

Galaxy

Daniel Blankenberg

The Galaxy Team

<http://GalaxyProject.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

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 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, EMBOS, 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 configuration for the 'Map with Bowtie for Illumina' tool. The settings are as follows:

- Will you select a reference genome from your history or use a built-in index?:** Use a built-in index (selected). Built-ins were indexed using default options.
- Select a reference genome:** mm9 (selected).
- If your genome of interest is not listed - contact Galaxy team**
- Is this library mate-paired?:** Paired-end (selected).
- Forward FASTQ file:** 1: E18 PE.1 Reads (selected).
- Reverse FASTQ file:** 1: E18 PE.1 Reads (selected).
- 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) (selected).
- Bowtie settings to use:** Commonly used (selected).
- Suppress the header in the output SAM file:** ☒ (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:

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

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order

- Select lines that match an

Operate on Genomic Intervals

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

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases

Galaxy Analysis Workspace

The screenshot displays the Galaxy Analysis Workspace interface. The main panel shows a workflow titled "Bowtie for Illumina" with a form to select a reference genome. The right sidebar contains a "History" panel listing 15 steps, including "Imported: SNP Pileup Analysis for Sample E18", "Variants from sample E18, consensus different, in RefSeq Genes", "UCSC mm9 RefSeq Genes", "Variants from sample E18 where consensus base different than ref. base", "Variants from sample E18", "Generate pileup on data 8", "SAM-to-BAM on data 7", "Map with Bowtie for Illumina on data 6 and data 5", "E18 PE.2 Reads Groomed, Trimmed", "E18 PE.1 Reads Groomed, Trimmed", "E18 PE.2 Reads Groomed", "E18 PE.1 Reads Groomed", "E18 PE.2 Reads", and "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

Operate on Genomes

- Intersect the intersection of two queries

- Subtract the intersection of two queries

- Merge the overlap of two queries

- Filter SAM or BAM files

NGS: SAM Tools

- Filter SAM or BAM files

- Convert SAM to BAM

- SAM-to-BAM format to BAM

- BAM-to-SAM format to SAM

- Merge BAM files together

- Generate pileup dataset

- Filter pileup on coverage and SNPs

- Pileup-to-Interval condenses pileup format into ranges of bases

Filter pileup

Select dataset:

10: Variants from sample E18

which contains:

Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

3

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

No

See "Output format" below for explanation

Print total number of differences?:

No

See "Example 3" below for explanation

Print quality and base string?:

Yes

See "Example 4" below for explanation

Execute

ce

aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead. 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

Operate on Genomes

- Intersect the intersection of two queries

- Subtract the intersection of two queries

- Merge the overlap of a query

NGS: SAM Tools

- Filter SAM on values

- Convert SAM to BAM

- SAM-to-BAM format to BAM

- BAM-to-SAM format to SAM

- Merge BAM files together

- Generate pileup dataset

- Filter pileup on coverage and SNPs

- Pileup-to-Interval condenses pileup format into ranges of bases

Filter pileup

Select dataset:

10: Variants from sample E18

which contains:

Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

3

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

No

See "Output format" below for explanation

Print total number of differences?:

No

See "Example 3" below for explanation

Print quality and base string?:

Yes

See "Example 4" below for explanation

Execute

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

aligner designed to be ultrafast and memory-efficient. It is developed by Langmead B, Trapnell C, Pop M, Salzberg SL. U.S. Department of short DNA sequences to the human genome. Genome Biol

Filter and Sort

- Filter data on any complex or simple expressions
- Sort data in ascending or descending order

Select lines that match an expression

- Intersect the results of two queries
- Subtract the results of two queries
- Merge the results of two queries

NGS: SAM To BAM

- Filter SAM values
- Convert SAM to BAM
- SAM-to-BAM format to BAM
- BAM-to-SAM format to BAM
- Merge BAM files together
- Generate BAM dataset



This dataset is large and only the first megabyte is shown below.
[Show all](#) | [Save](#)

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

aligner designed to be ultrafast and memory-efficient. It is developed by Langmead B, Trapnell C, Pop M, Salzberg SL. U.S. Department of Health and Human Services, Genome Biology Division.

Analysis

Options

Analysis for Sample E18

Intersect to get Variants from Sample E18, consensus different, Genes

mm9 RefSeq Genes

to get Variants from Sample E18 where consensus base than ref. base

pileup to get from sample E18

ate pileup on data 8

to-BAM on data 7





with Bowtie for on data 6 and data 5

6: E18 PE.2 Reads Groomed, Trimmed

5: E18 PE.1 Reads Groomed, Trimmed


User Metadata

History
Options




Variant Analysis for Sample E18

Tags:

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






Annotation / Notes:

Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.


10: Variants from sample E18




26,742 regions, format: interval, database: mm9

Info:

Tags:


pileup ×
sample:e18 ×
snps ×


Annotation:

Find variants with coverage ≥ 30 and quality score ≥ 20 .

| display at UCSC [main](#) | view in [GeneTrack](#) | display at Ensembl [Current](#)

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



Datasources

Upload file from your computer

- ✦ FTP support for large datasets

Files directly from a sequencer

- ✦ Sample Tracking System

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

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

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
<input type="checkbox"/> All SNPs in personal genomes ▼	Summary table of SNPs in all individuals	greg@bx.psu.edu	2010-01-28	676.8 Mb
<input type="checkbox"/> Alu insertions in KB1 ▼		greg@bx.psu.edu	2010-02-10	14.9 Kb
<input type="checkbox"/> Alu insertions in NB1 ▼		greg@bx.psu.edu	2010-02-10	6.5 Kb
<input type="checkbox"/> KB1 microsatellites.txt ▼		greg@bx.psu.edu	2010-02-15	3.5 Mb
<input type="checkbox"/> NB1 microsatellites.txt ▼		greg@bx.psu.edu	2010-02-15	828.5 Kb
<input type="checkbox"/> amino acid differences with functional predictions ▼		greg@bx.psu.edu	2010-02-05	1.1 Mb
<input type="checkbox"/> gene copy number (HCP) and other variants ▼		greg@bx.psu.edu	2010-02-15	2.1 Mb
<input type="checkbox"/> indels in ABT ▼		greg@bx.psu.edu	2010-02-03	105.3 Kb
<input type="checkbox"/> indels in KB1 ▼		greg@bx.psu.edu	2010-02-03	14.2 Mb
<input type="checkbox"/> indels in MD6 ▼		greg@bx.psu.edu	2010-02-03	109.8 Kb
<input type="checkbox"/> indels in NB1 ▼		greg@bx.psu.edu	2010-02-03	51.5 Kb
<input type="checkbox"/> indels in TK1 ▼		greg@bx.psu.edu	2010-02-03	123.2 Kb
<input type="checkbox"/> novel SNPs in ABT ▼		greg@bx.psu.edu	2010-02-09	9.4 Mb
<input type="checkbox"/> novel SNPs in KB1 ▼		greg@bx.psu.edu	2010-02-09	16.9 Mb
<input type="checkbox"/> novel SNPs in MD6 ▼		greg@bx.psu.edu	2010-02-09	594.1 Kb
<input type="checkbox"/> novel SNPs in NB1 ▼		greg@bx.psu.edu	2010-02-09	4.1 Mb
<input type="checkbox"/> novel SNPs in TK1 ▼		greg@bx.psu.edu	2010-02-09	722.6 Kb
<input type="checkbox"/> 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 Workflows

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 is a "Tools" sidebar on the left containing categories like Get 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.

