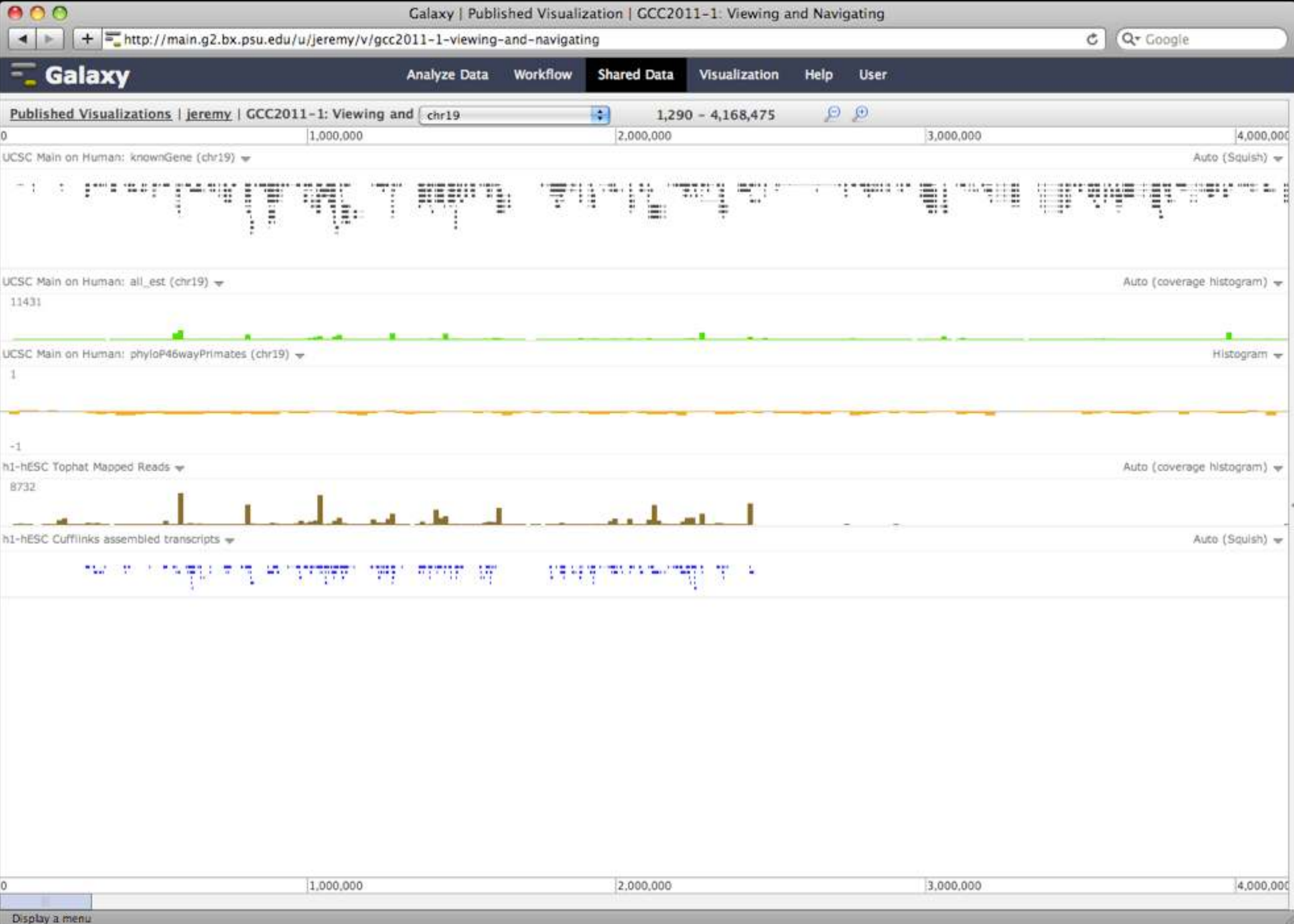
- ✦ No data transfers to external site
- ✦ Use it locally, even without internet access

