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

Galaxy 101 Exercise

The Vision

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

What is Galaxy?

GUI for genomics

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

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

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

Session Overview

What is Galaxy?

What you can do in Galaxy

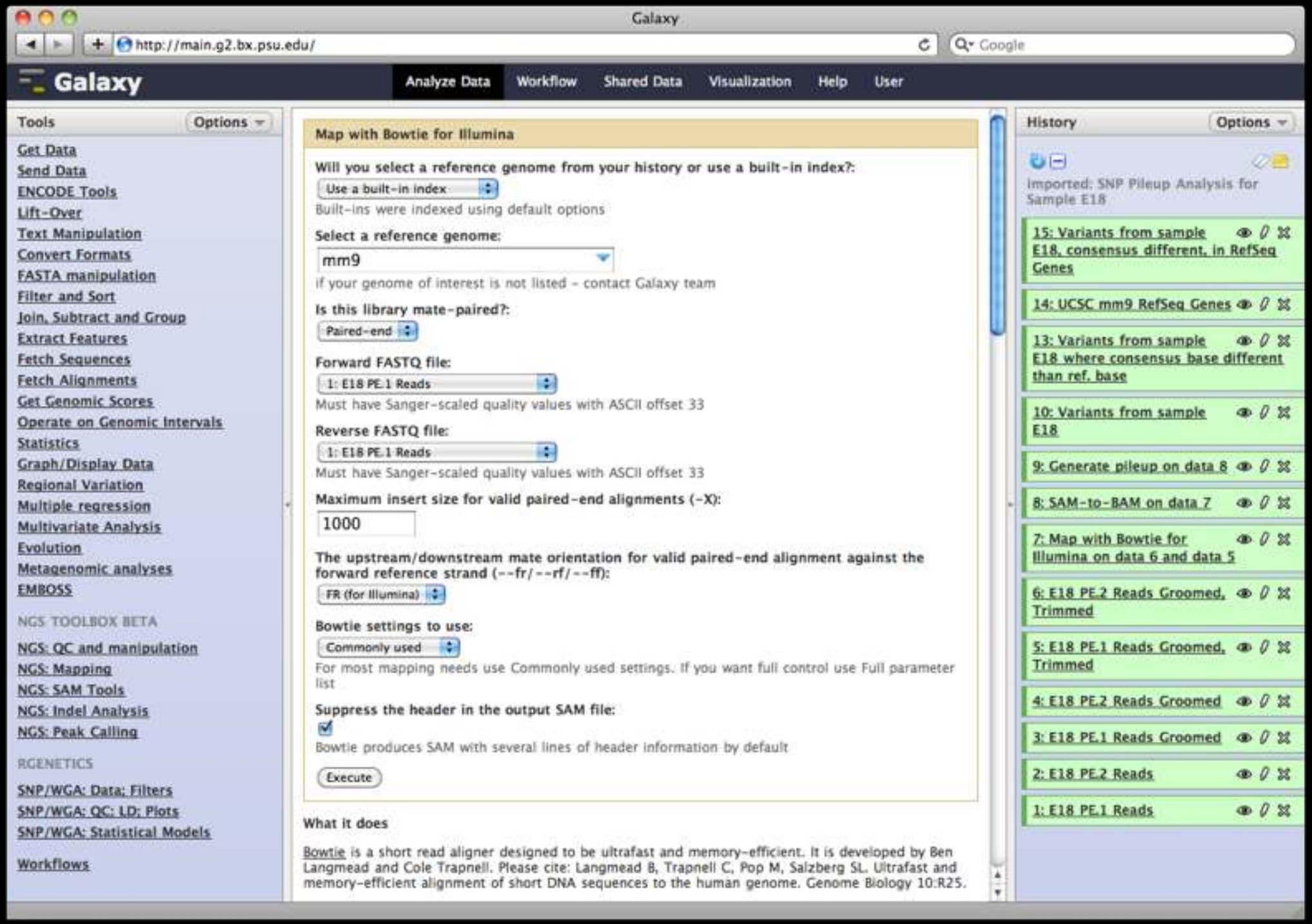
- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