The main panel shows a dataset view with a warning icon and the message: "This dataset is large and only the first megabyte is shown below. Show all | Save". Below this is a table of genomic coordinates:

chr	pos	ref	alt	qual	phred	depth	coverage
chr10	6882036	A	A	107	0	60	32
chr10	14243075	C	G	0	0	96	0
chr10	14243079	T	C	0	0	106	0
chr10	14465082	C	X	173	176	60	35
chr10	14465083	G	X	144	144	60	35
chr10	14465084	T	T	117	0	60	38
chr10	14465085	G	G	70	0	60	38
chr10	14465257	C	C	79	0	60	42
chr10	14465258	A	A	137	0	60	46
chr10	14465263	A	A	136	0	60	61
chr10	14465366	A	A	101	0	60	38
chr10	14465371	G	G	137	0	60	50
chr10	14465410	G	G	184	0	60	69
chr10	14465447	T	T	186	0	60	65
chr10	14465456	T	G	193	0	60	70
chr10	14465465	T	T	177	0	60	63
chr10	14465485	A	A	129	129	60	34
chr10	14465557	C	T	219	0	60	84
chr10	14465581	G	G	240	0	60	84
chr10	14465586	C	C	248	0	60	82
chr10	14465621	C	C	134	0	60	49
chr10	14465658	C	C	134	0	60	49
chr10	14465660	T	T	153	0	60	55
chr10	14465691	G	G	128	0	60	42
chr10	14465778	C	C	89	0	60	34
chr10	14465792	G	G	104	0	60	33
chr10	14465881	G	G	110	0	60	41
chr10	17445088	A	A	103	0	60	34
chr10	17445271	A	A	55	0	60	34
chr10	17731269	T	T	113	0	60	42
chr10	19928287	G	A	135	135	60	36
chr10	19928468	C	T	132	132	60	35
chr10	19928488	A	A	119	0	60	44
chr10	19928494	C	T	138	138	60	37
chr10	19928527	A	A	134	0	60	45
chr10	19928538	G	G	144	0	60	52
chr10	19928543	A	G	147	147	60	40
chr10	19928741	T	T	80	0	60	30
chr10	20799826	G	G	117	0	60	37
chr10	28750217	C	T	138	138	60	37
chr10	28750397	A	A	154	211	60	64
chr10	28750401	A	A	128	0	60	47
chr10	28750423	A	T	113	113	60	35
chr10	28750438	A	A	95	0	60	36
chr10	28750446	A	G	165	165	60	46
chr10	28750487	A	A	80	0	60	31
chr10	28750512	G	G	220	0	60	72
chr10	28750548	G	C	255	255	60	83
chr10	28750574	T	T	233	0	60	82
chr10	28750577	T	T	234	0	60	76
chr10	28750578	T	T	242	0	60	75
chr10	28750593	G	G	220	0	60	75
chr10	28750640	T	C	165	165	60	46
chr10	28750746	G	A	202	202	60	58
chr10	28750766	A	G	205	205	60	59
chr10	28750769	T	C	175	175	60	49
chr10	28750787	T	T	225	0	60	90
chr10	28750797	C	C	180	0	6	

Galaxy Workflows

Galaxy

Tools

search tools

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- EMBOSS
- NGS TOOLBOX BETA
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Indel Analysis
- NGS: Peak Calling
- RGENETICS
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Workflows

Tool

History items created

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

History Lists

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

Delete

Generate pileup on

Map to-BAM on data

Map with Bowtie for

Map on data 6 and data 5

928 lines, format: sam,
se: mm9
sequence file aligned.

2: FLAG: 3:1

S269:3:1:1449:913 99 cba

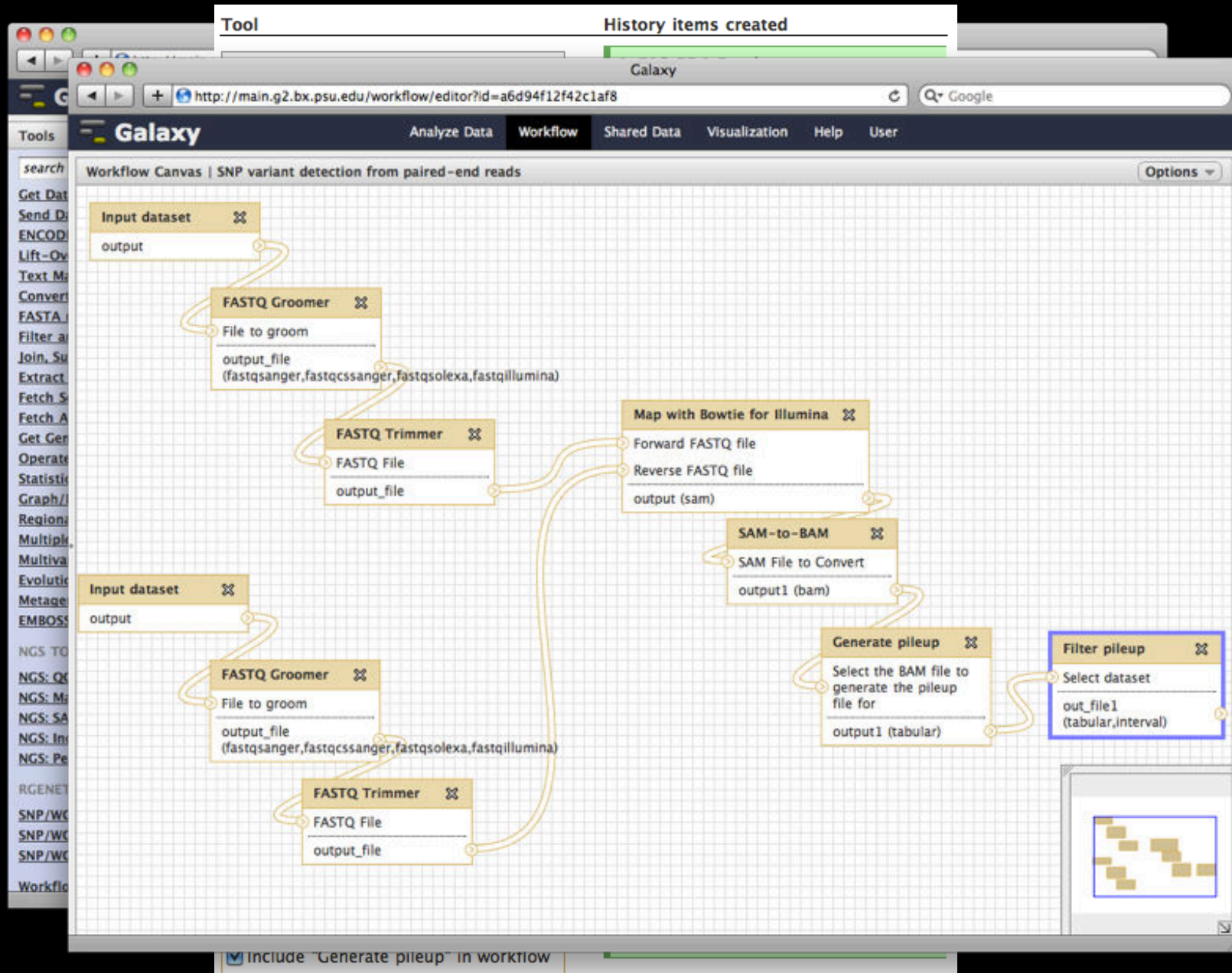
S269:3:1:1449:913 147 cba

S269:3:1:709:832 99 cba

S269:3:1:709:832 147 cba

S269:3:1:1422:1087 99 cba

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

ENCODE

Lift-Over

Text Manipulation

Conversion

FASTA

Filter

Join

Subtract

Extract

Fetch

Fetch

Get

Operate

Statistical

Graph

Regions

Multiple

Multivariate

Evolution

Metagenomics

EMBOSS

NGS TO

NGS: QC

NGS: M

NGS: SA

NGS: In

NGS: Pe

RGENET

SNP/WC

SNP/WC

SNP/WC

Workflow

Workflow Canvas | SNP variant detection from paired-end reads

Input dataset

output

FASTQ Groomer

File to groom

output_file (fastqsanger, fastqc, 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

Select

file fo

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 image displays the Galaxy Workflows interface. In the background, a workflow canvas titled "Workflow Canvas | SNP variant detection" is visible, showing a sequence of steps: "Input dataset" (output), "FASTQ Groomer" (output_file), "FASTQ Trimmer" (output_file), and "FASTQ File" (output_file). The "FASTQ Groomer" step is highlighted with a blue box.