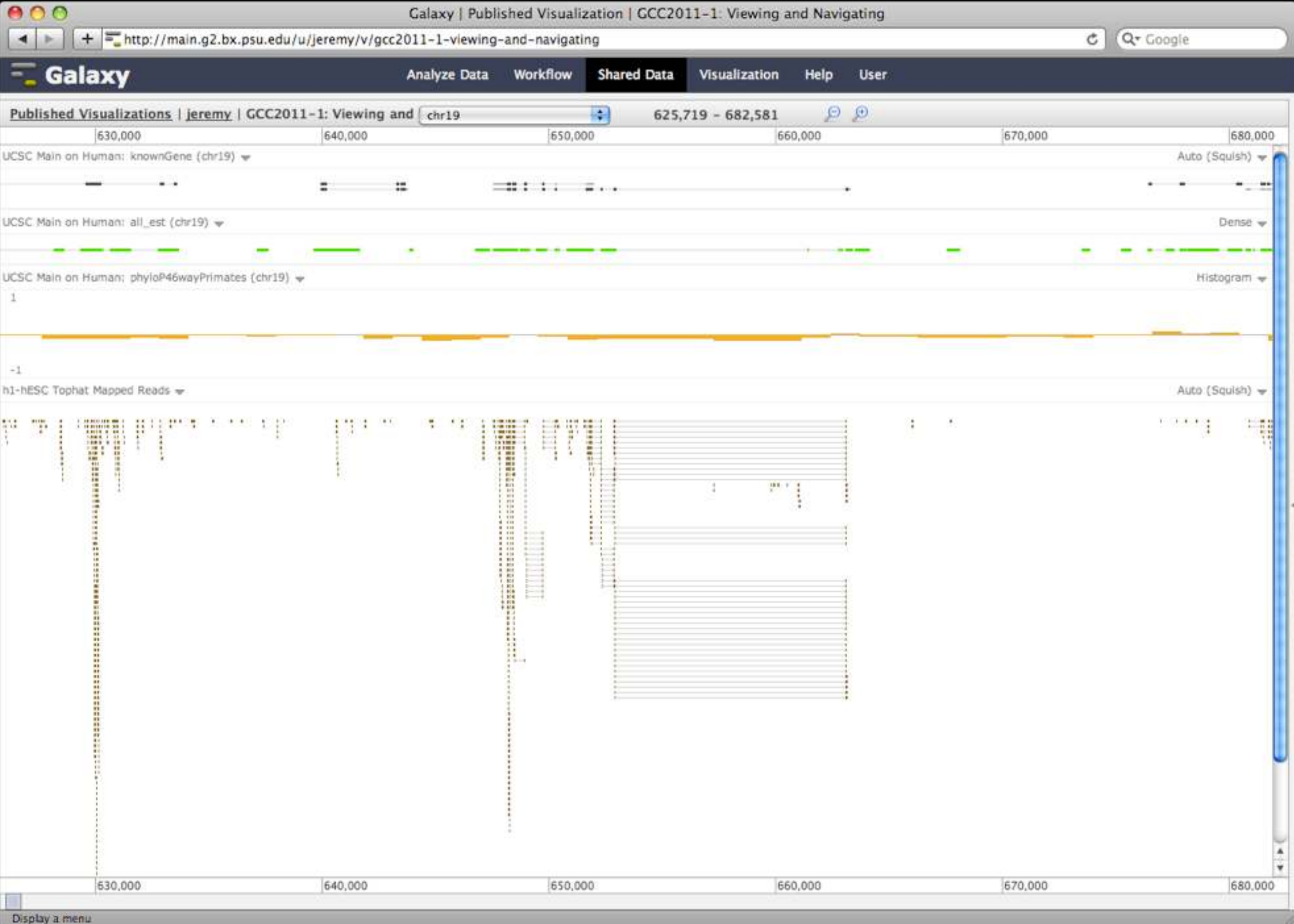
Supports common filetypes

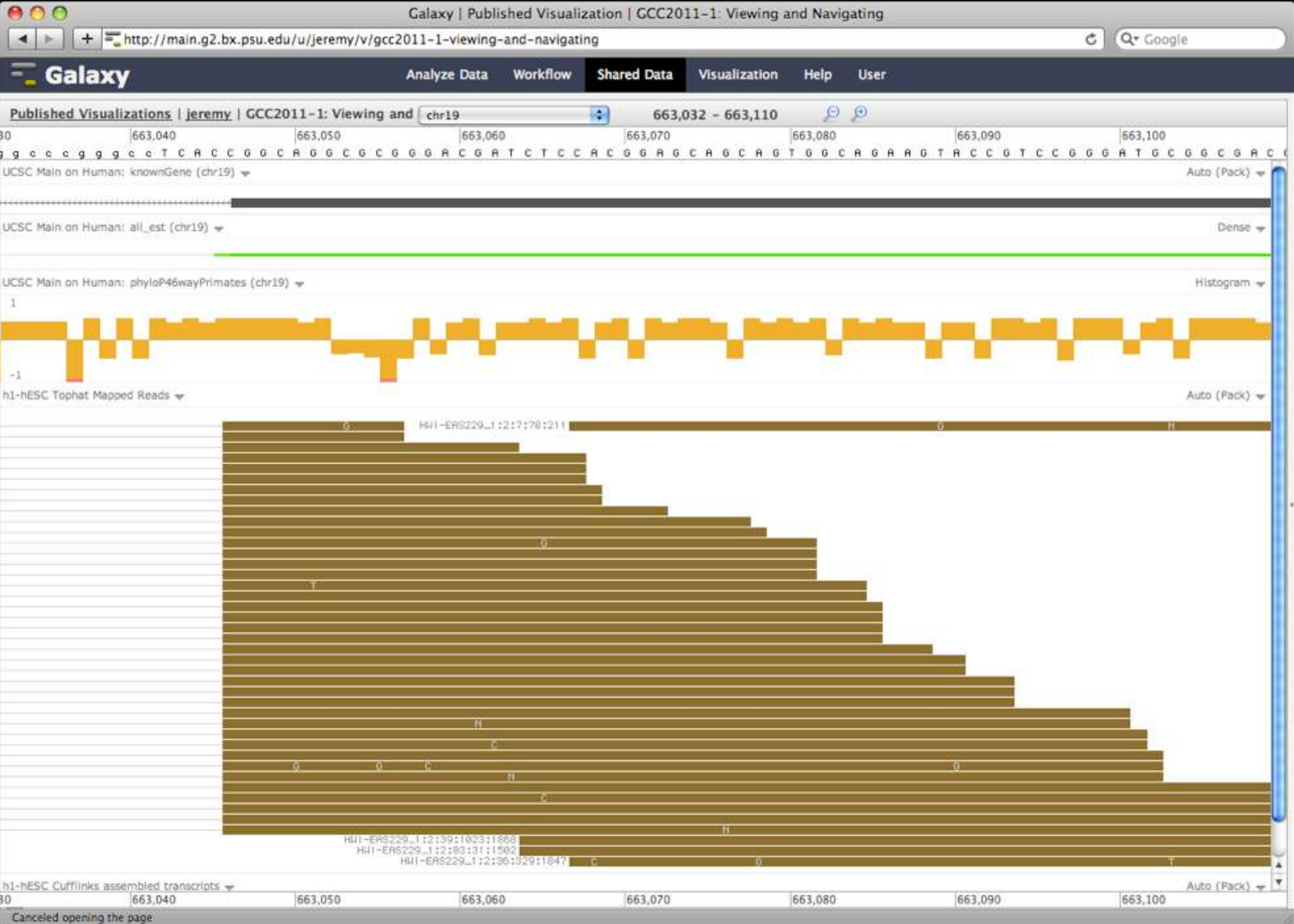
- ✦ BAM, BED, GFF/GTF, WIG

Unique features

- ✦ custom genomes
- ✦ highly interactive







But really, why *another* genome browser

From static browsing to **visual analysis**

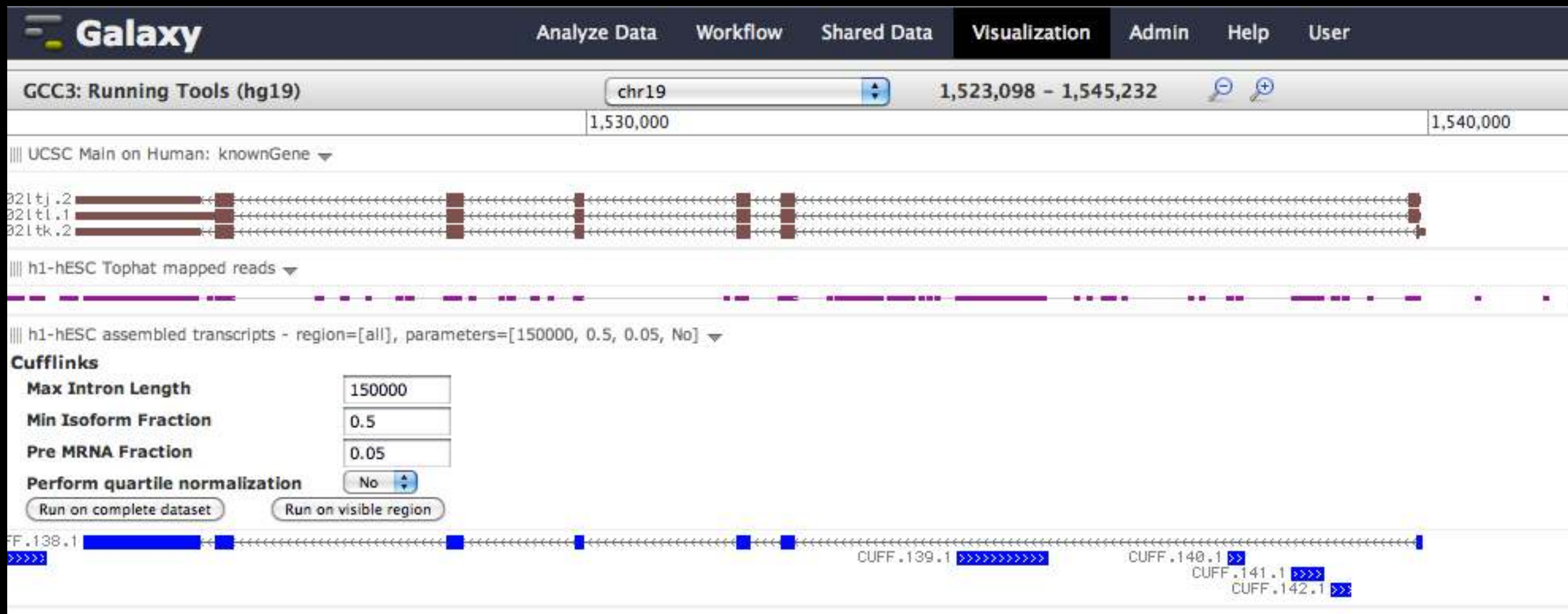
Visual feedback and experimentation needed for complex tools with many parameters

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

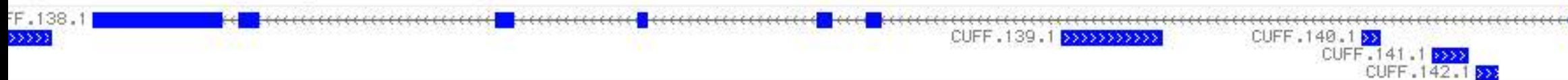
0.05

Perform quartile normalization

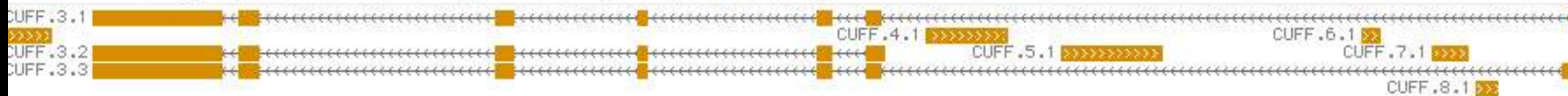
No

Run on complete dataset

Run on visible region



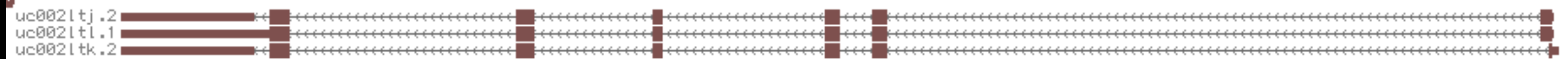
➔ Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No] ▼



1,530,000

1,540

UCSC Main on Human: knownGene ▼



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



Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No] ▼



Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.001, No] ▼



Overview

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

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

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.q2.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 History 'Variant Analysis for Sample E18'

 [Import history](#)

Dataset

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

Annotation

Variants with consensus different that occur in RefSeq genes.

Author



[All published histories](#)
[Published histories by jgoecks](#)

Community
(1 rating, 4.0 average)

Yours

Tags

Community:

[snp](#) [pileup](#) [bowtie](#) [demo](#)

sample

Yours

[snp](#) [x](#) [pileup](#) [x](#) [bowtie](#) [x](#)

demo x sample:e18 x

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](#)

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

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

Step 6: FASTQ Trimmer

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

Trim reads to remove low-quality bases.

Step 7: 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

Select a reference genome

/galaxy/data/apiMe13/bowtie_index/apiMe13

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

Map reads using default parameter values.