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

Galaxy 101 Exercise

Galaxy Analysis Workspace



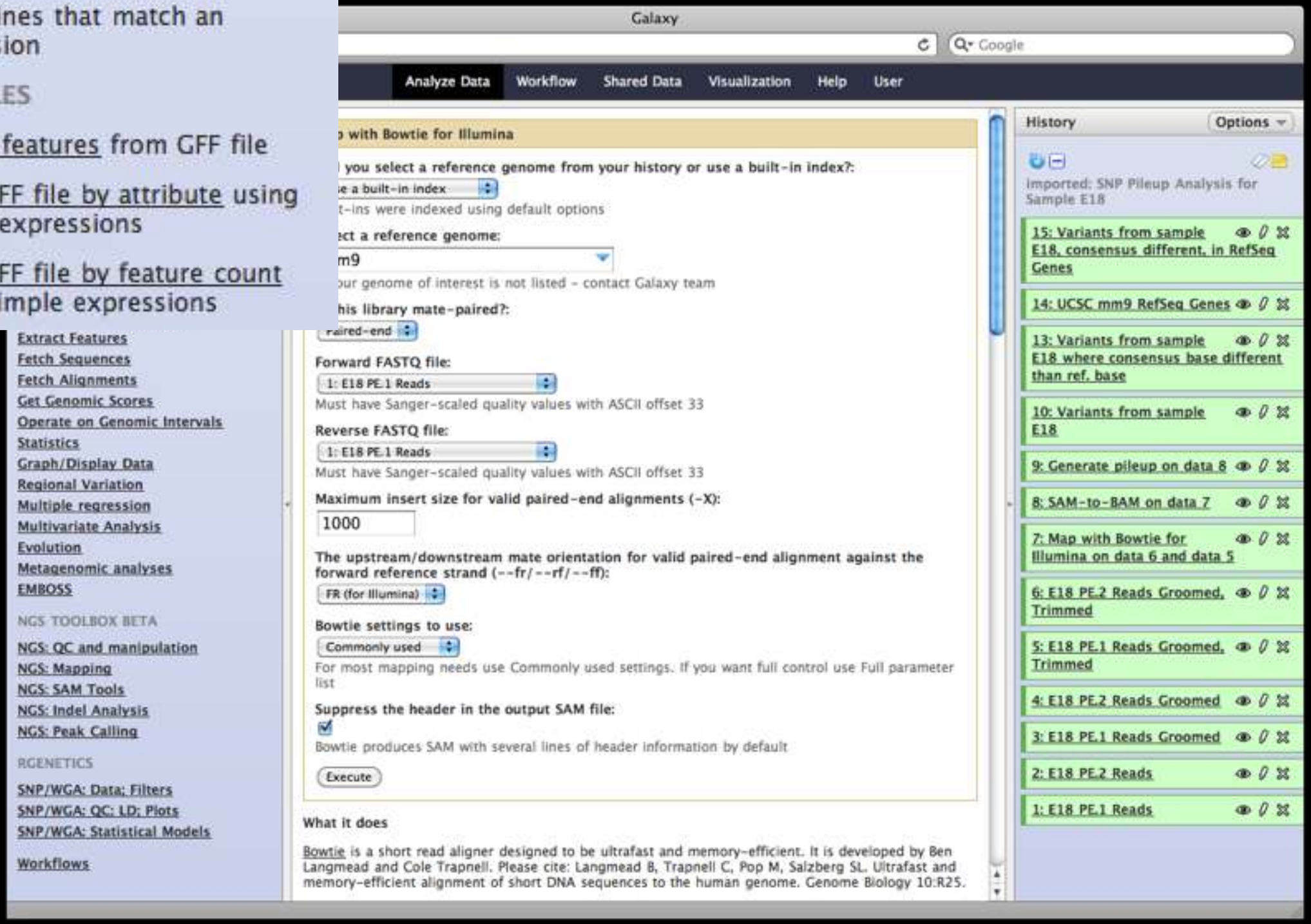
Filter and Sort

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

GFF FILES

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

xy Analysis Workspace



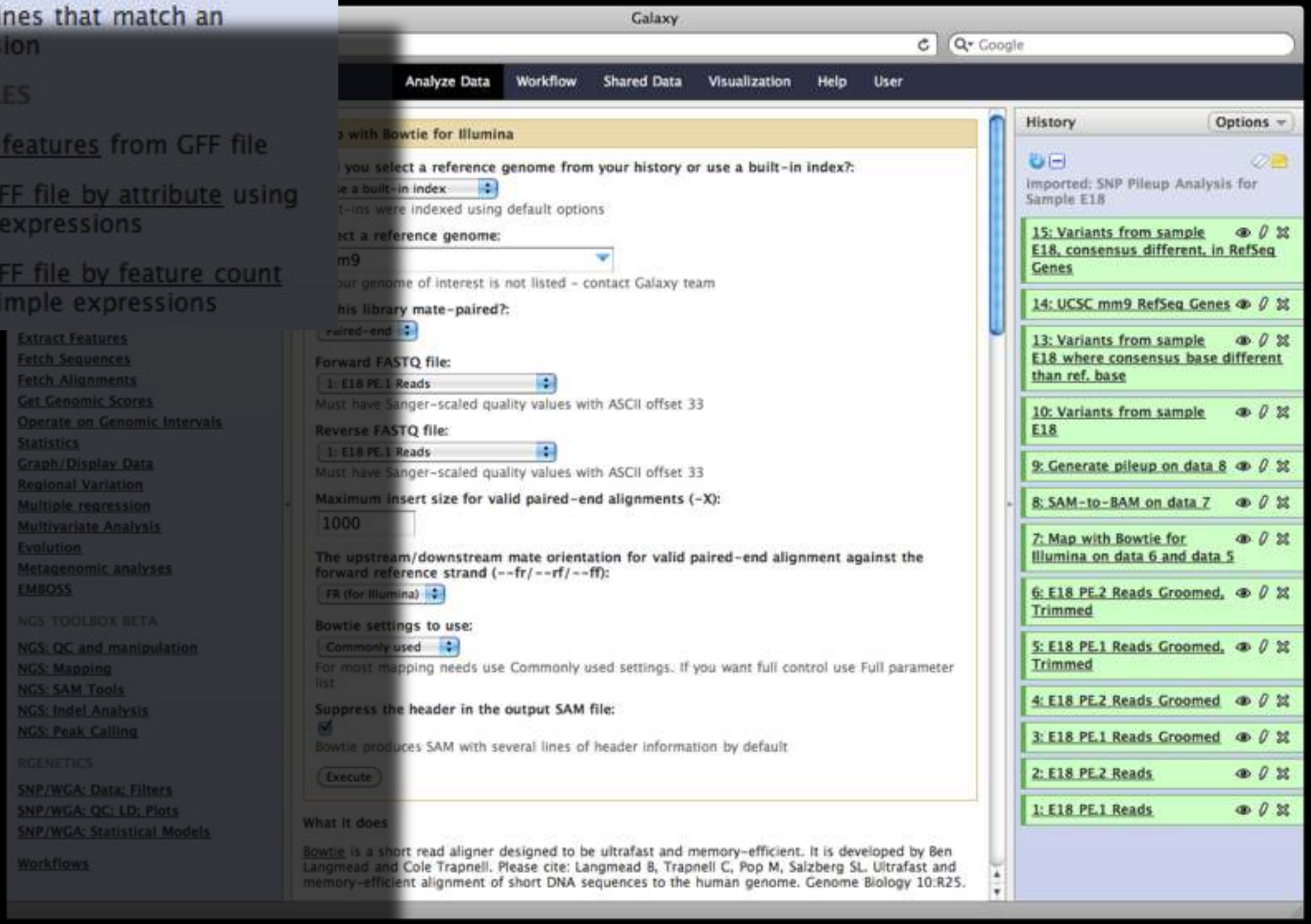
xy Analysis Workspace

Filter and Sort

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

GFF FILES

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



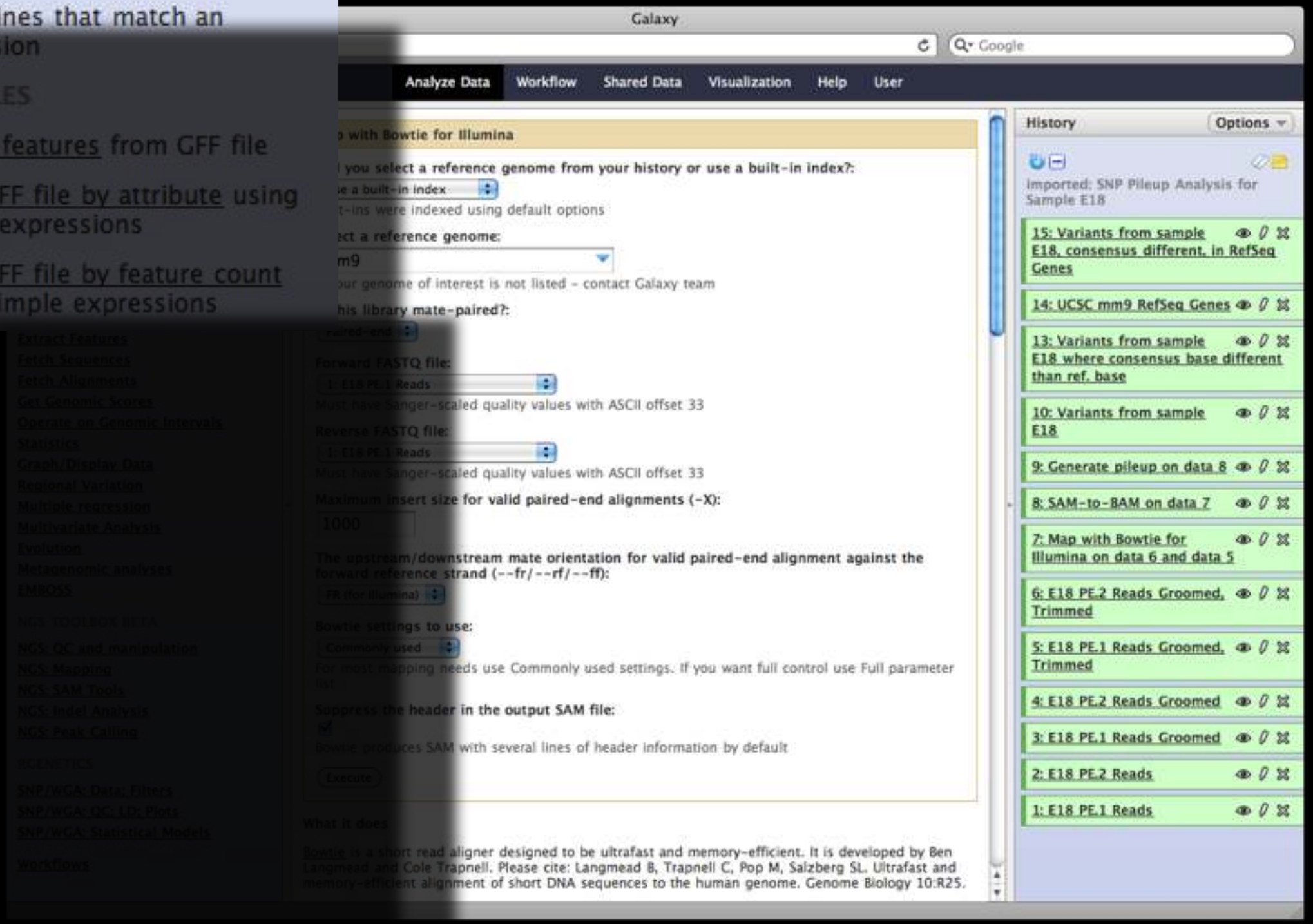
Filter and Sort

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

GFF FILES

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

xy Analysis Workspace



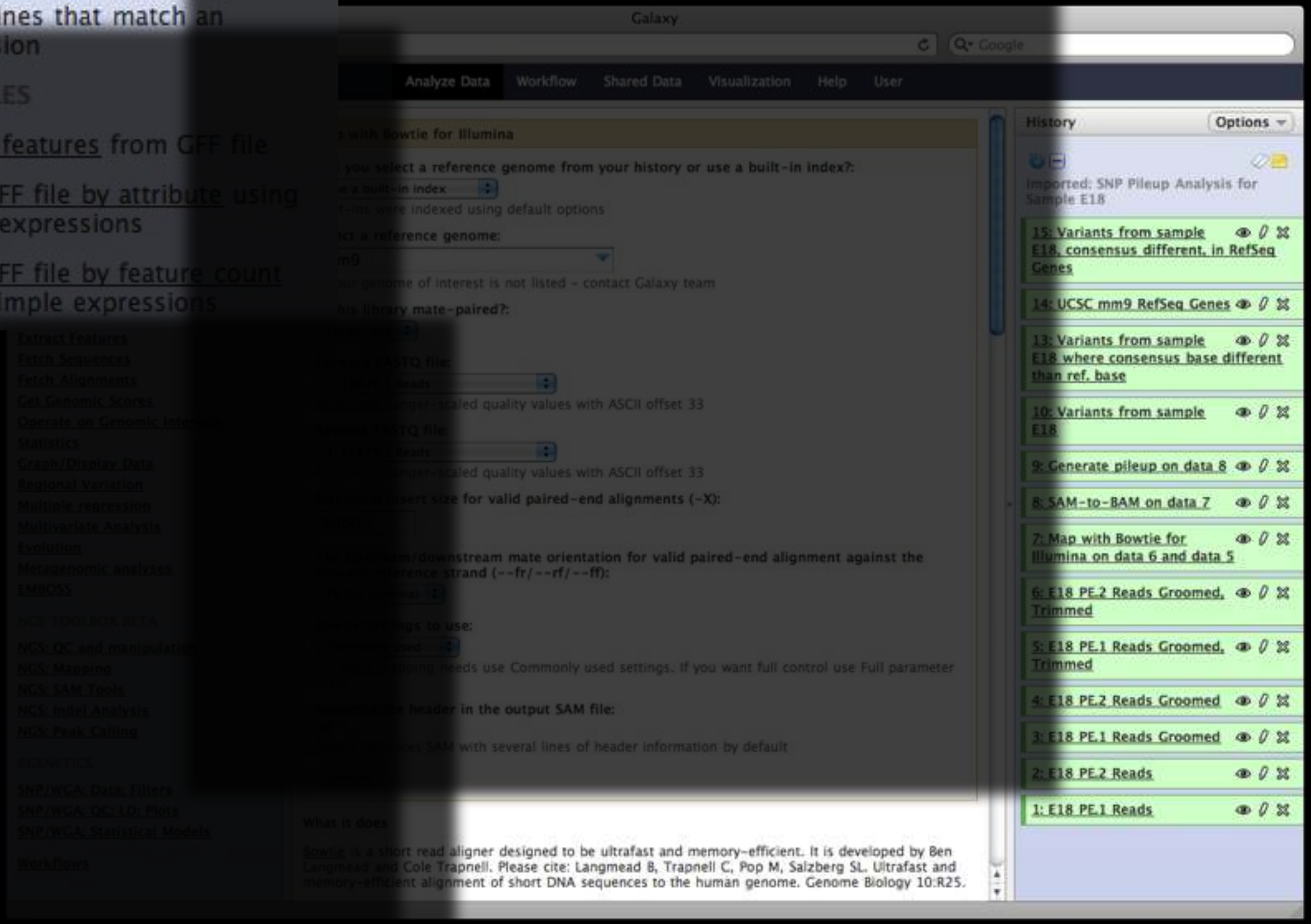
xy Analysis Workspace

Filter and Sort

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

GFF FILES

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



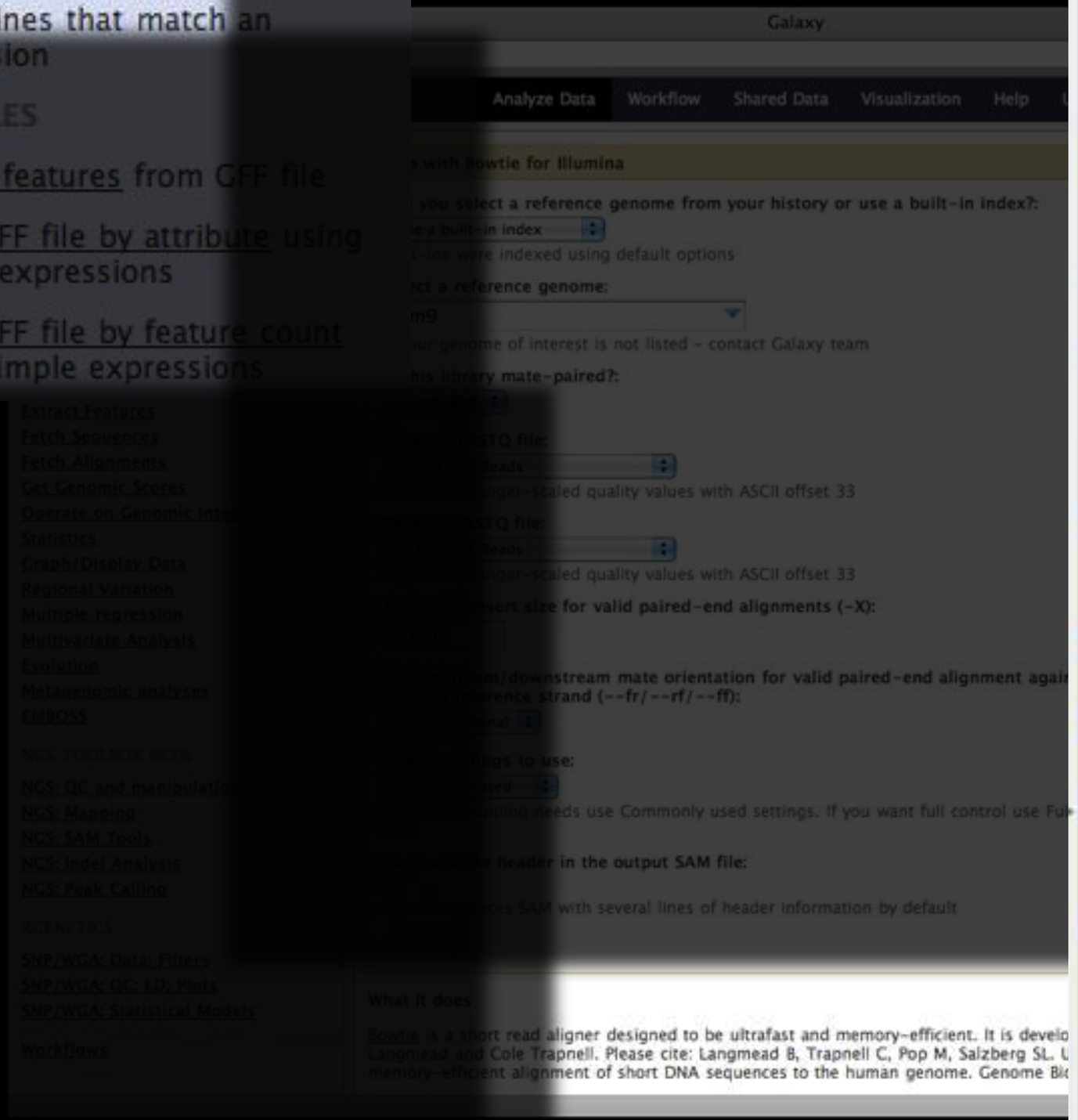
Filter and Sort

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

GFF FILES

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

xy Analysis Wor



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

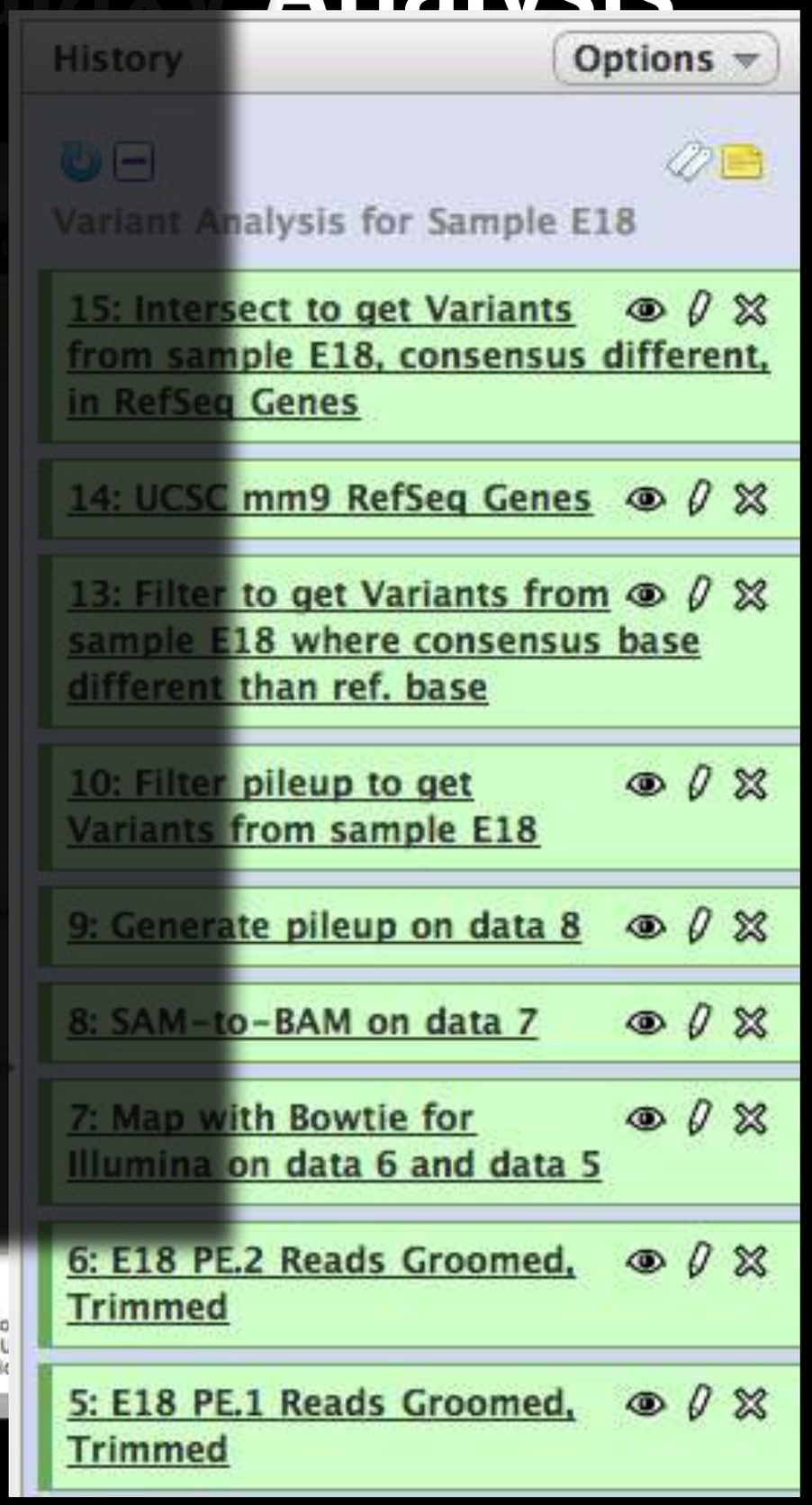
g Users with Galaxy Analysis Workspace

Filter and Sort

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

GFF FILES

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



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 in KB1 and other personal genomes ▼		greg@bx.psu.edu	2010-02-15	2.1 Mb
indels in ABT ▼		greg@bx.psu.edu	2010-02-03	105.3 Kb
indels in KB1 ▼		greg@bx.psu.edu	2010-02-03	14.2 Mb
indels in MD6 ▼		greg@bx.psu.edu	2010-02-03	109.8 Kb
indels in NB1 ▼		greg@bx.psu.edu	2010-02-03	51[illegible] Kb
indels in TK1 ▼		greg@bx.psu.edu	2010-02-03	123.2 Kb
novel SNPs in ABT ▼		greg@bx.psu.edu	2010-02-09	9.4 Mb
novel SNPs in KB1 ▼		greg@bx.psu.edu	2010-02-09	16.9 Mb
novel SNPs in 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: Import into your current history Go

http://usegalaxy.org/bushman

Managing Libraries

Loading Data

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

Accessing Data

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

Annotating Library Data: Library Templates

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

Overview

What is Galaxy?