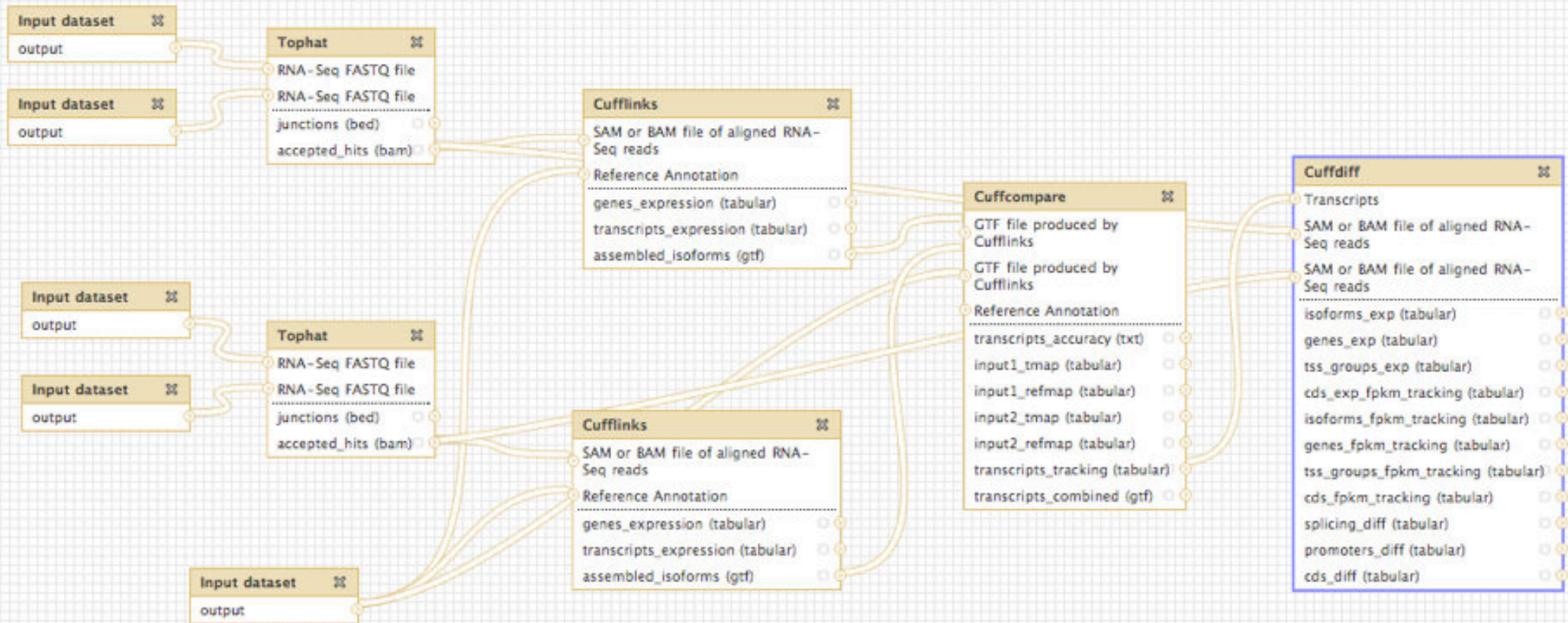
Overlaid on the canvas is the "Edit Workflow Attributes" dialog box. It contains the following information:

- Name:** SNP identification within annotated genes from NGS PE Data
- Tags:** snp, ngs, pileup, bowtie
- Annotation / Notes:** Identify variants in annotated genes from NGS paired-end data.

Below the dialog, a "Tool: SAM-to-BAM" configuration panel is shown. It includes the following options:

- Choose the source for the reference list:** Locally cached
- SAM File to Convert:** Data input 'input1' (sam)
- Edit Step Actions:** Assign Columns, output1, Create
- Edit Step Attributes:** Annotation / Notes: Convert Bowtie SAM output to BAM format so that pileup can be run.

At the bottom of the interface, a checkbox labeled "Include 'Generate pileup' in workflow" is visible.



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



1

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

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

12,000 bp

configure

14: Tag Counts (bigWig)

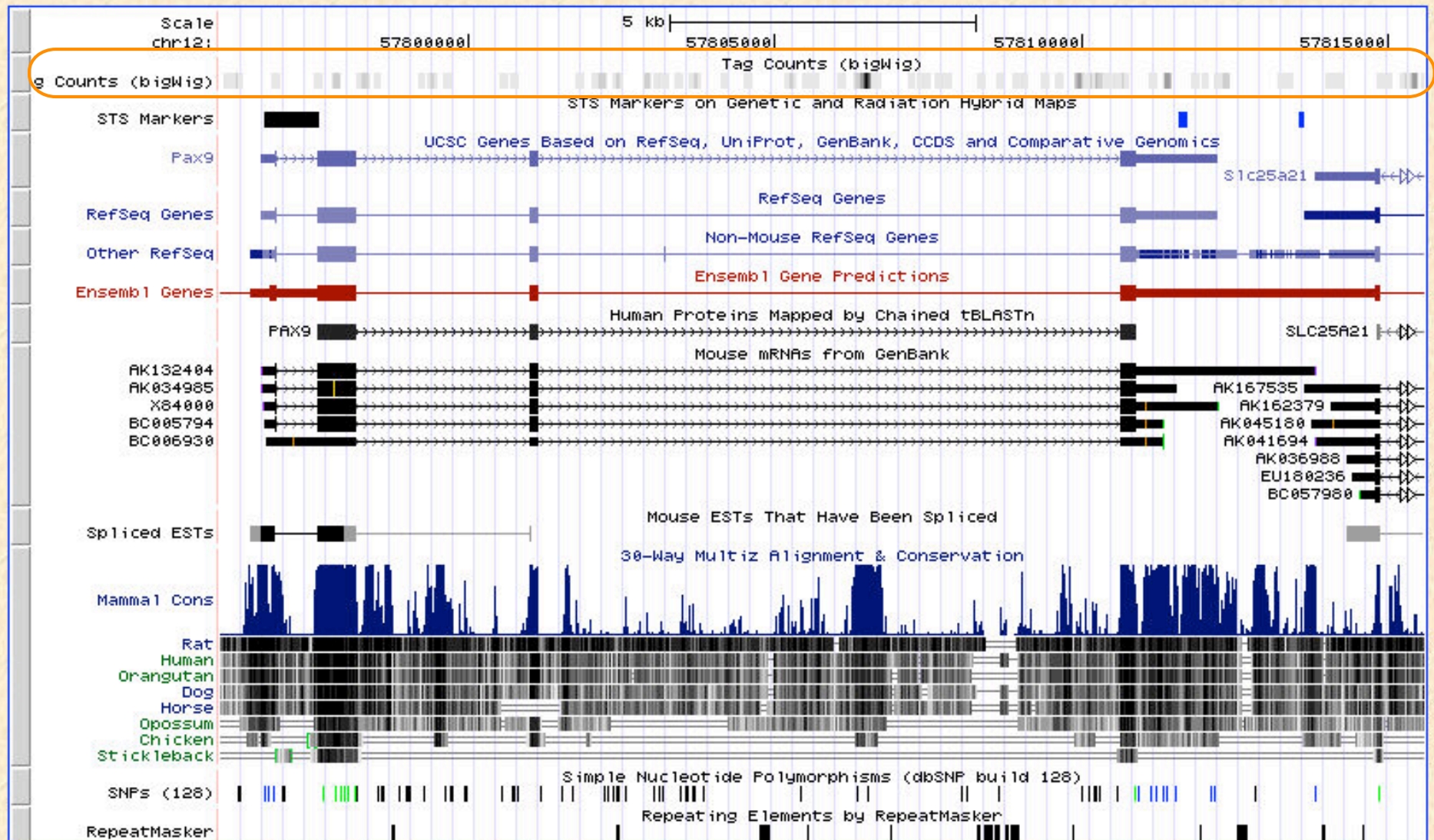
2.4 Gb, format: bigwig, database: mm9

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.