Step 8: SAM-to-BAM

Choose the source for the reference list

Locally cached

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

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 

Published Histories


[Advanced Search](#)

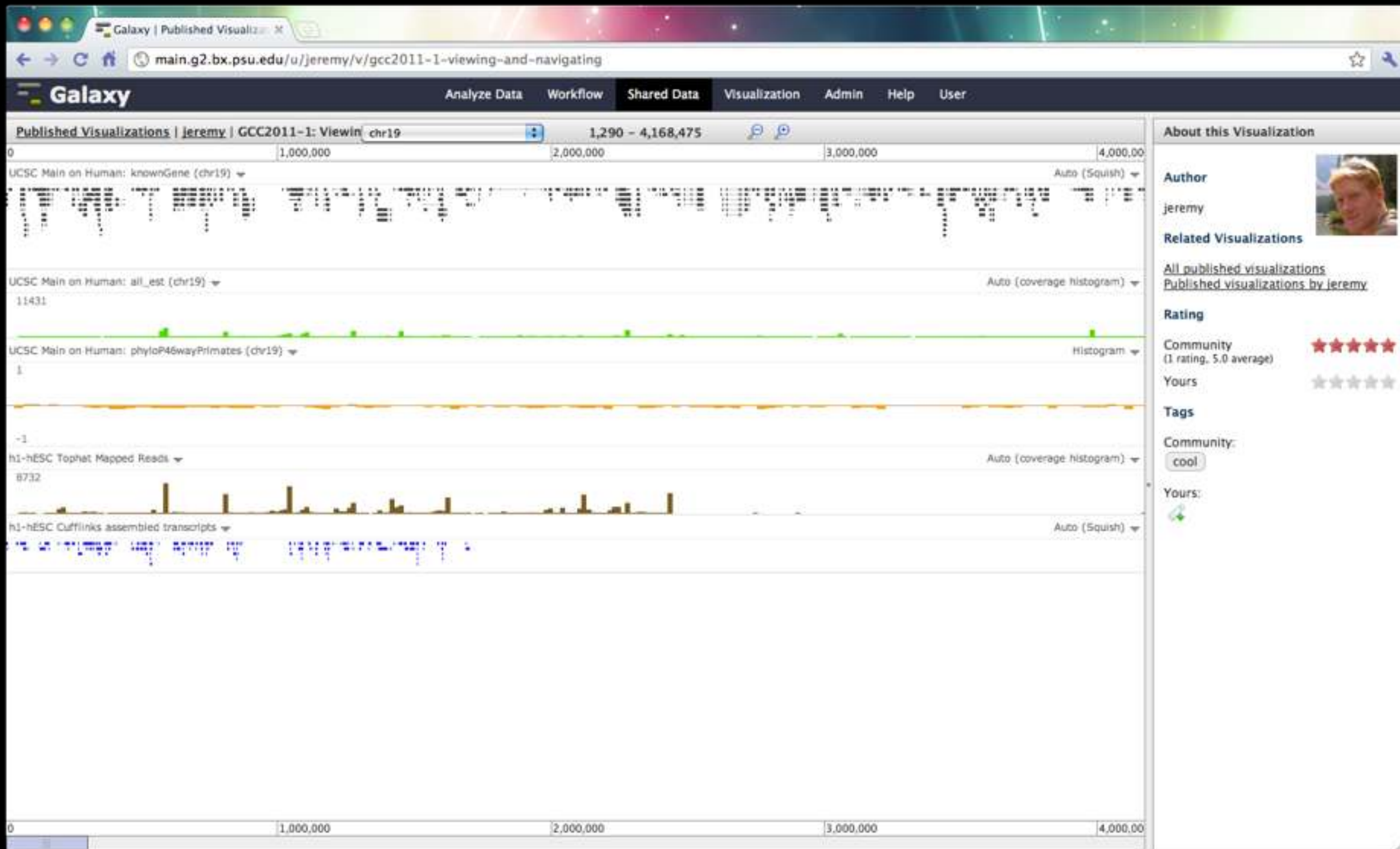
Name	Annotation	Owner	Community Rating 1	Community Tags	Last Updated
Galaxy vs MEGAN	Comparison of Galaxy vs. MEGAN pipeline.	aun1	★★★★★	metagenomics megan galaxy	Mar 19, 2010
metagenomic analysis		aun1	★★★★★	metagenomics galaxy	Mar 19, 2010
SM_1186088	Datasets correspond to our paper published in Science 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		jl	★★★★★		Aug 26, 2010
EKLE		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



Overview

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

Galaxy Pages

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

Galaxy Pages

The screenshot shows a web browser window displaying a Galaxy page. The browser's address bar shows the URL <http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18>. The Galaxy navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The page title is 'Variant Analysis of Embryonic Mouse Brain Tissue' by Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. The 'Results' section describes a variant analysis experiment on mm9 brain tissue. A green box highlights a 'Galaxy Dataset' link. The 'Method' section describes the bioinformatics pipeline. A blue box highlights a 'Galaxy History' link. An orange box highlights a 'Galaxy Workflow' link. The 'References' section lists three scientific papers. On the right, the 'About this Page' sidebar shows the author's profile, related pages, and a rating section.

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).

[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:

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/fdcdbcf1207fbd2>
Google

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	G G G G G . G G G G G G G . G G
chr10	14465083	14465084	G	K	144	144	60	35	 T T . . T T T T
chr10	14465485	14465486	C	T	129	129	60	34	t \$ T T T T T T T T T T T T T T T T T
chr10	19928287	19928288	G	A	135	135	60	36	A A A A A A A A A A A A A A A A A A
chr10	19928468	19928469	C	T	132	132	60	35	T \$ t t t t t T T t t t t t T t t T T T
chr10	19928494	19928495	C	T	138	138	60	37	T T T T t t t t t t t t t t T T t T t T
chr10	19928543	19928544	A	G	147	147	60	40	G g G g G g g g G G G G G G g G G G G
chr10	28750217	28750218	C	T	138	138	60	37	T T t t t t T T T t T t T t T T T T T
chr10	28750397	28750398	A	C	154	211	60	64	C \$. \$ C \$ C \$ C \$ C \$ C \$ C C C C C
chr10	28750423	28750424	C	T	113	113	60	35	T \$ t t T T T t t t t T T A T T T T t T
chr10	28750446	28750447	A	G	165	165	60	46	G \$ G G g g G G g g G G g g G G G G g
chr10	28750548	28750549	G	C	255	255	60	97	C \$ C \$ C \$ C \$ C \$ C C C C C C C C C
chr10	28750640	28750641	T	C	165	165	60	46	C e e e e e e e C C C C C C C C C C C
chr10	28750746	28750747	G	A	202	202	60	58	A A a a a a a a a a a A a a a a a a a
chr10	28750766	28750767	A	G	205	205	60	59	G \$ g \$ G \$ g \$ G \$ g g g g g g g g g
chr10	28750769	28750770	T	C	175	175	60	49	c e e c C e c C C C C C C c C e c c c c
chr10	28750924	28750925	C	T	182	217	60	64	T \$ T \$ t t T t t t T t T t T t T t T g
chr10	28751092	28751093	a	A	147	0	60	123	 , g g , , g g g g ,
chr10	28751096	28751097	a	A	212	0	60	119	g \$ g \$, , , , , , , , , , , , , , ,
chr10	28751114	28751115	g	A	225	235	60	85	a a a a , a a a a a a a a a a a a a a
chr10	28751117	28751118	T	A	191	190	60	79	a \$ a \$ a \$ a \$ a \$ a \$ a \$ a \$ a \$ a \$
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 \$ a a a a a a a a a a a a a a a a a
chr10	33613489	33613490	G	A	82	114	60	30	. \$ A \$ A \$ A \$ A \$ A \$ A \$ A \$ A \$ A \$
chr10	36721501	36721502	G	K	129	129	60	43	T T T T , , , , , , , , , ,
chr10	36721507	36721508	C	Y	51	51	60	54	. \$. , , , , , , , , , , , , , , , ,
chr10	36721695	36721696	T	A	120	120	60	31	a \$ a \$ a \$ a \$ a \$ a \$ a \$ a \$ a \$ a \$
chr10	36805412	36805413	A	G	126	126	60	33	G G G G G G G G G G G G G G G G G G
chr10	36805605	36805606	A	G	120	120	60	31	g g g g g g g g g g g g g g g g g g
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	 G G G
chr10	36855350	36855351	a	R	156	156	60	56	 G G G
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	c \$ c \$ c \$ c \$ c \$, , , , , , , ,
chr10	39144591	39144592	G	C	126	126	60	33	C C C C C C C C C C C C C C C C C C
chr10	42722938	42722939	C	Y	63	63	60	34	t \$ t \$ t \$ t \$ t \$, , , , , , , , ,

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 DataWorkflowShared DataVisualizationHelpUser

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

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

The screenshot displays the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The left sidebar lists various tools under categories like 'Get Data', 'Text Manipulation', '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', and 'SNP/WGA: Statistical Models'.

The main content area shows the 'Filter pileup' tool configuration. The 'Select dataset' dropdown is set to '9: Generate pileup on data 8'. The 'which contains' dropdown is set to 'Pileup with ten columns (with consensus)'. The 'Do not consider read bases with quality lower than' is set to 20. The 'Do not report positions with coverage lower than' is set to 30. The 'Only report variants?' dropdown is set to 'Yes'. The 'Convert coordinates to intervals?' dropdown is set to 'Yes'. The 'Print total number of differences?' dropdown is set to 'Yes'. The 'Print quality and base string?' dropdown is set to 'Yes'. The 'Execute' button is visible at the bottom of the configuration panel.

The right sidebar shows the 'History' panel with a list of jobs. The jobs are listed in descending order of age. The top job is '15: Variants from sample E18, consensus different, in RefSeq Genes'. Below it are '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', and '6: E18 PE,2 Reads'.

Below the job list, there is a table with 6 columns: '1. Chrom', '2. Start', '3. End', '4. A', '5. G', and '6. C'. The table contains 6 rows of data, each representing a variant. The first row is 'chr10 6882036 6882037 A A 107'. The second row is 'chr10 14243075 14243076 G G 96'. The third row is 'chr10 14243079 14243080 C C 106'. The fourth row is 'chr10 14465082 14465083 T K 173'. The fifth row is 'chr10 14465083 14465084 G K 144'. The sixth row is 'chr10 14465084 14465085 T T 117'.

At the bottom of the page, there is a footer that says 'Open "http://main.g2.bx.psu.edu/tool_runner/rerun?id=1703758" in a new tab'.

Galaxy Pages

The screenshot shows a web browser window displaying a Galaxy page. The browser's address bar shows the URL <http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18>. The Galaxy logo is in the top left, and navigation links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' are in the top right. The page title is 'Variant Analysis of Embryonic Mouse Brain Tissue' by 'Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team'. The 'Results' section describes a variant analysis experiment on mm9 brain tissue. A green box highlights a 'Galaxy Dataset' for intersecting variants. The 'Method' section details the bioinformatics pipeline. A blue box highlights a 'Galaxy History' entry for the variant analysis. Below, an 'Import workflow' button is shown next to a 'Galaxy Workflow' for identifying variants in annotated genes. The 'References' section lists three scientific papers. On the right, an 'About this Page' sidebar shows the author's profile, related pages, and a rating section.

Galaxy | Published Page | Variant Analysis for sample E18

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

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.

Import workflow

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).

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

Creating a Page

The screenshot shows the Galaxy web interface in a browser window. The address bar displays the URL: http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427. The Galaxy logo is in the top left, and navigation links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' are in the top right. Below the navigation bar, the page editor title is 'Page Editor | Title : Variant Analysis for sample E18'. The editor toolbar includes buttons for bold, italic, text color, background color, bulleted list, numbered list, link, unlink, undo, redo, and icons for inserting a link to a Galaxy object or embedding a Galaxy object. The main content area contains the following text:

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

The screenshot shows the Galaxy web interface with the 'Page Editor' for a page titled 'Variant Analysis for sample E18'. The page content includes sections for 'Results', 'Method', and 'References'. An 'Embed Histories' dialog is open, showing a list of available histories for embedding into the page.

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: History 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

The screenshot shows the Galaxy web interface in a browser window. The address bar displays the URL: http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427. The Galaxy logo is in the top left, and navigation links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' are in the top right. Below the navigation bar, the page title is 'Variant Analysis for sample E18'. A toolbar contains icons for text formatting (bold, italic, list, link, etc.) and buttons for 'Paragraph type', 'Insert Link to Galaxy Object', and 'Embed Galaxy Object'. The main content area has a title 'Variant Analysis of Embryonic Mouse Brain Tissue' and authors 'Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team'. It includes sections for 'Results' and 'Method'. The 'Results' section describes a variant analysis experiment. The 'Method' section describes the workflow from read grooming to variant identification. A blue box contains an embedded Galaxy history titled 'Variant Analysis for Sample E18' with a warning not to edit it. Below this, it says 'Here is a workflow for performing this analysis:'. A 'References' section at the bottom lists a paper by Han et al. (2009). At the very bottom, a footer says 'Open # on this page in a new tab'.