What you can do in Galaxy

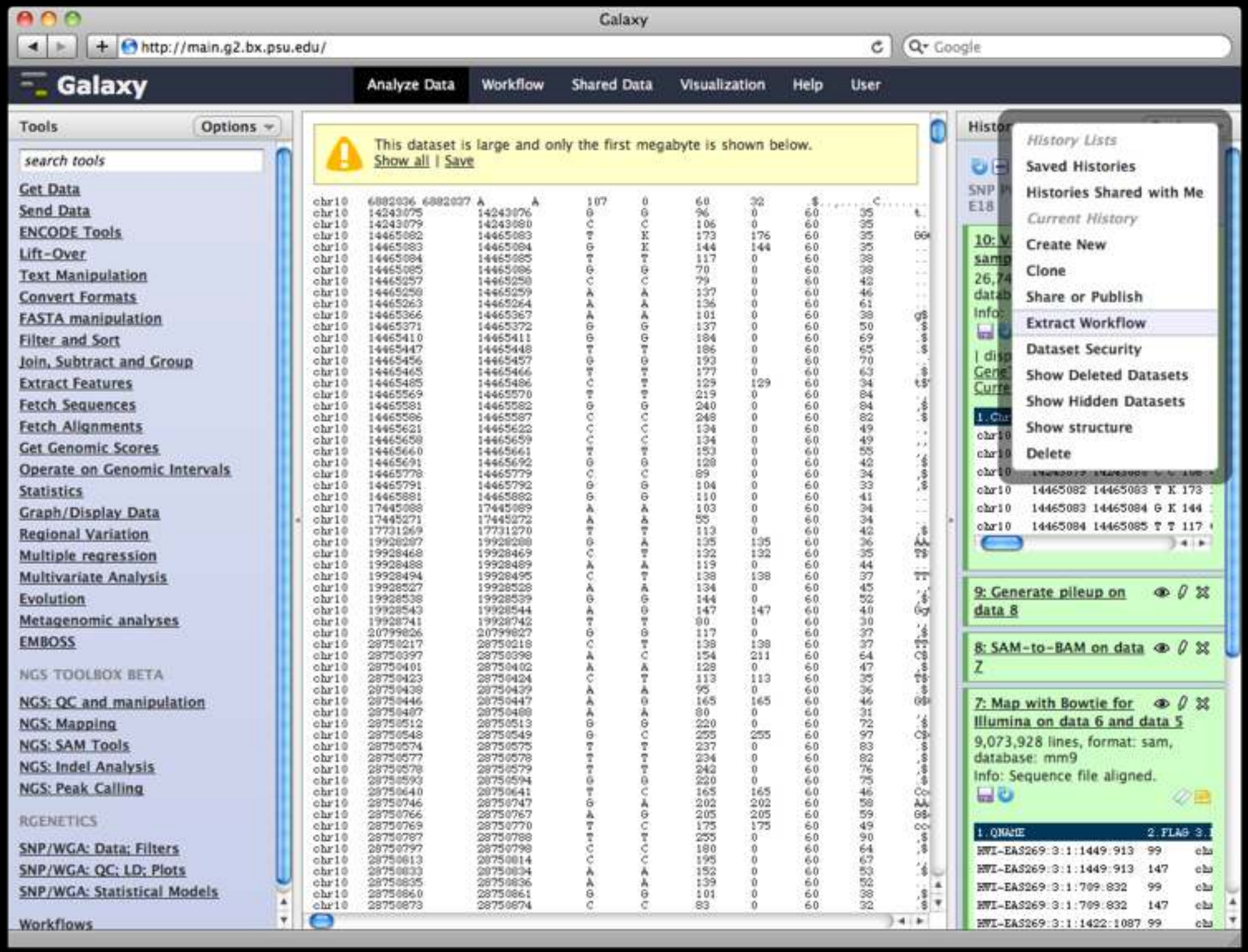
- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

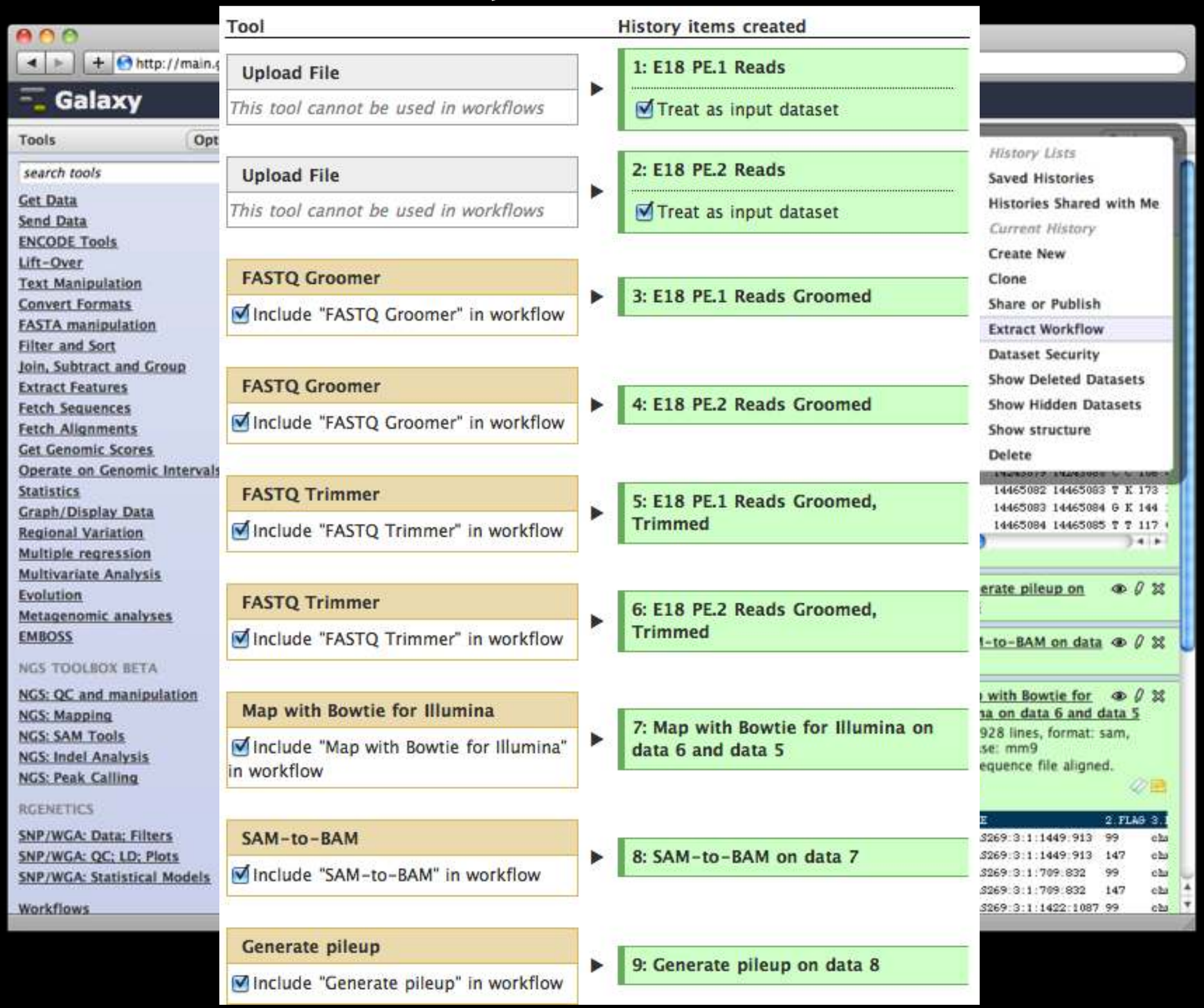
- public website
- local instance
- on the cloud
- tool shed/contributing tools