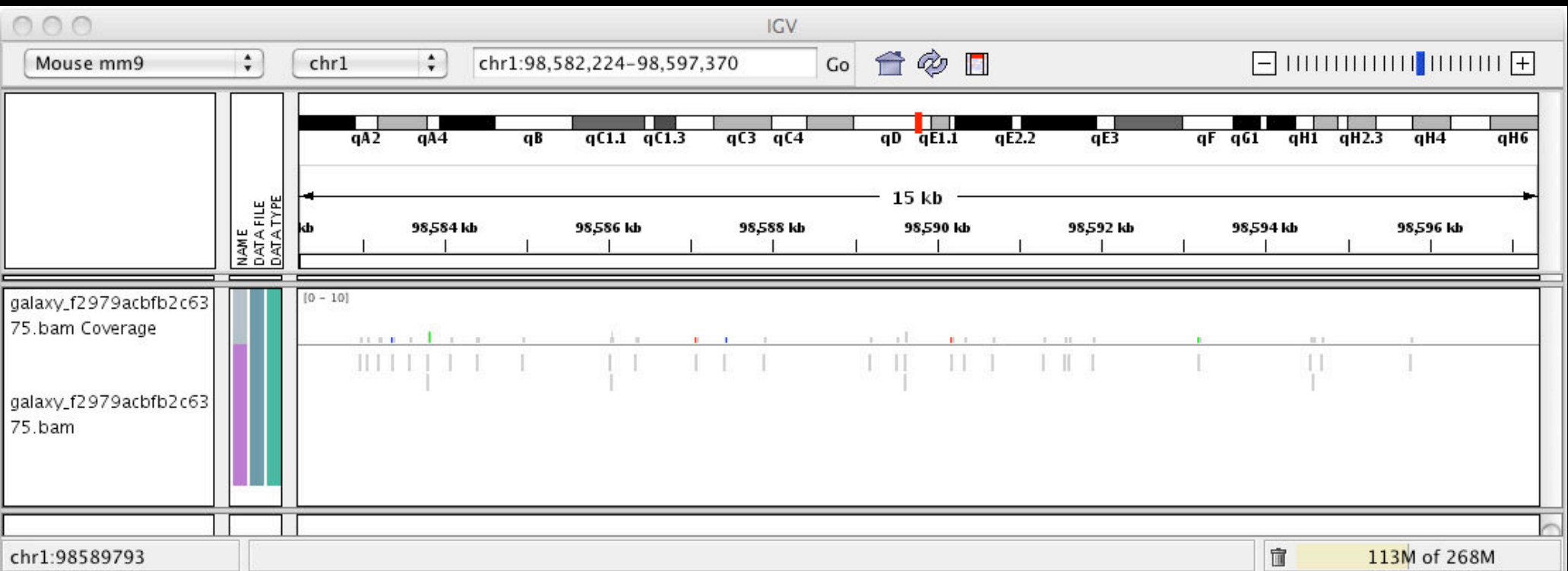
☐ Allow all applications from "www.broadinstitute.org" with this signature



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

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

[Published Visualizations](#) | [jeremy](#) | GCC2011-1: Viewing and chr19

chr19

1,290 - 4,168,475



0	1,000,000	2,000,000	3,000,000	4,000,000
---	-----------	-----------	-----------	-----------