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

B I x₁ x₂ 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

The screenshot shows the Galaxy web interface in a browser window. The address bar shows the URL: http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427. The Galaxy logo is in the top left, and navigation links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' are in the top right. The page title is 'Page Editor | Title : Variant Analysis for sample E18'. Below the title bar is a toolbar with various icons and three dropdown menus: 'Paragraph type', 'Insert Link to Galaxy Object', and 'Embed Galaxy Object'. The main content area contains the following text:

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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

How to get into Galaxy

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

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

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

Lack IT knowledge or resources?

No problem, just use the Cloud

Overview

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

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

The screenshot displays the AWS Management Console interface. At the top, the Amazon Web Services logo is visible, along with navigation links for AWS, Products, Developers, Community, Support, and Account. The 'Account' link is highlighted with a red box. Below the navigation bar, the 'Your Account' section is visible, containing links for Account Activity, Consolidated Billing, DevPay Activity, and Payment Method. The 'Personal Information' section is also visible, containing links for Security Credentials and Usage Reports. The 'Security Credentials' link is highlighted with a red box. The main content area shows the 'Amazon EC2 Console Dashboard' for the 'US East' region. The dashboard includes a 'Getting Started' section with a 'Launch Instance' button, a 'My Resources' section showing 0 Running Instances, 0 Elastic IPs, 6 EBS Volumes, and 12 EBS Snapshots, and a 'Service Health' section showing the current status of Amazon EC2 (US East Virginia) as 'OK'. A 'Request Instances Wizard' modal is open, showing the 'CHOOSE AN AMI' step. The wizard displays the selected AMI as 'Other Linux AMI ID ami-ed03ed84 (x86_64)' and provides options to edit the AMI, instance details, advanced details, key pair, and firewall. The 'Launch' button is visible at the bottom right of the wizard.

amazon web services

Sign in to the AWS Management Console | Create an AWS Account

AWS | Products | Developers | Community | Support | **Account**

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Enis Afgan | Settings | Sign Out

Amazon S3 | **Amazon EC2** | Amazon Elastic MapReduce | Amazon CloudFront | Amazon RDS

Navigation

Region: US East

EC2 Dashboard

INSTANCES

- Instances
- Spot Requests

IMAGES

- AMIs
- Bundle Tasks

ELASTIC BLOCK STORE

- Volumes
- Snapshots

NETWORKING & SECURITY

- Elastic IPs
- Security Groups
- Key Pairs
- Load Balancers

Amazon EC2 Console Dashboard

Getting Started

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Launch Instance

Note: Your instances will be launched in the US East (Virginia) region.

My Resources

You are using the following Amazon EC2 resources in the US East (Virginia) region:

0 Running Instances | 0 Elastic IPs | 6 EBS Volumes | 12 EBS Snapshots

Service Health

Current Status

Amazon EC2 (US East Virginia) **OK** [View](#)

Request Instances Wizard

CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | CONFIGURE FIREWALL | REVIEW

Please review the information below, then click **Launch**.

AMI: Other Linux AMI ID ami-ed03ed84 (x86_64) [Edit AMI](#)

Number of Instances:

Availability Zone: No Preference

Monitoring: Disabled

Instance Type: Large (m1.large)

Instance Class: On Demand [Edit Instance Details](#)

Kernel ID: Use Default

Ramdisk ID: Use Default

User Data: testGC1|AKIAJKQI3RT... [Edit Advanced Details](#)

Key Pair Name: galaxy_keypair [Edit Key Pair](#)

Security Group(s): default, galaxyWeb [Edit Firewall](#)

[Back](#) **Launch**

3. Configure Your Cluster

The screenshot shows a web browser window with the URL `ec2-50-16-1-149.compute-1.amazonaws.com/cloud`. The page title is "Galaxy Cloudman". In the top right corner, there are links for "Info: [report bugs](#) | [wiki](#) | [screencast](#)".

The main content area is partially obscured by a modal dialog titled "Initial Cluster Configuration". The dialog contains the following text:

Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.

Below the text, there is a radio button selected for "Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)". The storage size is set to "1000 GB" with a green "OK" status. A link "Show more startup options" is visible below the storage size input.

At the bottom of the dialog, there is a "Start Cluster" button.

In the background, the "Galaxy Cloudman" interface is visible. It includes a "Status" section with labels for "Cluster name", "Disk status:", "Worker status:", "Service status:", and "External Logs". There is also a "Terminal" button and a "Cluster status log" section at the bottom.

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 able to 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



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.

The image displays two overlapping screenshots of the Galaxy Cloud interface. The top-left screenshot shows the 'Saved Histories' table, which lists various history entries with columns for Name, Datasets (by state), Tags, Sharing, Created, and Last Updated. The bottom-right screenshot shows the 'Galaxy Cloud Console' with a 'Scale' section containing 'Add more instances' and 'Remove idle instances' buttons, and a 'Status' section showing cluster details and a log of events.

Name	Datasets (by state)	Tags	Sharing	Created	Last Updated
mt replicates pair 3	8	96	0 Tags	about 1 hour ago	2 m ago
mt replicates pair 2	8	96	0 Tags	about 1 hour ago	15 min ago
mt replicates pair 1 testing	35	3	66	0 Tags	about 2 hours ago
mt datasets	24		0 Tags	about 2 hours ago	about 2 hours ago

Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

Scale

[Add more instances](#) [Remove idle instances](#)

Status

Cluster name: james-galaxy-cluster-9May2010-1
Cluster status: Ready
Instance status: Idle: 0 Available: 4 Requested: 4

[Access Galaxy](#)

Cluster status log

- 14:54:40 - Instance 'i-ade7b2c6' ready
- 14:54:40 - Setting up Galaxy
- 14:54:40 - Starting Galaxy...
- 14:54:45 - Instance 'i-a1e7b2ca' ready
- 14:54:49 - Instance 'i-afe7b2c4' ready
- 14:54:56 - Instance 'i-a3e7b2c8' reported alive
- 14:55:00 - Sent master public key to worker instance 'i-a3e7b2c8'.
- 14:55:00 - Adding instance i-a3e7b2c8 to SGE Execution Host list
- 14:55:01 - Successfully added instance 'i-a3e7b2c8' to SGE
- 14:55:01 - Waiting on worker instance 'i-a3e7b2c8' to configure itself...
- 14:55:09 - Instance 'i-a3e7b2c8' ready
- 14:55:16 - Galaxy started successfully!
- 14:55:16 - Ready for use

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

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/

Google

AWS Management Console

Galaxy Cloud

Galaxy Cloudman

Info: [report bugs](#) | [wiki](#) | [screencast](#)

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. Your previous data store has been reconstructed and nodes for running jobs.