Galaxy 101 Exercise

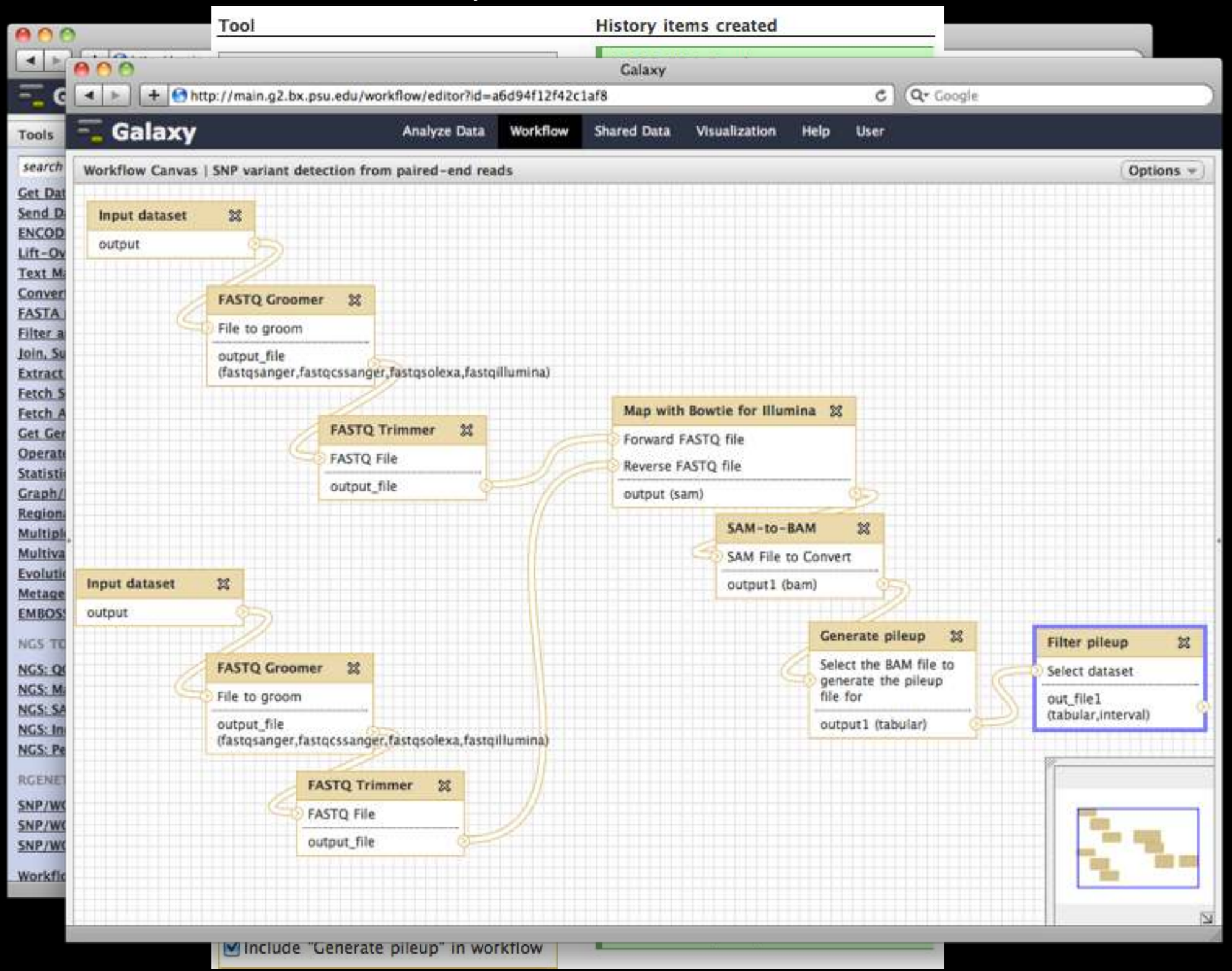
Galaxy Workflows



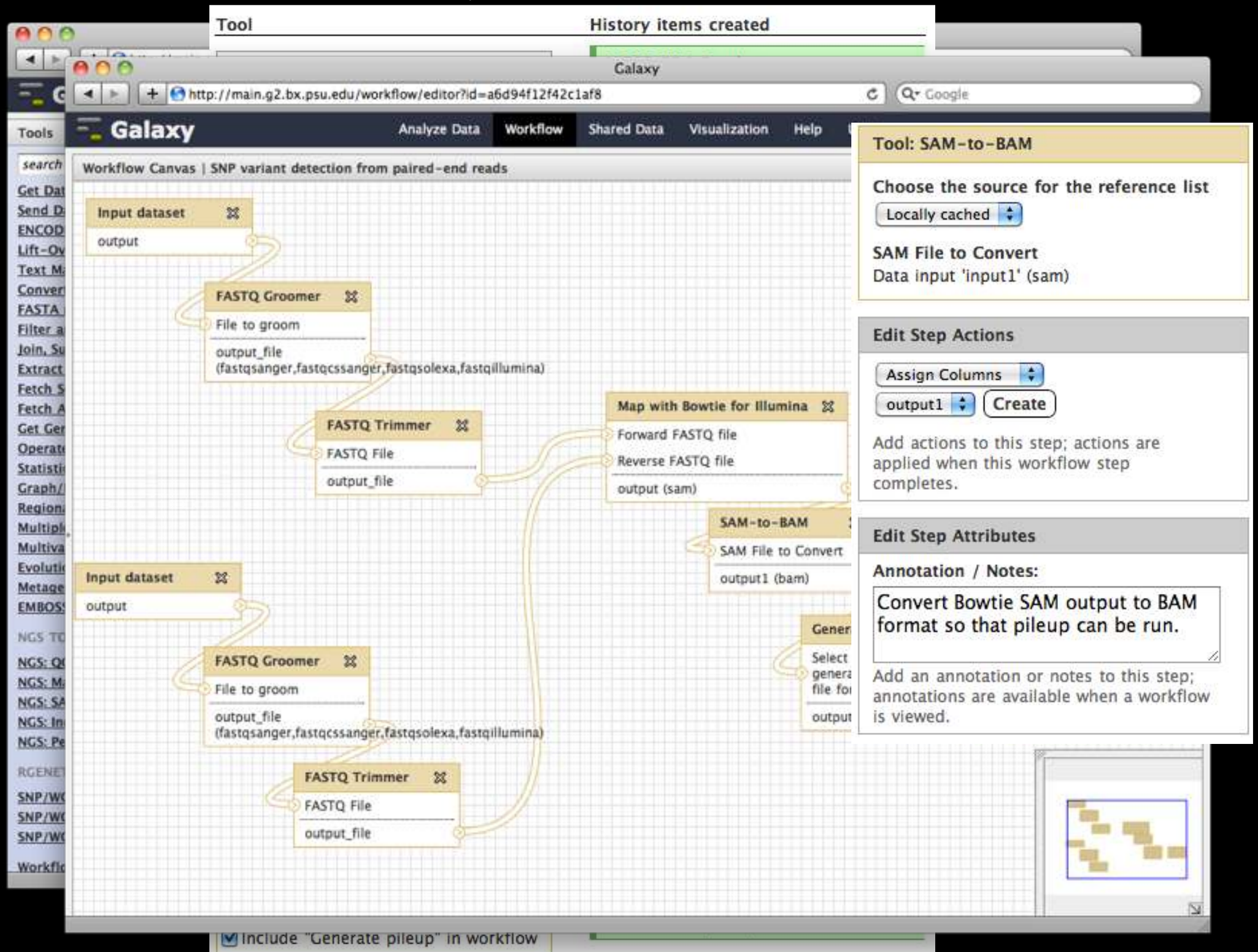
Galaxy Workflows



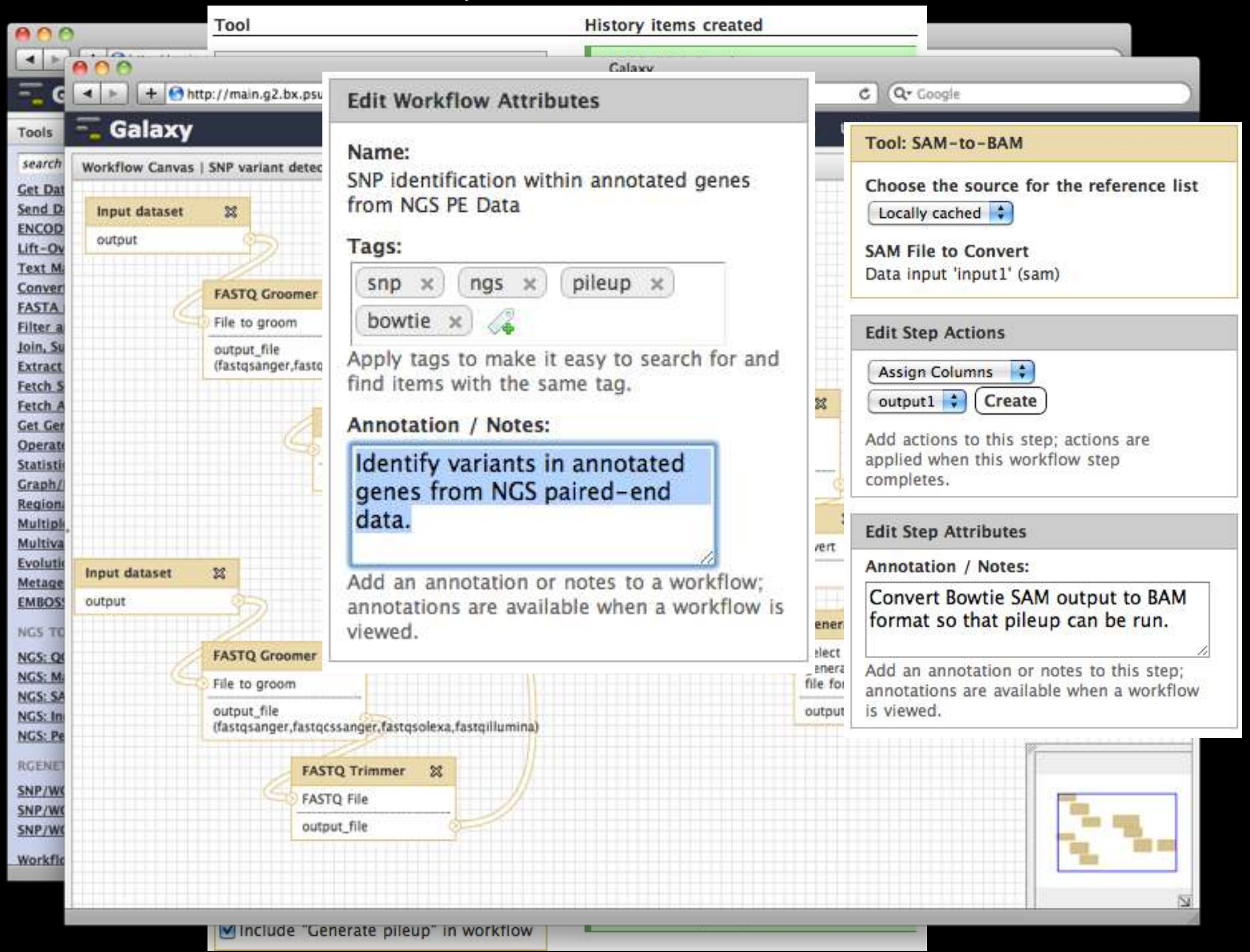
Galaxy Workflows

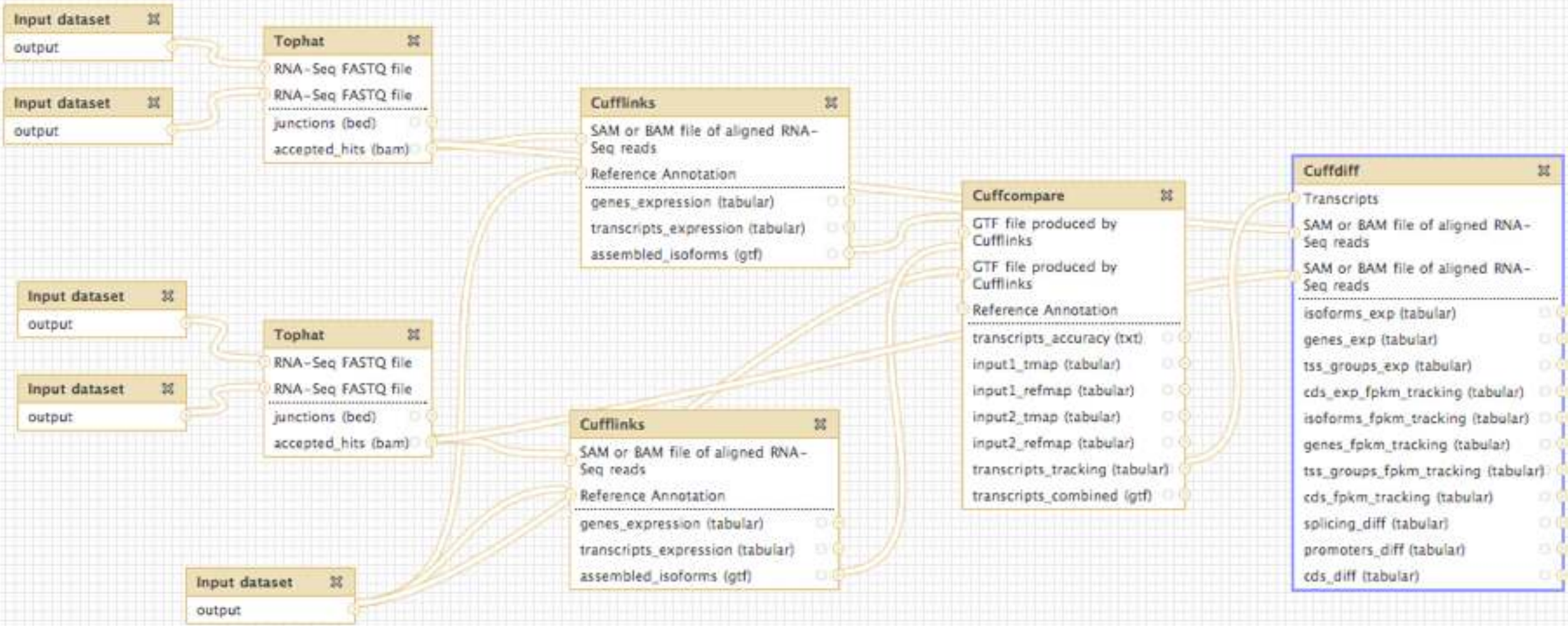


Galaxy Workflows

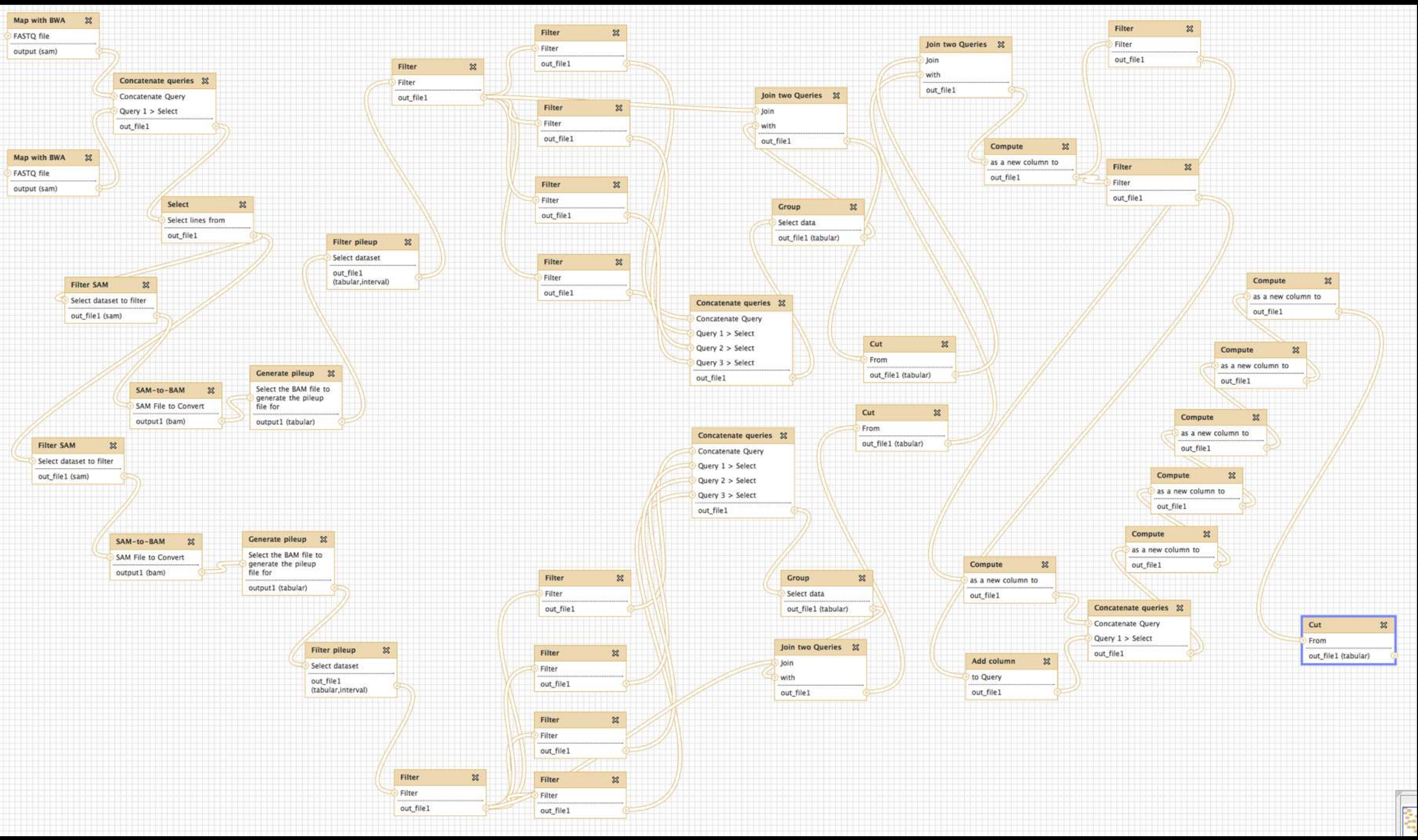


Galaxy Workflows





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



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

Overview

What is Galaxy?