UCSC Main on Human: knownGene (chr19) ▼

Auto (Squish) ▼



UCSC Main on Human: all_est (chr19) ▾

Auto (coverage histogram) ▾

11431

UCSC Main on Human: phyloP46wayPrimates (chr19) ▾

Histogram ▼

1

-1

h1-hESC Tophat Mapped Reads ▾

Auto (coverage histogram) ▾

8732

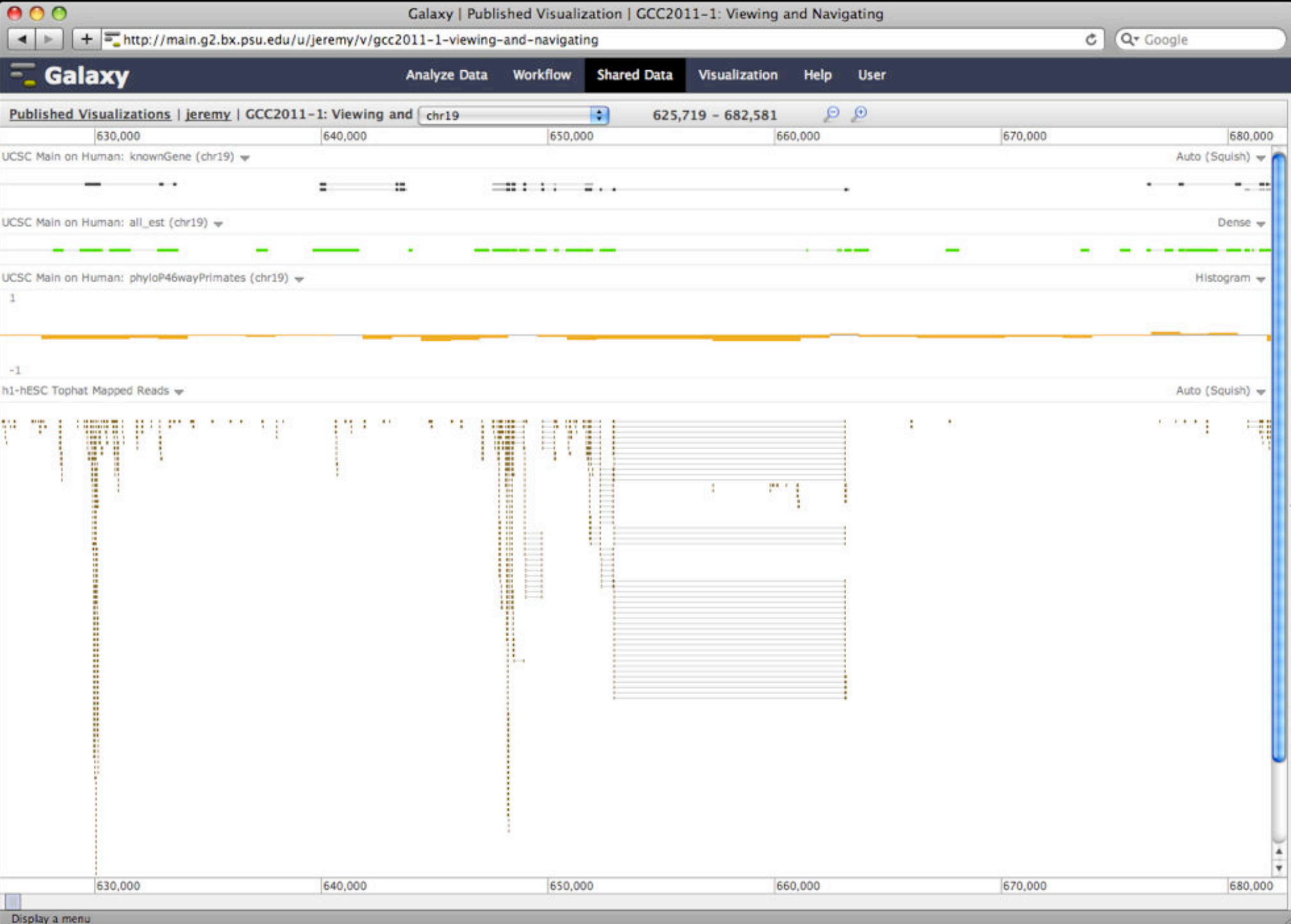


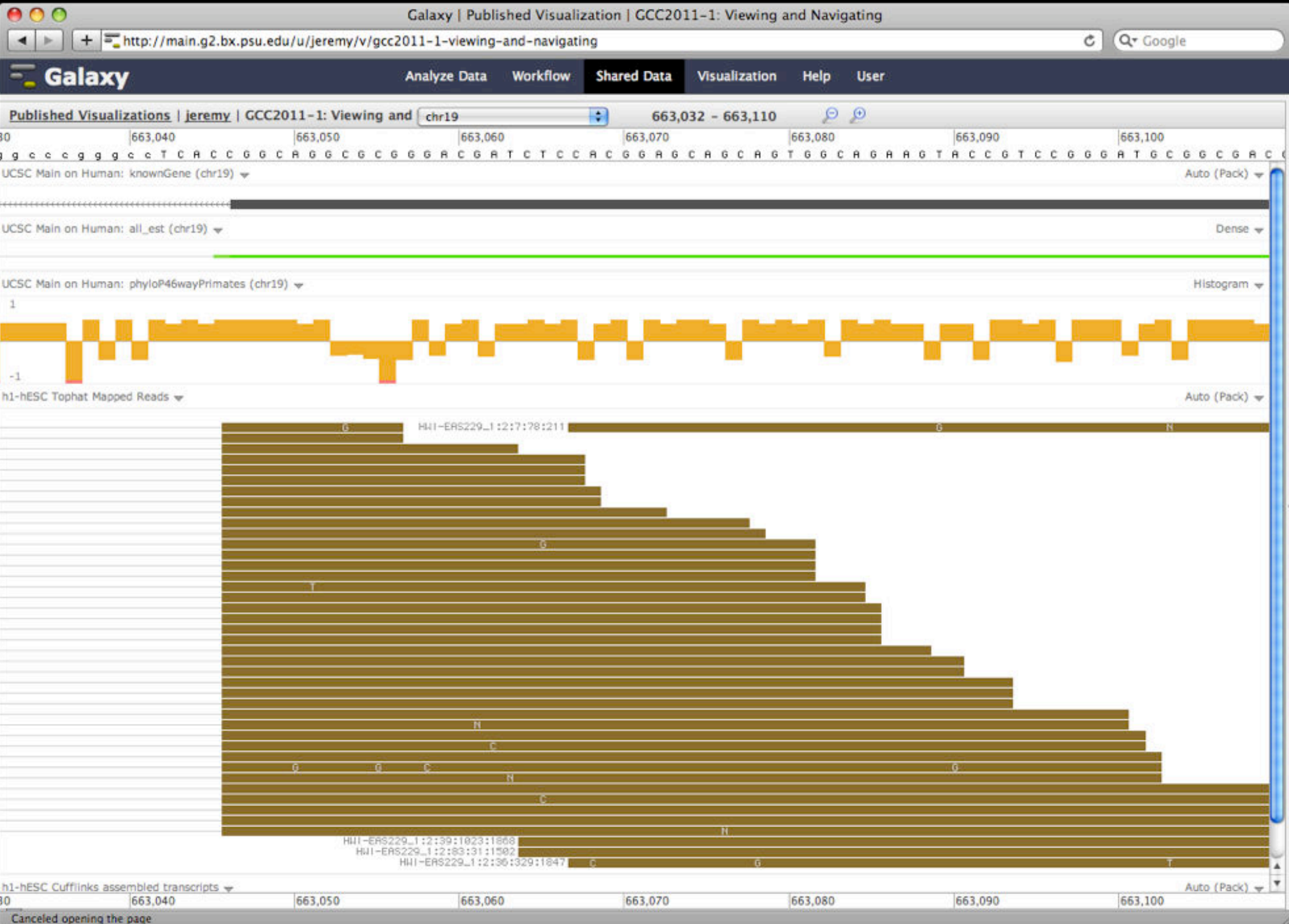
h1-hESC Cufflinks assembled transcripts ▾

Auto (Squish) ▼



0	1,000,000	2,000,000	3,000,000	4,000,000
---	-----------	-----------	-----------	-----------





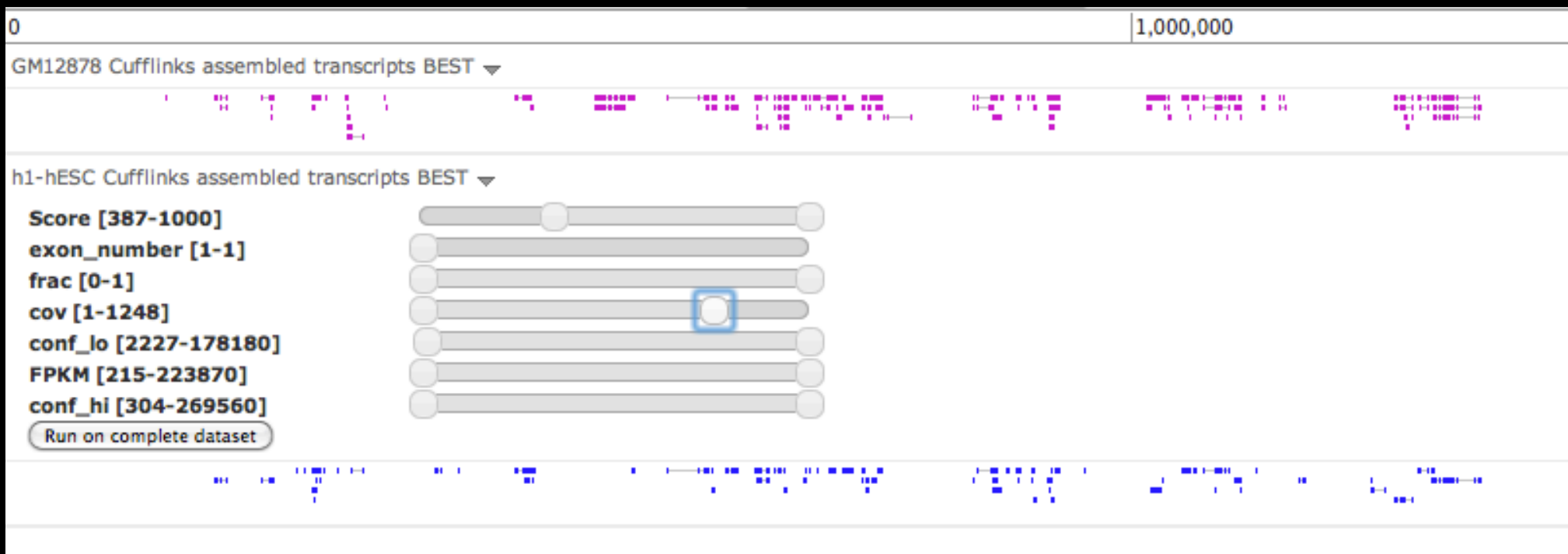
But really, why *another* genome browser