Terminate cluster

Status

Cluster name:

James

Disk status:

1811

Worker status:

Idle

Service status:

Applied

External Logs:

Galaxy Cloudman

Cluster status log

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

cloudMan. Your previous

and remove 'worker'

Access Galaxy

Autoscaling is off.

Turn on?

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://s3.amazonaws.com/usegalaxy/UseGalaxy.ova>

Overview

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

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:

max distance between intervals: (bp)

min number of intervals per cluster:

Return type:

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

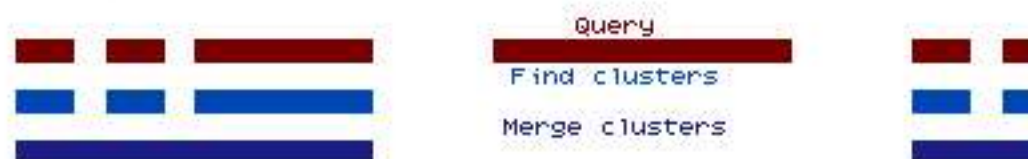
Screencasts!

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 output.
- **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



```

1 <tool id="gops_cluster_1" name="Cluster">
2   <description>[[Cluster]] the intervals of a query</description>
3   <command interpreter="python2.4">
4     gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
5       -d $distance -m $minregions -o $returntype
6   </command>
7   <inputs>
8     <param format="interval" name="input1" type="data">
9       <label>Cluster intervals of</label>
10    </param>
11    <param name="distance" size="5" type="integer" value="1" help="(bp
12      <label>max distance between intervals</label>
13    </param>
14    <param name="minregions" size="5" type="integer" value="2">
15      <label>min number of intervals per cluster</label>
16    </param>
17    <param name="returntype" type="select" label="Return type">
18      <option value="1">Merge clusters into single intervals</option>
19      <option value="2">Find cluster intervals; preserve comments and
20      <option value="3">Find cluster intervals; output grouped by clus
21      <option value="4">Find the smallest interval in each cluster</op
22      <option value="5">Find the largest interval in each cluster</opt
23    </param>
24  </inputs>
25  <help>
26
27  .. class:: infomark
28
29  **TIP:** If your query does not appear in the pulldown menu -> it is n
30
31  -----
32
33  **Screencasts!**
34
35  See Galaxy Interval Operation Screencasts (right click to open this l
36
37  .. \_Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc
38
39  -----
40
41  **Syntax**
42
43  - Maximum distance is greatest distance in base pairs allowed betw
44  - Minimum intervals per cluster allow a threshold to be set on the
45  - Merge clusters into single intervals outputs intervals that span
46  - Find cluster intervals; preserve comments and order filters out
47  - 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`

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

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)}&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>
```

2: SAM-to-BAM on data 1   

660.5 Mb, format: bam, database: mm9

Info:

| [display at UCSC main](#)

Binary bam alignments file

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

chr12 (qC1) 12qA1.1 qA2 12qA3 qB1 12qB3 12qC1 12qC2 12qC3 qD1 qD2 12qD3 12qE 12qF1 qF2

Scale 5 kb

to-BAM on data 1

STS Markers

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

Pax9

RefSeq Genes

Other RefSeq

Ensembl Gene Predictions

Human Proteins Mapped by Chained tBLASTn

Mouse mRNAs from GenBank

Spliced ESTs

Mammal Cons

Rat

Human

Orangutan

Dog

Horse

Opossum

Chicken

Stickleback

SNPs (126)