What you can do in Galaxy

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

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

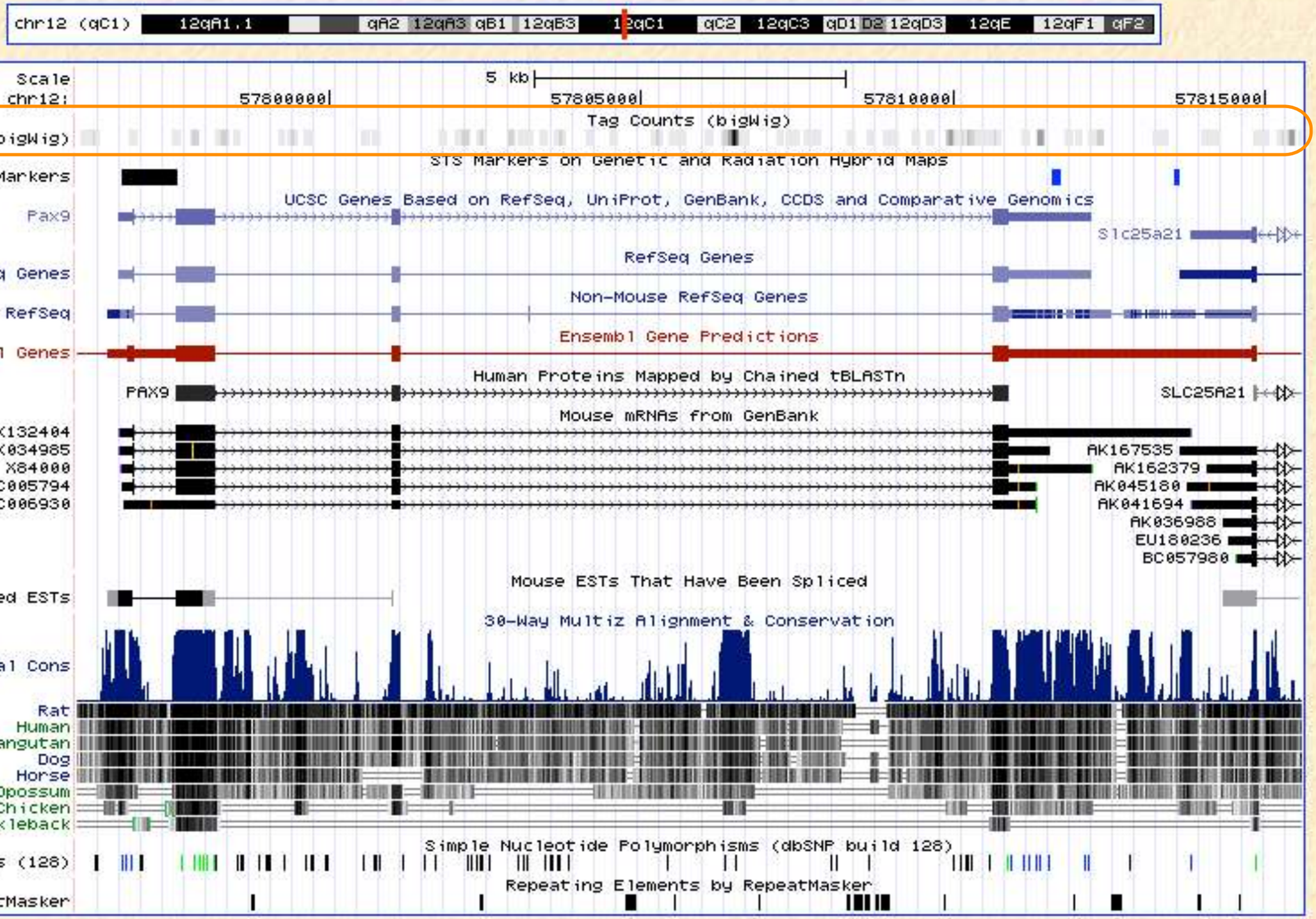
IGV

UCSC Genome Browser on Mouse July 2007 (NCBI37/

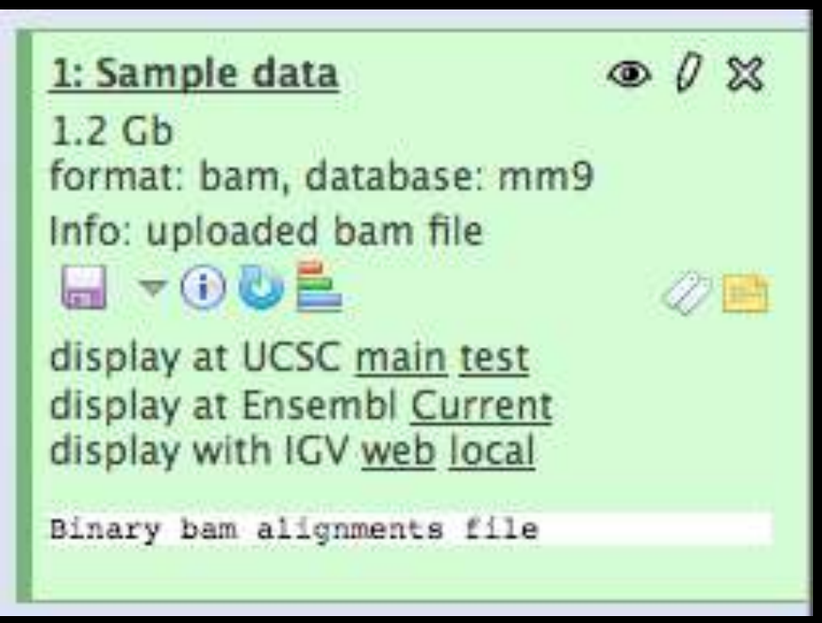
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom

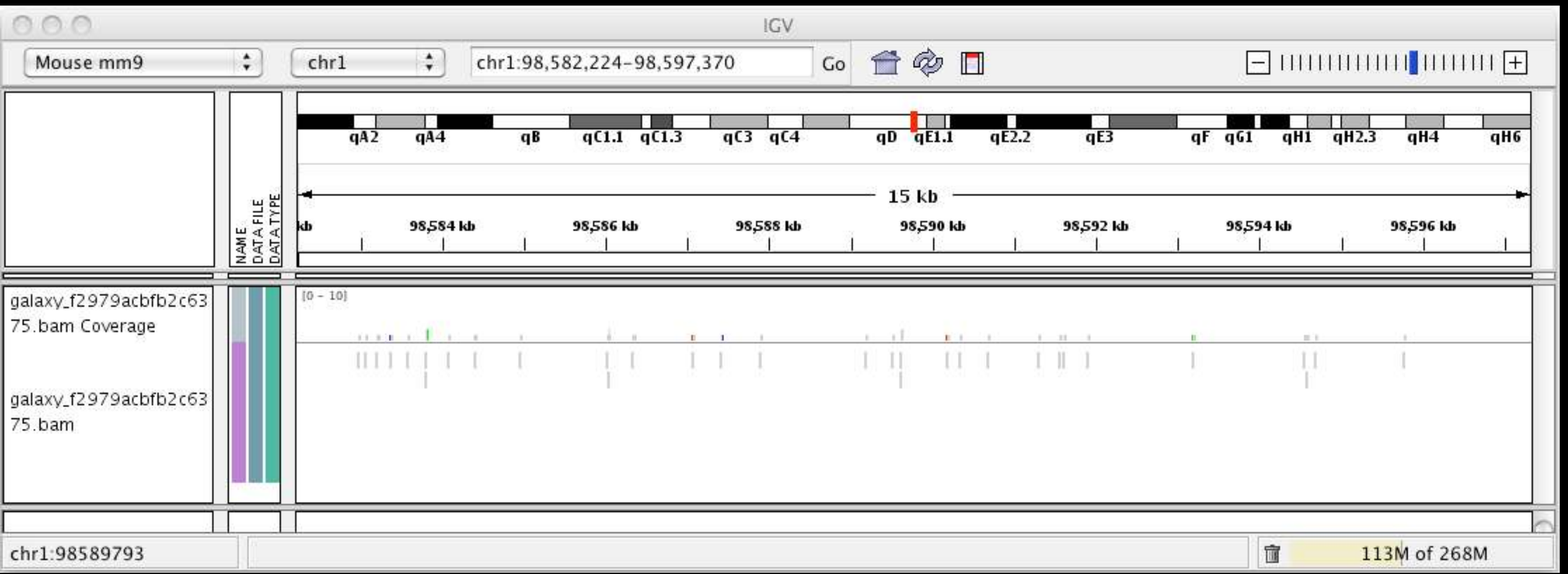
position/search chr12:57,795,963-57,815,592 gene jump clear

14: Tag Counts (bigWig)
2.4 Gb, format: bigwig, database: mm9
Info:
| display at UCSC main
Binary UCSC BigWig file



Integrative Genomics Viewer (IGV)





Galaxy

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

Genome Browser

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

Trackster

Trackster

View your data from within Galaxy

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

Supports common filetypes

- BAM, BED, GFF/GTF, WIG

Unique features

- custom genomes
- highly interactive

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

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

Google

Galaxy

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

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

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

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

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

11431

UCSC Main on Human: phyloP46wayPrimates (chr19) — Histogram

1

-1

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

8732

h1-hESC Cufflinks assembled transcripts — Auto (Squish)

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

Display a menu

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

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

Google

Galaxy

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

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

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

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

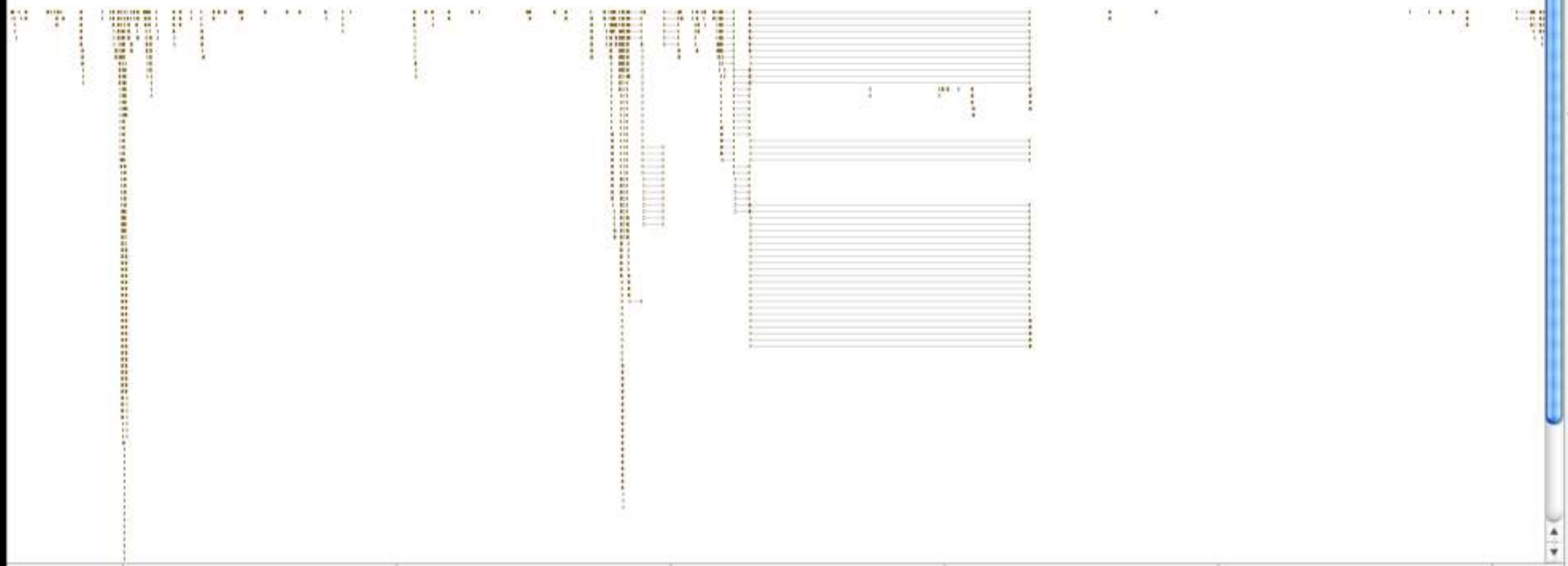
UCSC Main on Human: all_est (chr19) Dense

UCSC Main on Human: phyloP46wayPrimates (chr19) Histogram

1

-1

h1-hESC Tophat Mapped Reads Auto (Squish)



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

Display a menu

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

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

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

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 663,032 - 663,110

30 663,040 663,050 663,060 663,070 663,080 663,090 663,100

g g c c c g g g c c T C A C C G G C A G G C G C G G G A C G A T C T C C A C G G A G C A G C A G T G G C A G A A G T A C C G T C C G G G A T G C G G C G A C

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

UCSC Main on Human: all_est (chr19) Dense

UCSC Main on Human: phyloP46wayPrimates (chr19) Histogram

1

-1

h1-hESC Tophat Mapped Reads Auto (Pack)

HWI-EAS229_1:2:7:78:211

HWI-EAS229_1:2:39:1023:1868

HWI-EAS229_1:2:83:31:1502

HWI-EAS229_1:2:30:329:1847

h1-hESC Cufflinks assembled transcripts Auto (Pack)