From static browsing to **visual analysis**

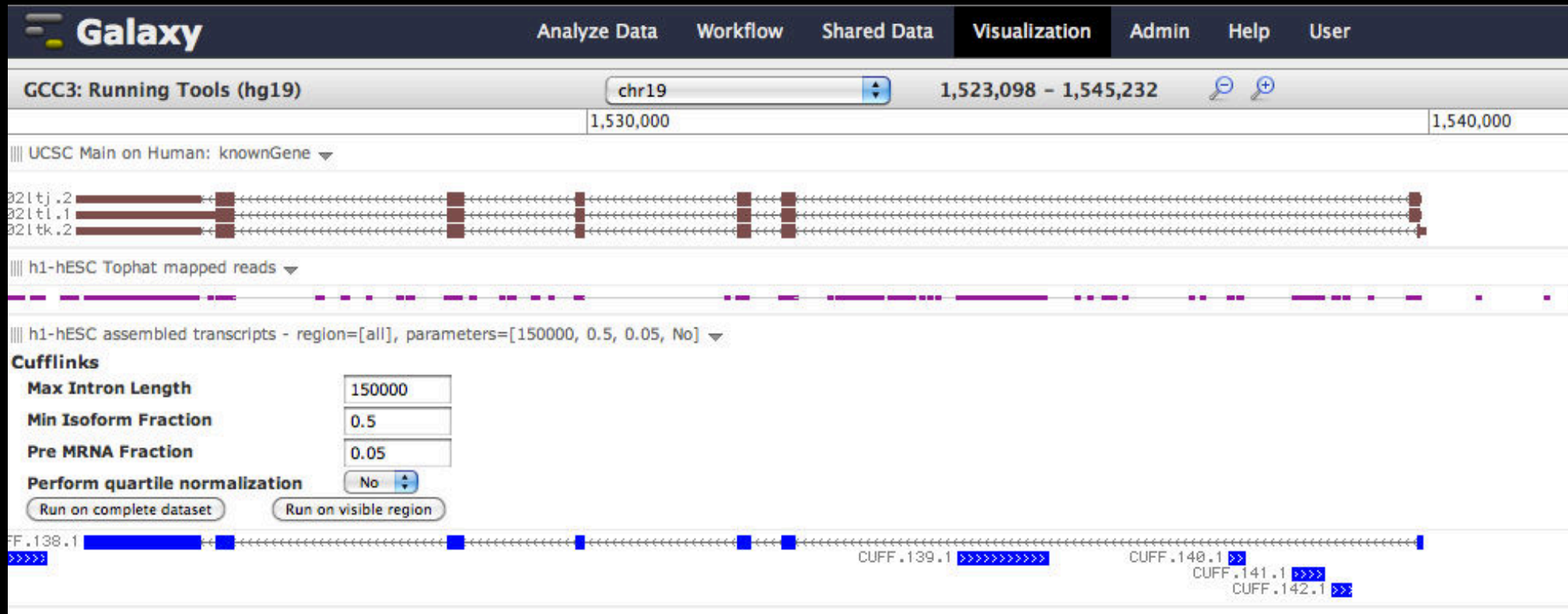
Visual feedback and experimentation needed for complex tools with many parameters

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



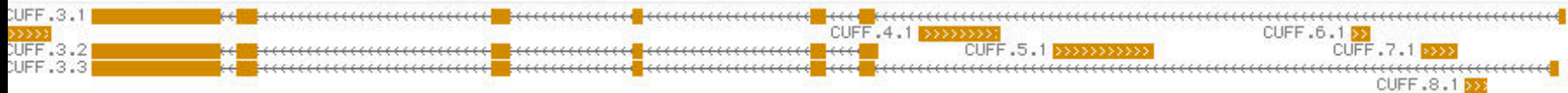
||| h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼

Cufflinks

Max Intron Length	<input type="text" value="150000"/>
Min Isoform Fraction	<input type="text" value="0.05"/>
Pre MRNA Fraction	<input type="text" value="0.05"/>
Perform quartile normalization	<input type="button" value="No"/>
<input type="button" value="Run on complete dataset"/> <input type="button" value="Run on visible region"/>	



➔ Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, 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

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](#)

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

Annotation

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

About this History

Author

joecks



Related Histories

[All published histories](#)

Published histories by jgoecks

Rating

Community
(1 rating, 4.0 average)



Yours

Tags

Community:

snp pileup bowtie demo
sample

Yours:

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

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

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

Google

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

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

Step 6: FASTQ Trimmer

FASTQ File

Output dataset 'output_file' from step 4

Define Base Offsets as

Absolute Values

Offset from 5' end

0

Offset from 3' end

9

Keep reads with zero length

False

Step 7: Map with Bowtie for Illumina

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

Use a built-in index

Select a reference genome

/galaxy/data/apiMel3/bowtie_index/apiMel3

Is this library mate-paired?

Paired-end

Forward FASTQ file

Output dataset 'output_file' from step 6

Reverse FASTQ file

Output dataset 'output_file' from step 5

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

1000

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

FR (for Illumina)

Bowtie settings to use

Commonly used

Suppress the header in the output SAM file

True

Step 8: SAM-to-BAM

Choose the source for the reference list

Locally cached

Trim reads to remove low-quality bases.

Map reads using default parameter values.

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 x bowtie x

43

Published Histories

search


[Advanced Search](#)

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

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

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 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' for intersecting variants. The 'Method' section details the bioinformatics pipeline. A blue box highlights the 'Galaxy History' entry for the analysis. An orange box highlights the 'Galaxy Workflow' for variant identification. The 'References' section lists two 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

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 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 G
chr10	14465083	14465084	G	K	144	144	60	35 T T T T T T T
chr10	14465485	14465486	C	T	129	129	60	34	t \$ T
chr10	19928287	19928288	G	A	135	135	60	36	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 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 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 C
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 T T
chr10	28750397	28750398	A	C	154	211	60	64	C \$. \$ 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 A T T 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 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 c C
chr10	28750640	28750641	T	C	165	165	60	46	C e c e c e c e c C 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 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 g G
chr10	28750769	28750770	T	C	175	175	60	49	c e c c C c C C C C C c C e c e c e 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 T g
chr10	28751092	28751093	a	A	147	0	60	123	/ / / / / / / / / / q q , q q q q ,
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 a a
chr10	28751117	28751118	T	A	191	198	60	79	a \$ 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 a a
chr10	33613489	33613490	G	A	82	114	60	30	. \$ a \$ a \$ a \$ a a a a 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 \$ 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 G G
chr10	36805605	36805606	A	G	120	120	60	31	q q q q q q q q q q q q q q q q q q q q
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	3							

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
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 links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The left sidebar lists various tools under categories like 'Tools', 'NGS TOOLBOX BETA', and 'RGENETICS'. 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, and 'Do not report positions with coverage lower than' is set to 30. The 'Only report variants?' and 'Convert coordinates to intervals?' are both set to 'Yes'. The 'Print total number of differences?' and 'Print quality and base string?' are also set to 'Yes'. The 'Execute' button is visible. The right sidebar shows a 'History' panel with a list of jobs, including '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 Z', '7: Map with Bowtie for Illumina on data 6 and data 5', and '6: E18 PE.2 Reads'. A table of variant data is also visible in the history panel.

Galaxy

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

Google

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools Options

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
RGENETICS
SNP/WGA: Data: Filters
SNP/WGA: QC: LD: Plots
SNP/WGA: Statistical Models

Filter pileup

Select dataset:
9: Generate pileup on data 8

which contains:
Pileup with ten columns (with consensus)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:
20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:
30

Pileup lines with coverage lower than this value will be skipped

Only report variants?:
Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:
Yes

See "Output format" below for explanation

Print total number of differences?:
Yes

See "Example 3" below for explanation

Print quality and base string?:
Yes

See "Example 4" below for explanation

Execute

What it does

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

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

History Options

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

14: UCSC mm9 RefSeq Genes

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

10: Variants from sample E18
26,742 regions, format: interval,
Run this job again
I display at UCSC main | view in GeneTrack | display at Ensembl Current

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

9: Generate pileup on data 8

8: SAM-to-BAM on data Z

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

6: E18 PE.2 Reads

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

Galaxy Pages

The screenshot shows a web browser window with the address bar displaying `http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18`. The page title is "Galaxy | Published Page | Variant Analysis for sample E18". The main content area is titled "Variant Analysis of Embryonic Mouse Brain Tissue" by Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. It includes a "Results" section with text about NGS re-sequencing and variant analysis. Below the text are two interactive buttons: "Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes" and "Galaxy History | Variant Analysis for Sample E18". A "Method" section follows, describing the workflow. At the bottom, there is a "References" section with three citations. On the right side, there is a sidebar titled "About this Page" containing information about the author (jgoecks), related pages, a rating (0 stars), and tags (none).