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($bam_file.dbkey)}&hgt.customText=${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
```

2: SAM-to-BAM on data 1   

660.5 Mb, format: bam, database:
mm9

Info:



| display at UCSC [main](#) [bx-main](#)

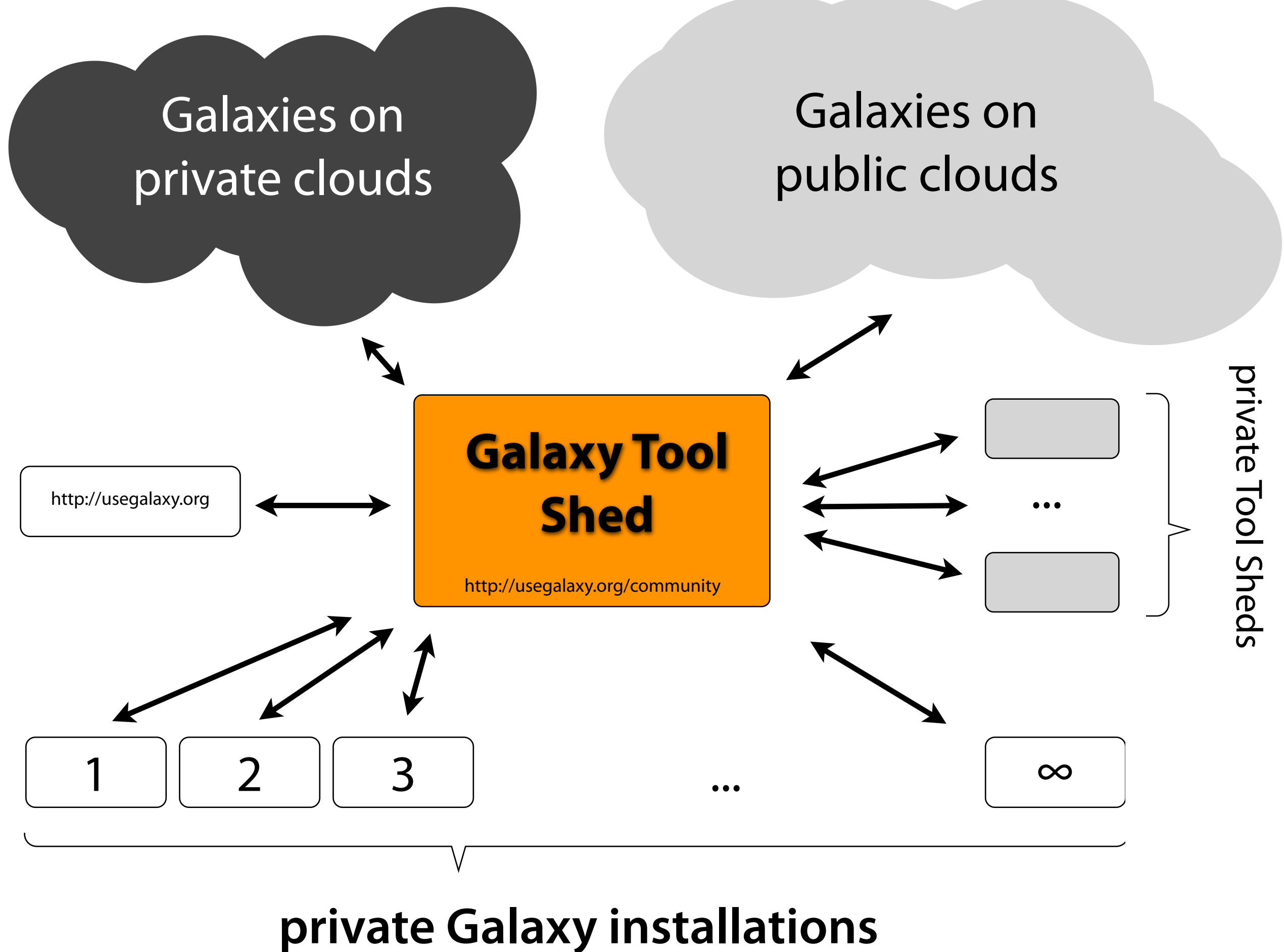
Binary bam alignments file

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



Get and Contribute Tools

 **Galaxy Tool Shed / (beta)** [Tools](#) [Help](#) [User](#)

Community

Tools

- [Browse by category](#)
- [Browse all tools](#)
- [Login to upload](#)

Categories

 [Advanced Search](#)

<u>Name ↓</u>	<u>Description</u>	<u>Tools</u>
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>

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



EMORY

PENNSTATE.



Enis Afgan



Dannon Baker



Dan Blankenberg



Nate Coraor



Dave Clements



Jeremy Goecks



Jennifer Jackson



Greg von Kuster



Kanwei Li



James Taylor



Kelly Vincent



Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

Overview

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

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

Galaxy 101 Exercise

Galaxy 101

<http://usegalaxy.org/galaxy101>

<http://ec2-50-16-165-27.compute-1.amazonaws.com/>

A simple question...

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

Galaxy 101

<http://usegalaxy.org/galaxy101>

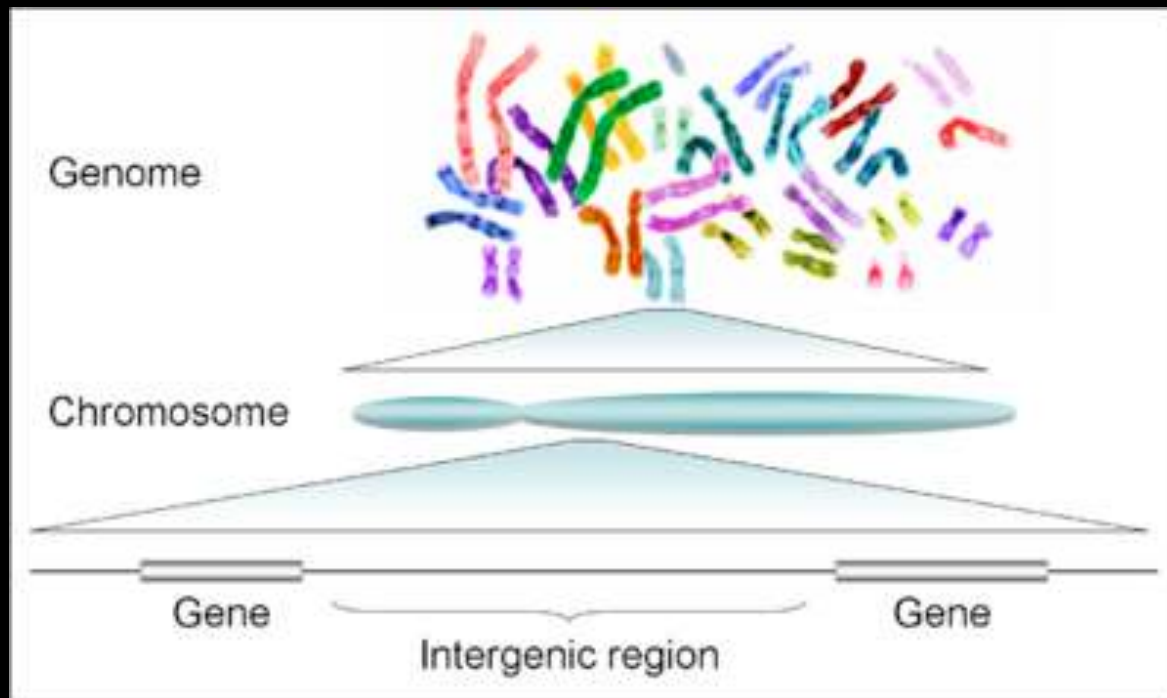
Overview

- ✦ 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  
taaccctaaccctaaccctaaccctaaccctaaccctaacccta  
accctaaccctaaccctaaccctaaccctaaccctaaccctaac
```

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