30 663,040 663,050 663,060 663,070 663,080 663,090 663,100

Canceled opening the page

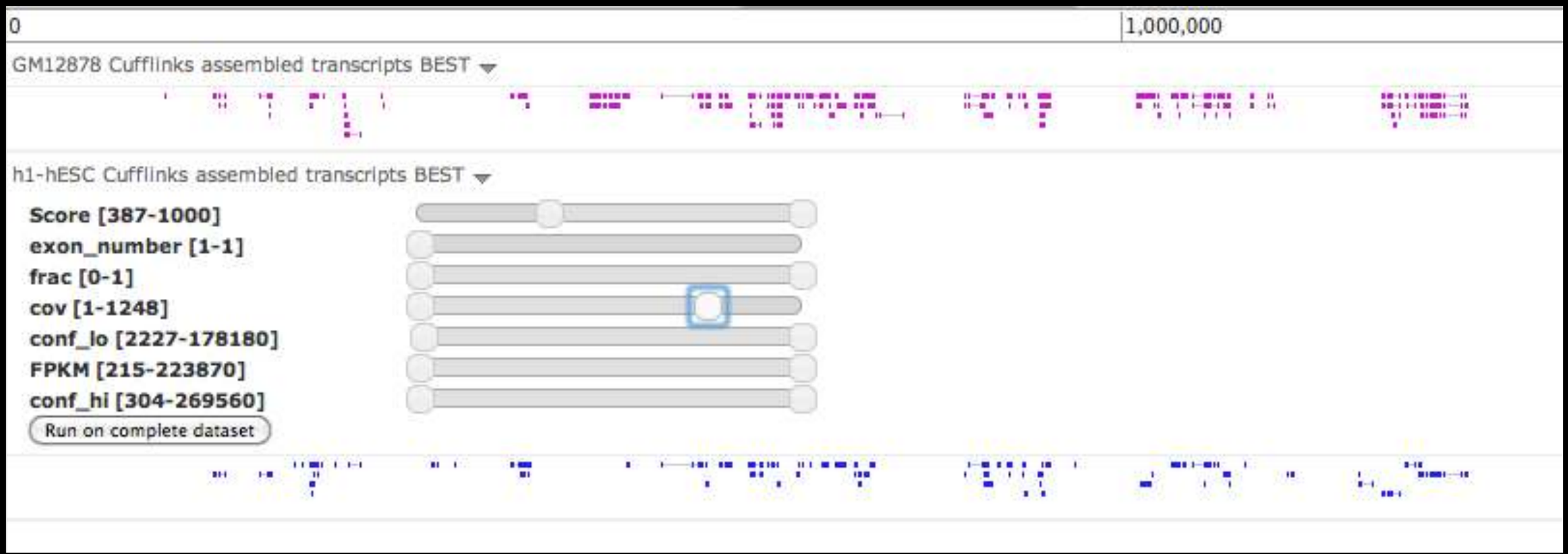
But really, why *another* genome browser

From static browsing to visual analysis

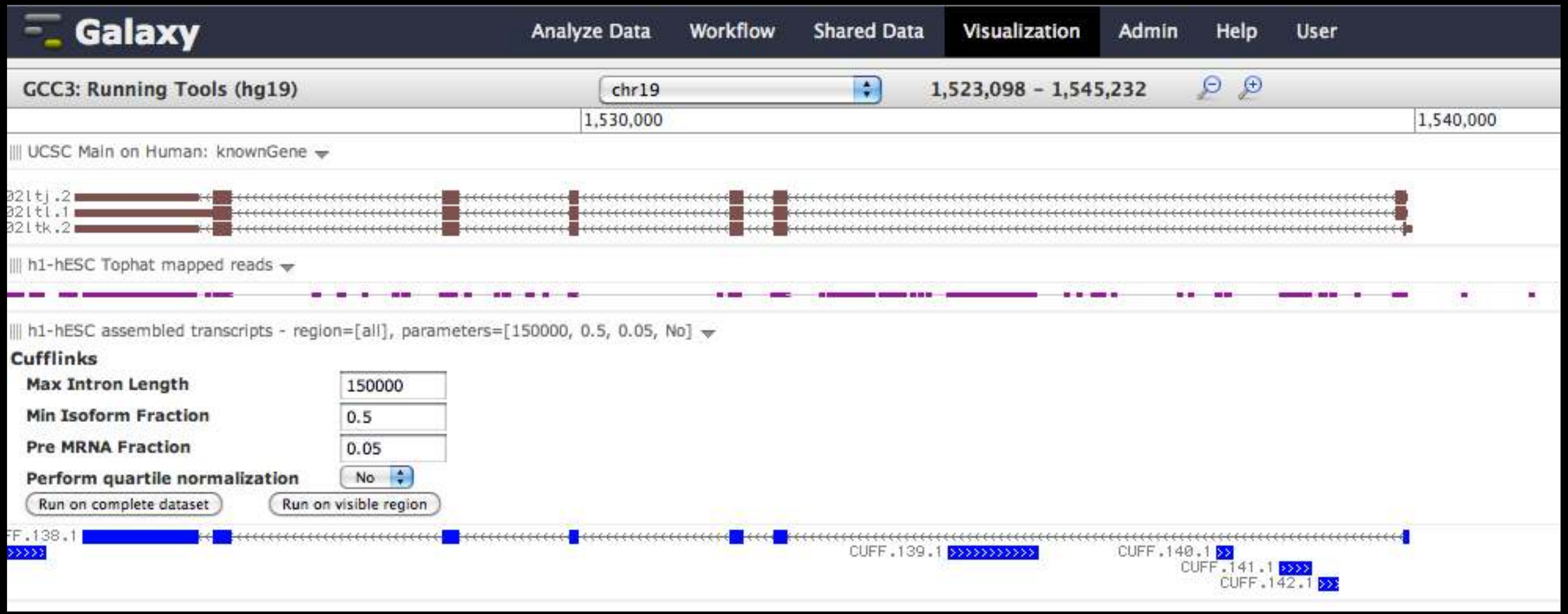
Visual feedback and experimentation needed for complex tools with many parameters

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

Dynamic Filtering



Integrating Tools and Visualization

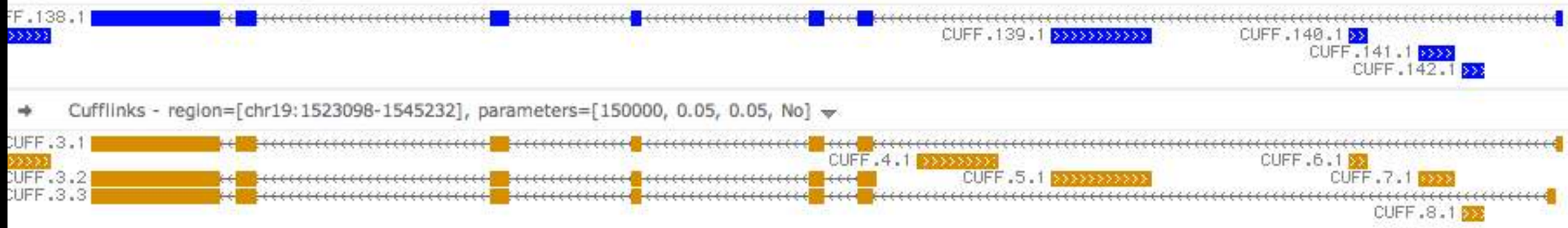


|||| h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No]

Cufflinks

Max Intron Length	150000
Min Isoform Fraction	0.05
Pre MRNA Fraction	0.05
Perform quartile normalization	No

Run on complete dataset Run on visible region



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

1,530,000 1,540,

UCSC Main on Human: knownGene

uc002ltj.2
uc002ltl.1
uc002ltk.2

h1-hESC Tophat mapped reads

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

Cufflinks

Max Intron Length	150000
Min Isoform Fraction	0.05
Pre MRNA Fraction	0.001
Perform quartile normalization	No

Run on complete dataset Run on visible region

CUFF.138.1
26.1
CUFF.139.1
CUFF.140.1
CUFF.141.1
CUFF.142.1

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

CUFF.3.1
2.1
CUFF.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
CUFF.8.1

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

CUFF.3.1
2.1
CUFF.3.2
CUFF.3.3
CUFF.4.1
CUFF.5.1
CUFF.6.1
CUFF.7.1
CUFF.8.1

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

Galaxy 101 Exercise

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

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

Make History Accessible via Link

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

Make History Accessible and Publish

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

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

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

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

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

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

You can:

Unpublish History

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

Disable Access to History via Link and Unpublish

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

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Back to Histories List

Galaxy | Published History | Variant Analysis for Sample E18

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

Google

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

Published Histories | jgoecks | Variant Analysis for Sample E18

Galaxy History ' Variant Analysis for Sample E18'

Import history

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

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

About this History

Author

jgoecks

Related Histories

All published histories
Published histories by jgoecks

Rating

Community
(1 rating, 4.0 average)

Yours

Tags

Community:
snp | pileup | bowtie | demo | sample

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

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

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

Google

Galaxy

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

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

Step 6: FASTQ Trimmer

Trim reads to remove low-quality bases.

FASTQ File
Output dataset 'output_file' from step 4

Define Base Offsets as
Absolute Values

Offset from 5' end
0

Offset from 3' end
9

Keep reads with zero length
False

Step 7: Map with Bowtie for Illumina

Map reads using default parameter values.

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

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

Is this library mate-paired?
Paired-end

Forward FASTQ file
Output dataset 'output_file' from step 6

Reverse FASTQ file
Output dataset 'output_file' from step 5

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

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

Bowtie settings to use
Commonly used

Suppress the header in the output SAM file
True

Step 8: SAM-to-BAM

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

Choose the source for the reference list
Locally cached

About this Workflow

Author

jgoecks

Related Workflows

All published workflows
Published workflows by jgoecks

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community:
snp | bowtie

Yours:
snp | bowtie

Galaxy | Published Histories

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

Google

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

Published Histories

search | Advanced Search

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

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

Sharing Trackster Visualizations

"A picture is worth a 1000 words."

A fully-interactive visualization is worth many more words

Galaxy | Published Visualiz

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

Galaxy

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

Published Visualizations | jeremy | GCC2011-1: Viewin

chr19 1,290 - 4,168,475

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

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

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

11431

UCSC Main on Human: phyloP46wayPrimates (chr19) — Histogram

1

-1

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

8732

h1-hESC Cufflinks assembled transcripts — Auto (Squish)

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

About this Visualization

Author

jeremy

Related Visualizations

All published visualizations
Published visualizations by jeremy

Rating

Community
(1 rating, 5.0 average)

Yours

Tags

Community:
cool

Yours:

Overview

What is Galaxy?

What you can do in Galaxy

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

Where you can use and build Galaxy

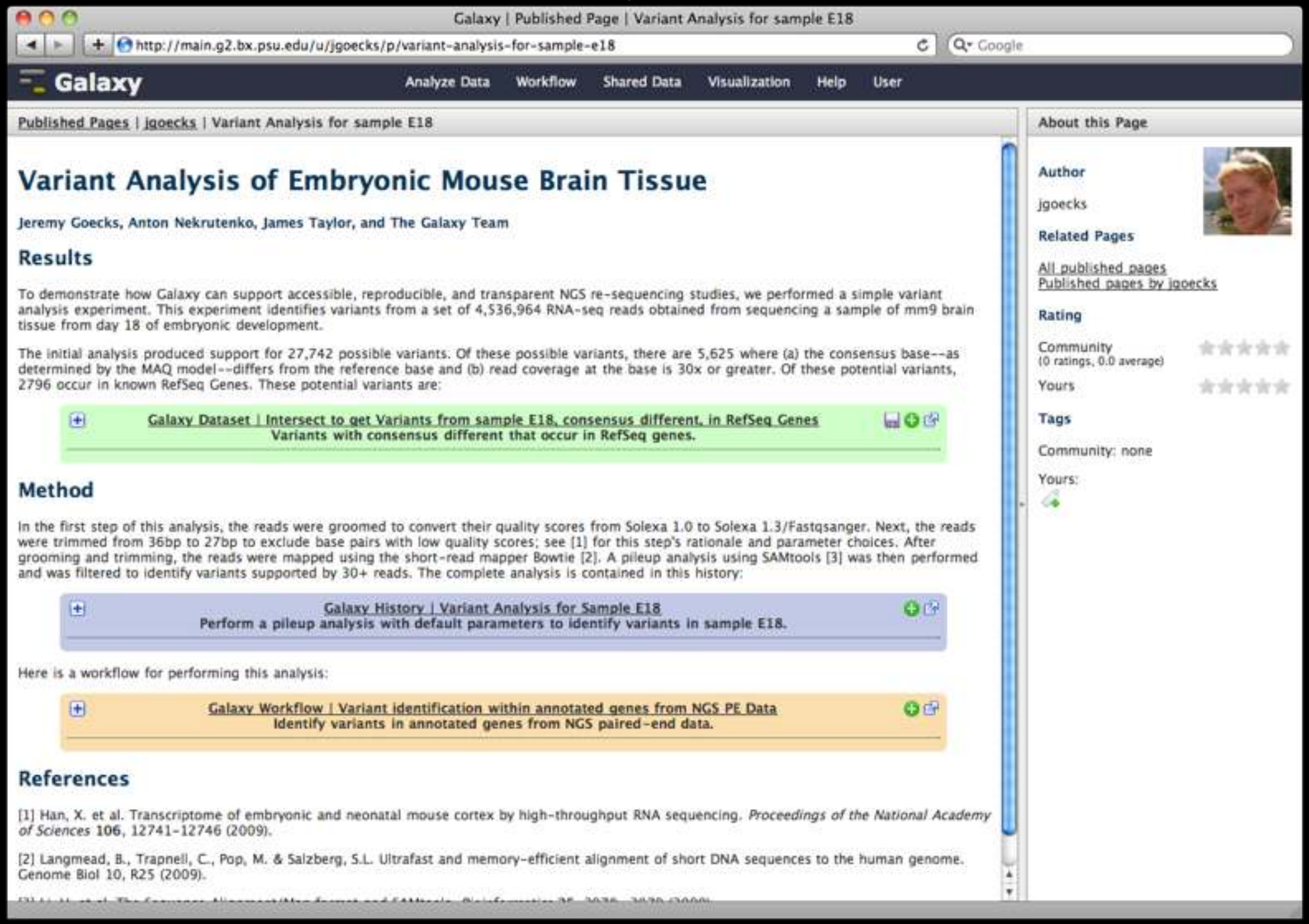
- public website
- local instance
- on the cloud
- tool shed/contributing tools