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.

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 a web browser window with the Galaxy logo and navigation menu. The page editor is titled 'Variant Analysis for sample E18'. The main content area displays a draft page with the following structure:

- Title:** Variant Analysis of Embryonic Mouse Brain Tissue
- Authors:** Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team
- Section: 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:
- Section: 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:
- Section: 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. The browser address bar displays `http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427`. The Galaxy navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The page editor title is "Variant Analysis for sample E18".

The background page content includes:

- Variant Analysis of Embryonic Development**
- Authors: Jeremy Goecks, Anton Nekrutenko, James Taylor, et al.
- Results**
To demonstrate how Galaxy can support accessible, identifies variants from a set of 4,536,964 RNA-seq...
- Method**
In the first step of this analysis, the reads were grouped and exclude base pairs with low quality scores; see [1] for Bowtie [2]. A pileup analysis using SAMtools [3] was...
- 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).

The "Embed Histories" dialog box is open, showing a search bar and a table of histories:

	Name	Tags	Last Updated ↑
<input checked="" type="checkbox"/>	Variant Analysis for Sample E18	5 Tags	15 minutes ago
<input type="checkbox"/>	Pileup analysis, sample E18	4 Tags	2 days ago
<input type="checkbox"/>	Unnamed history	0 Tags	Sep 07, 2010
<input type="checkbox"/>	Unnamed history	0 Tags	Dec 17, 2009
<input type="checkbox"/>	imported: Hsitory with ~100 items	5 Tags	Dec 10, 2009
<input type="checkbox"/>	imported: Galaxy vs MEGAN	0 Tags	Dec 04, 2009
<input type="checkbox"/>	imported: Galaxy vs MEGAN	2 Tags	Oct 06, 2009
<input type="checkbox"/>	imported: Galaxy vs MEGAN	0 Tags	Oct 06, 2009
<input type="checkbox"/>	imported: metagenomic analysis	0 Tags	Sep 30, 2009
<input type="checkbox"/>	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.

Buttons: 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 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 Galaxy objects. 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:

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 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 'Page Editor | Title : Variant Analysis for sample E18'. A toolbar contains icons for text formatting (bold, italic, subscript, superscript, bulleted list, numbered list, link, unlink, undo, redo) and dropdown menus for '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 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

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

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 navigation bar includes links for AWS, Products, Developers, Community, Support, and Account (highlighted with a red box). The main content area is divided into several sections:

- Your Account:** Includes links for Account Activity, Consolidated Billing, DevPay Activity, and Payment Method.
- Personal Information:** Includes links for Security Credentials and Usage Reports.
- Navigation:** A sidebar menu with categories like INSTANCES, IMAGES, ELASTIC BLOCK STORE, and NETWORKING & SECURITY. The 'EC2 Dashboard' link is highlighted.
- Amazon EC2 Console Dashboard:** The main area for managing EC2 instances. It 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.

The 'Request Instances Wizard' is open, showing the following configuration:

- CHOOSE AN AMI:** Other Linux AMI ID ami-ed03ed84 (x86_64)
- INSTANCE DETAILS:** Number of Instances: 1, Availability Zone: No Preference, Monitoring: Disabled, Instance Type: Large (m1.large), Instance Class: On Demand.
- CREATE KEY PAIR:** Key Pair Name: galaxy_keypair.
- CONFIGURE FIREWALL:** Security Group(s): default, galaxyWeb.

The wizard includes a 'Launch' button at the bottom right.

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 titled "Initial Cluster Configuration". It contains a welcome message: "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 message, 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 "OK" in green text next to it. A link "Show more startup options" is visible below the storage input.

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

In the background, the "Status" section is partially visible, showing fields for "Cluster name", "Disk status:", "Worker status", "Service status", and "External Logs".

Galaxy Cloud

+

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

Q

Google

Galaxy

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

Galaxy Cloudman Console

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

Terminate cluster

Add nodes ▼

Remove nodes

Access Galaxy

Status

Cluster name: ttt

Disk status: 0 / 0 (0%)

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

Service status: Applications Data

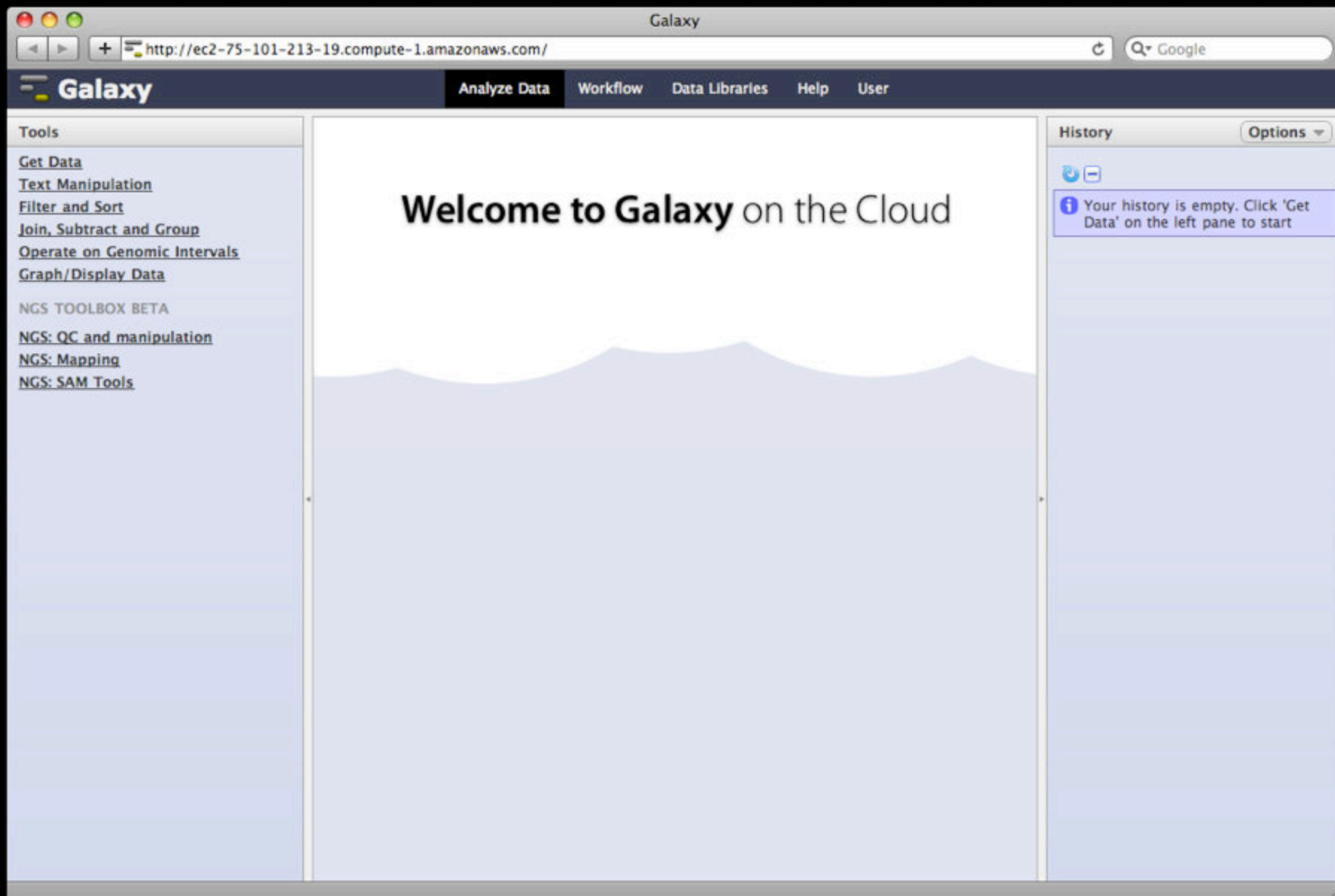
Pending

Starting

Ready

Error

Cluster status log



The image displays two screenshots of the Galaxy Cloud interface. The left screenshot shows the 'Saved Histories' table, which lists various history entries with their respective dataset counts and creation times. The right screenshot shows the 'Galaxy Cloud Console', which provides a high-level overview of the cluster's status and offers controls for scaling and terminating the instance.