Galaxy 101 Exercise

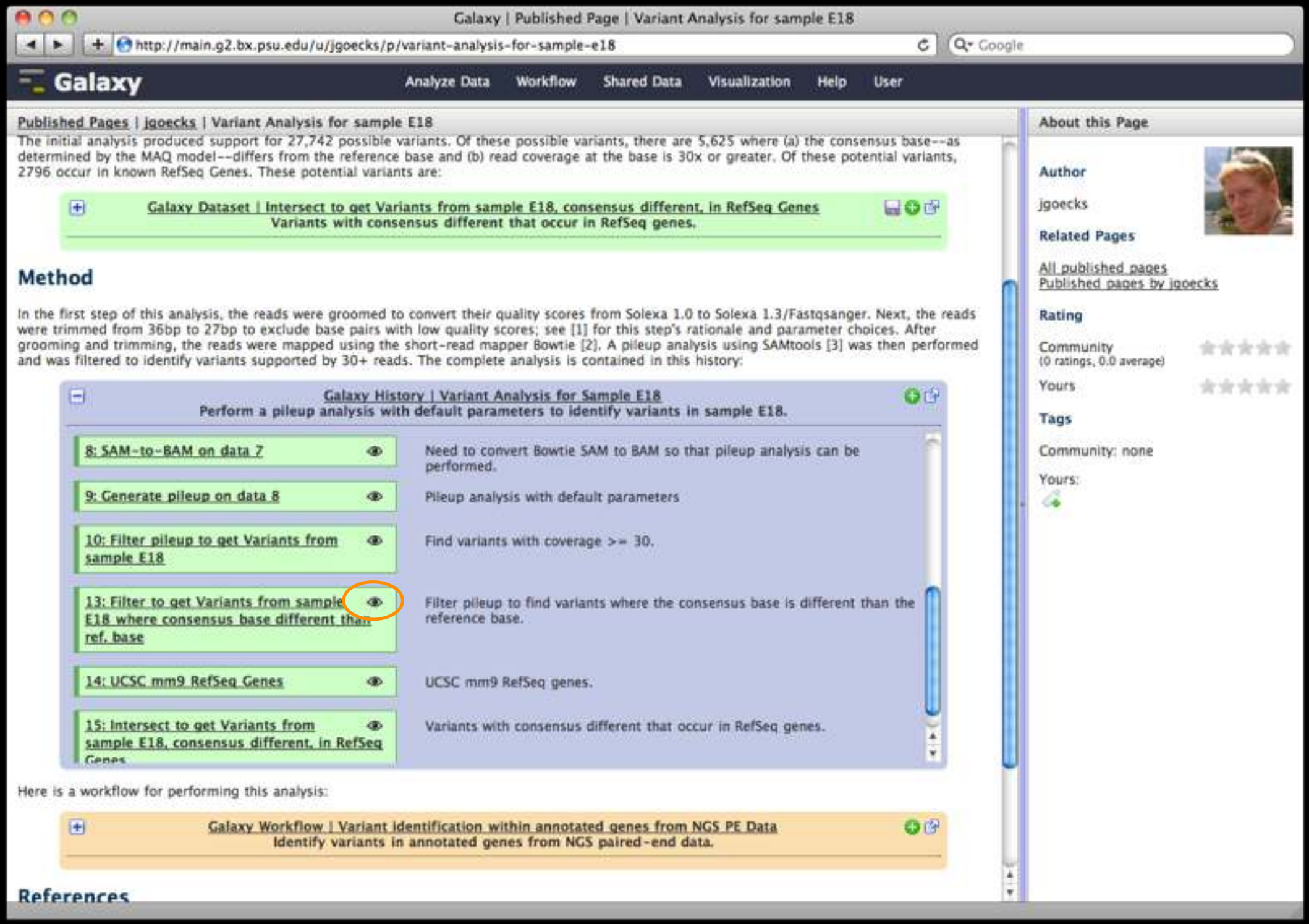
Galaxy Pages

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

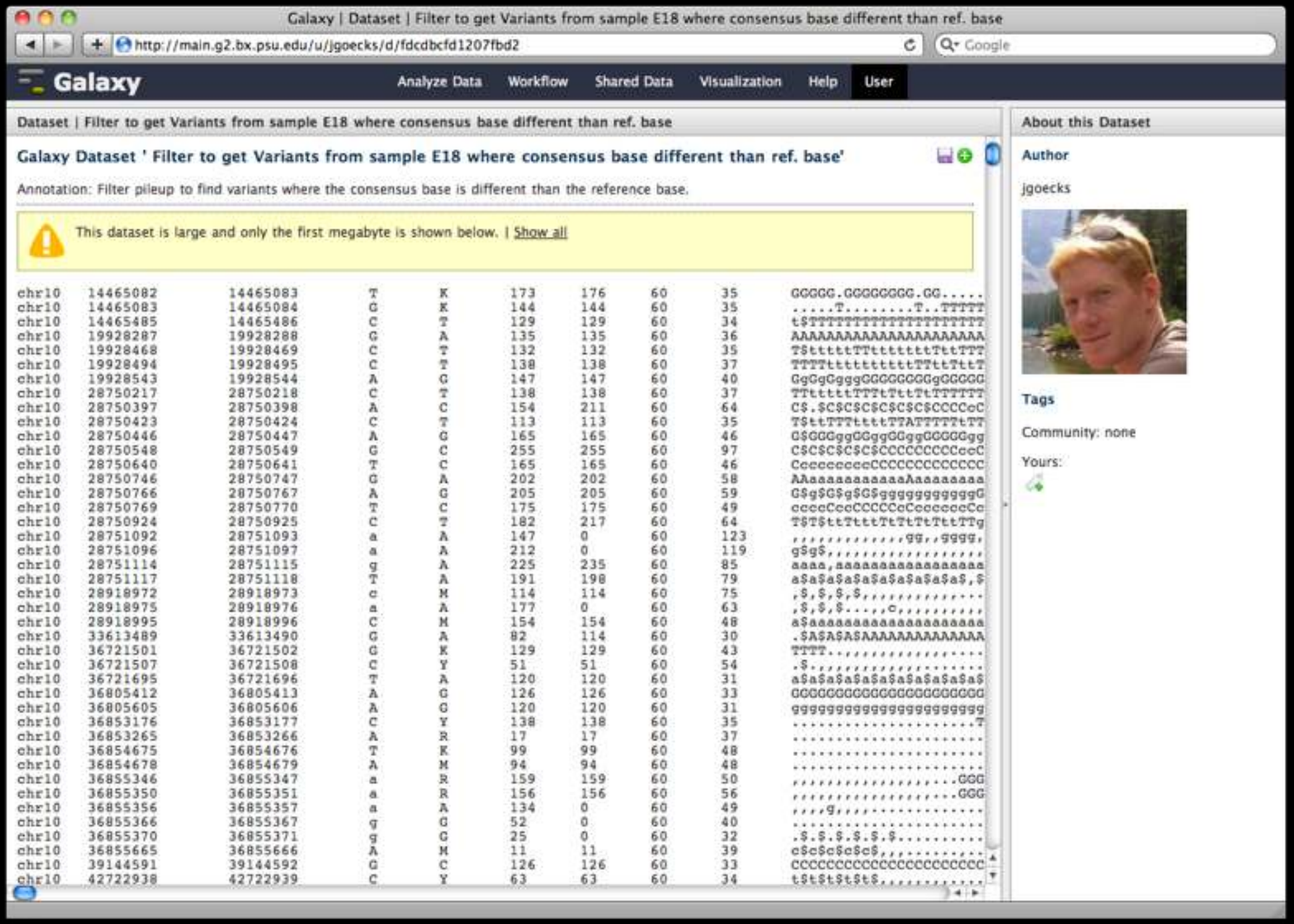
Galaxy Pages



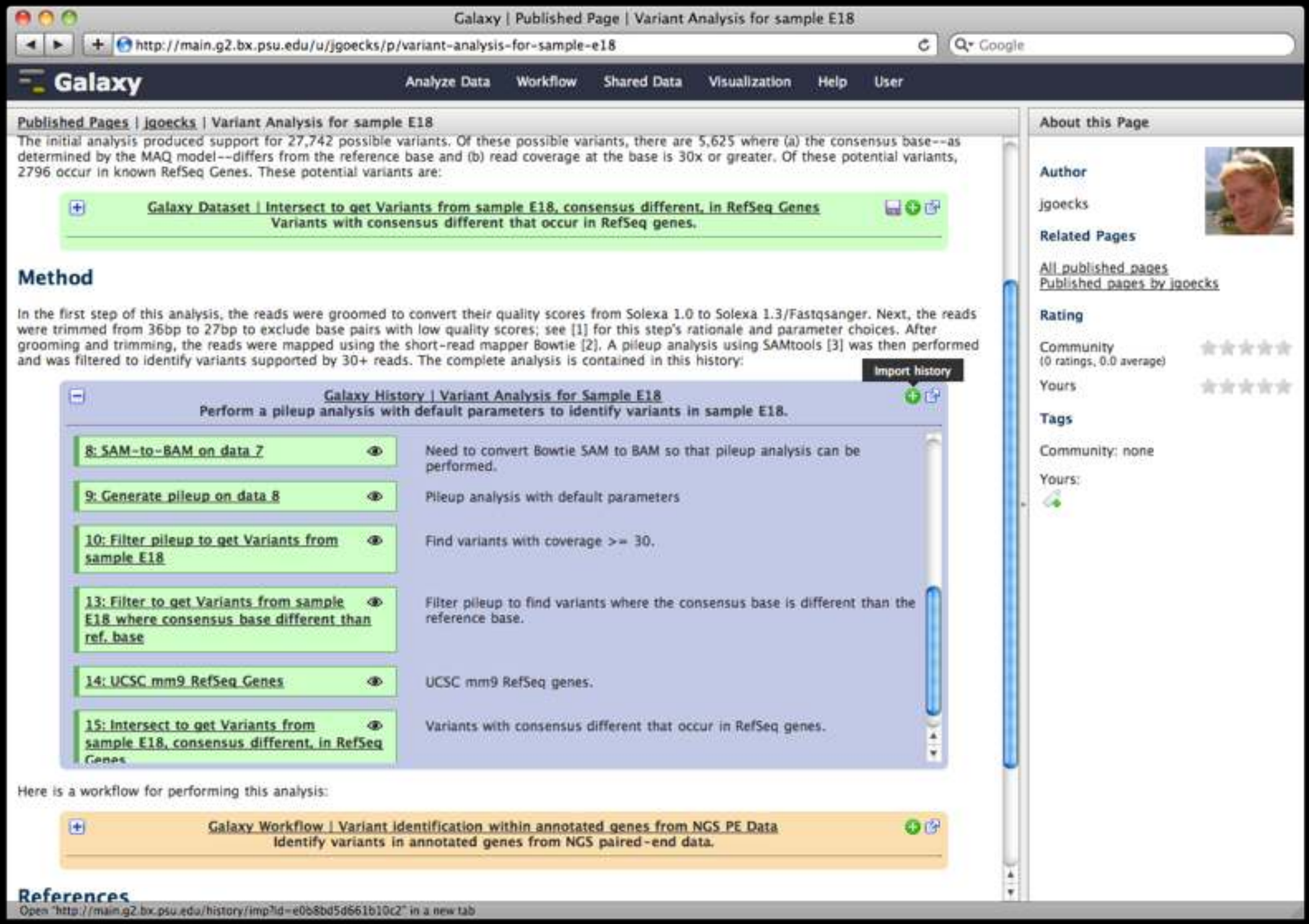