Name	Datasets (by state)	Tags	Sharing	Created	Last Updated	
mt replicates pair 1	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	21 min ago
mt datasets	24	0 Tags		about 2 hours ago	abo	

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-a3e7b2c8' 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
14:54:56 - 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

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 Problem

You have written a Perl 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

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



```
cluster.xml
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 -l $input1_chromCol,$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
48
49  Line: 87 Column: 8 XML
```


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 SAM-to-BAM on data 1

STS Markers on GenBank and Radiation Hybrid Maps

STS Markers

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

Pax9 Slc25a21

RefSeq Genes

Other RefSeq

Ensembl Gene Predictions

Human Proteins Mapped by Chained tBLASTn

Mouse mRNAs from GenBank

Spliced ESTs

Mouse ESTs That Have Been Spliced

36-way Multiz Alignment & Conservation

Mammal Cons

Rat Human Orangutan Dog Horse Opossum Chicken Stickleback

Simple Nucleotide Polymorphisms (dbSNP build 126)

Repeating Elements by RepeatMasker

move start move end

< 2.0 >

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](#)

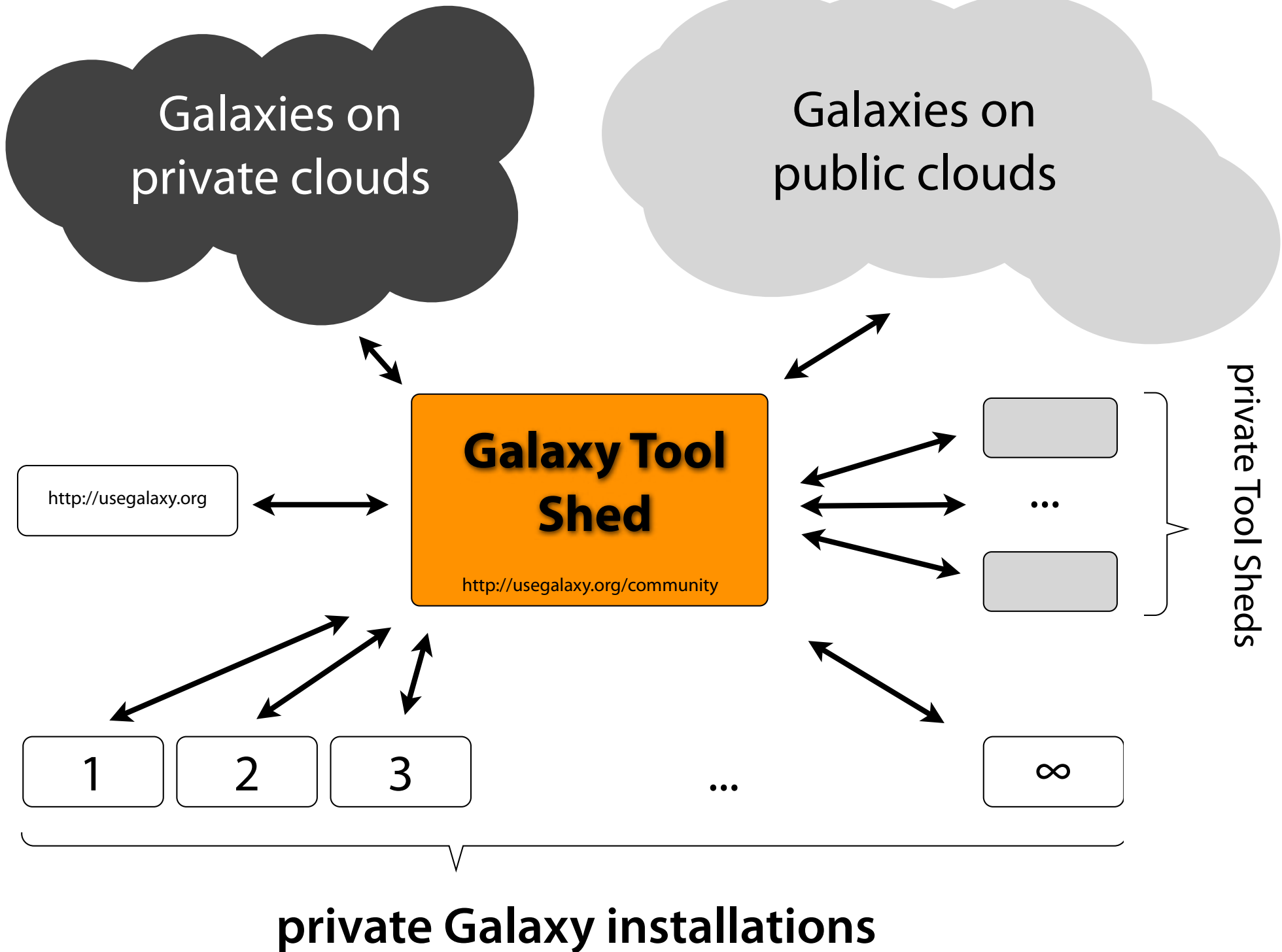
Binary bam alignments file

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)**


ToolsHelpUser

Community

Tools

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

Categories

 [Advanced Search](#)

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

<http://usegalaxy.org/community>

Some future challenges

- Capturing and automatically deploying tool dependencies, automatic tool acquisition in Galaxy instances
- Better interfaces for highly parallel analysis (e.g. running the same workflow across 192 individuals)
- Various workflow engine improvements, partial data streaming, combined experimental/computational workflows

Try it now:

<http://usegalaxy.org>

Develop and deploy:

<http://getgalaxy.org>

<http://galaxyproject.org>

Come do cool stuff, contact us at:

[http://wiki.g2.bx.psu.edu/News/Galaxy is Hiring](http://wiki.g2.bx.psu.edu/News/Galaxy%20is%20Hiring)

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



EMORY

PENNSTATE.



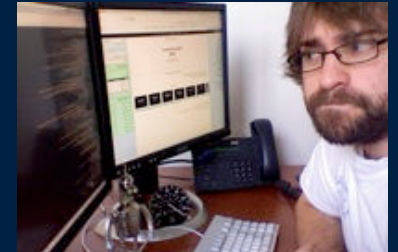
Enis Afgan



Dannon Baker



Dan Blankenberg



Nate Coraor



Dave Clements



Jeremy Goecks



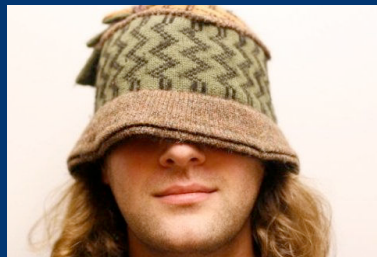
Jennifer Jackson



Greg von Kuster



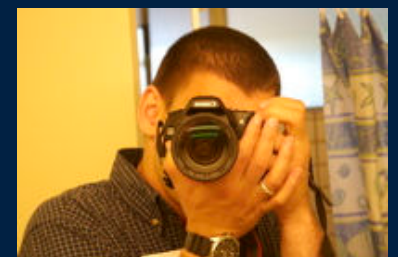
Kanwei Li



James Taylor



Guru Ananda



Anton Nekrutenko

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