Galaxy Pages



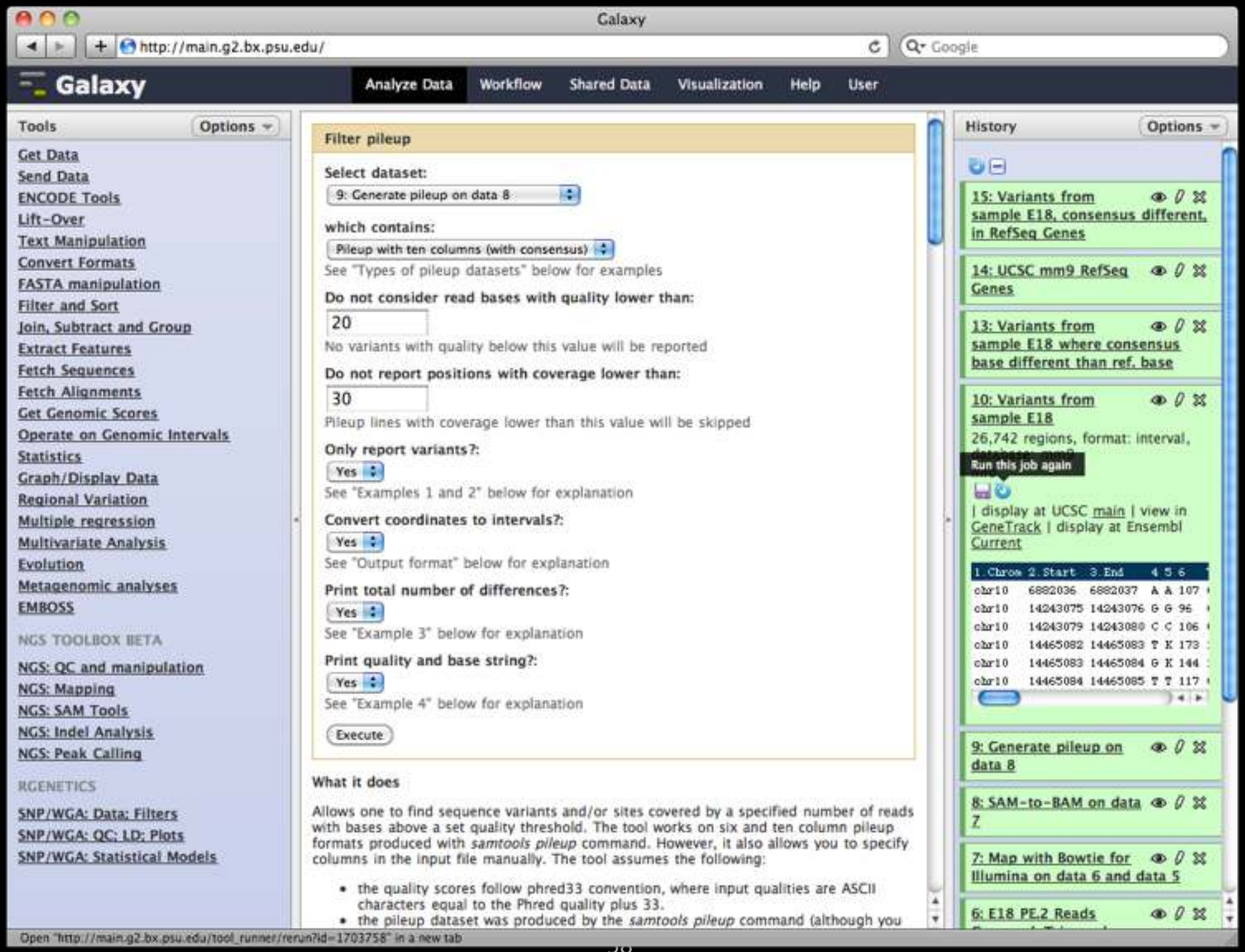
Galaxy Pages



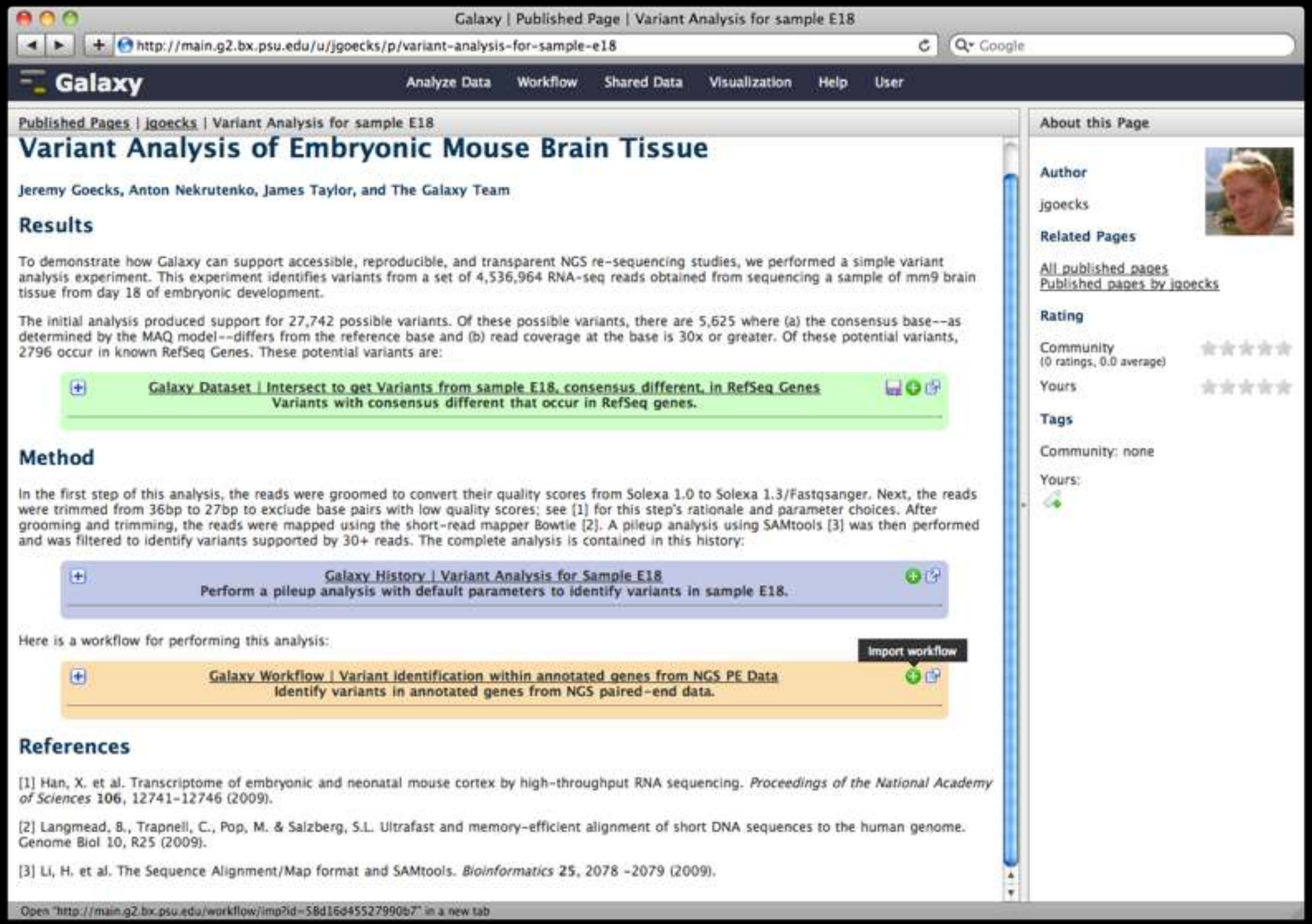
Galaxy Pages



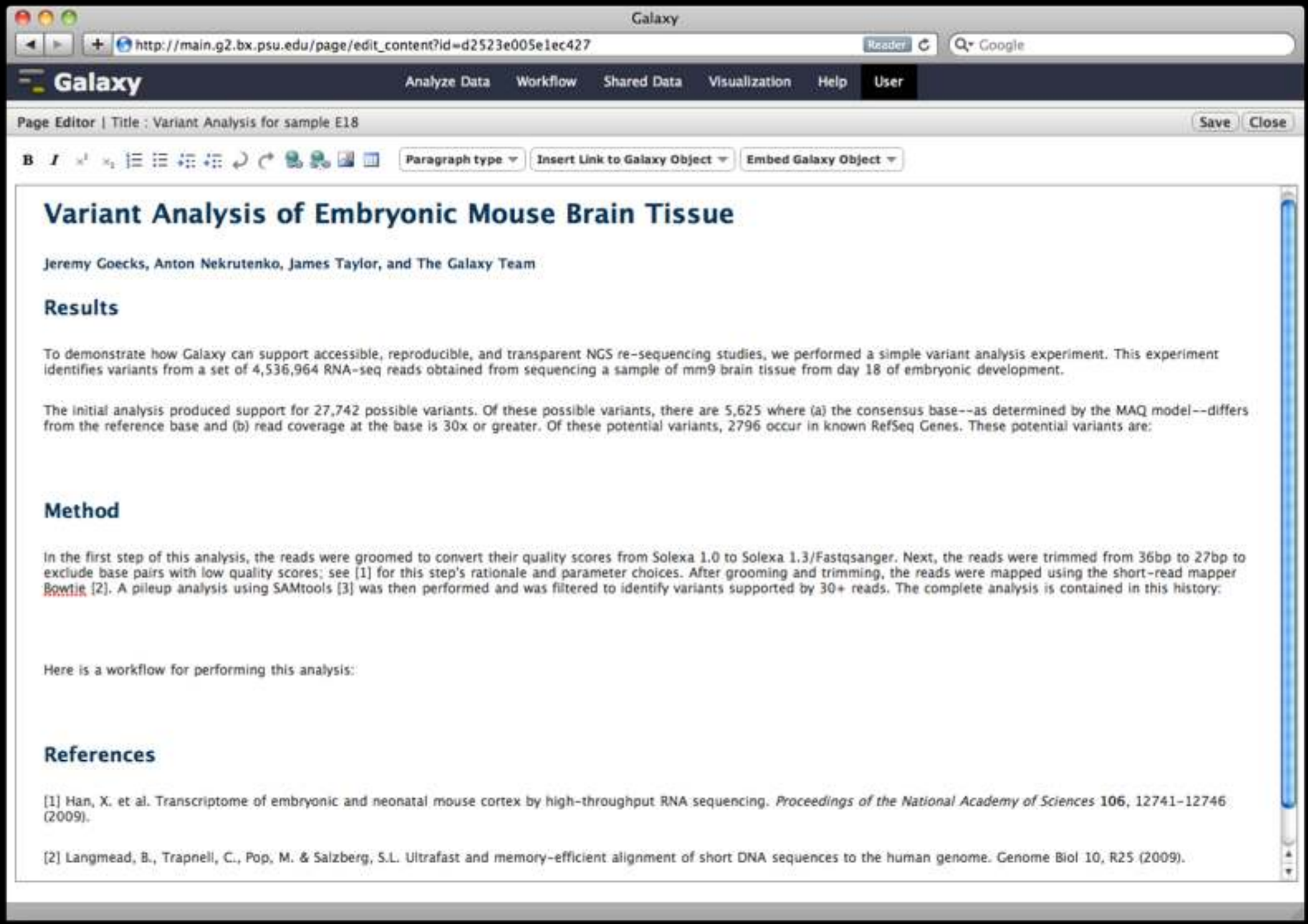
Galaxy Pages



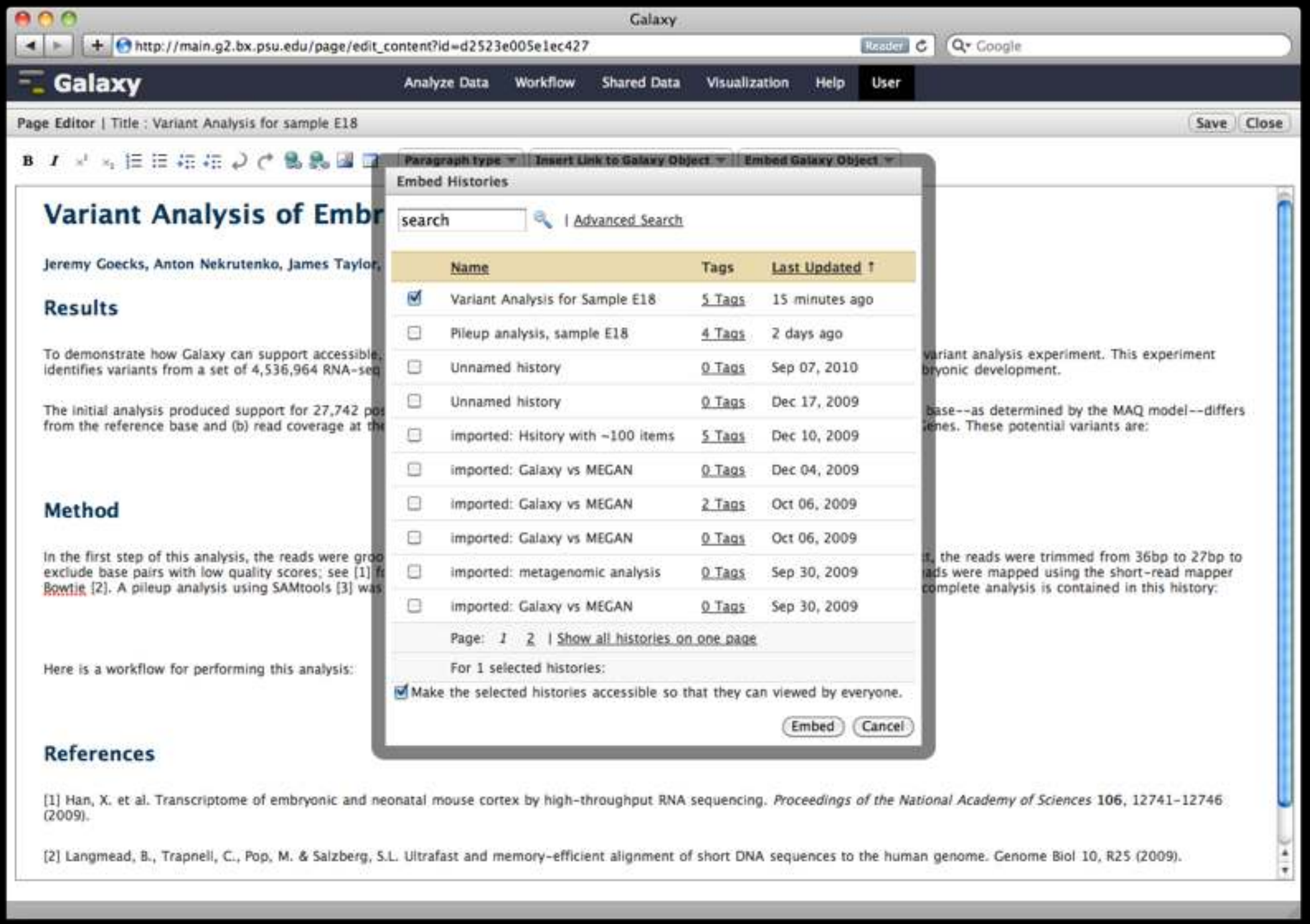
Galaxy Pages



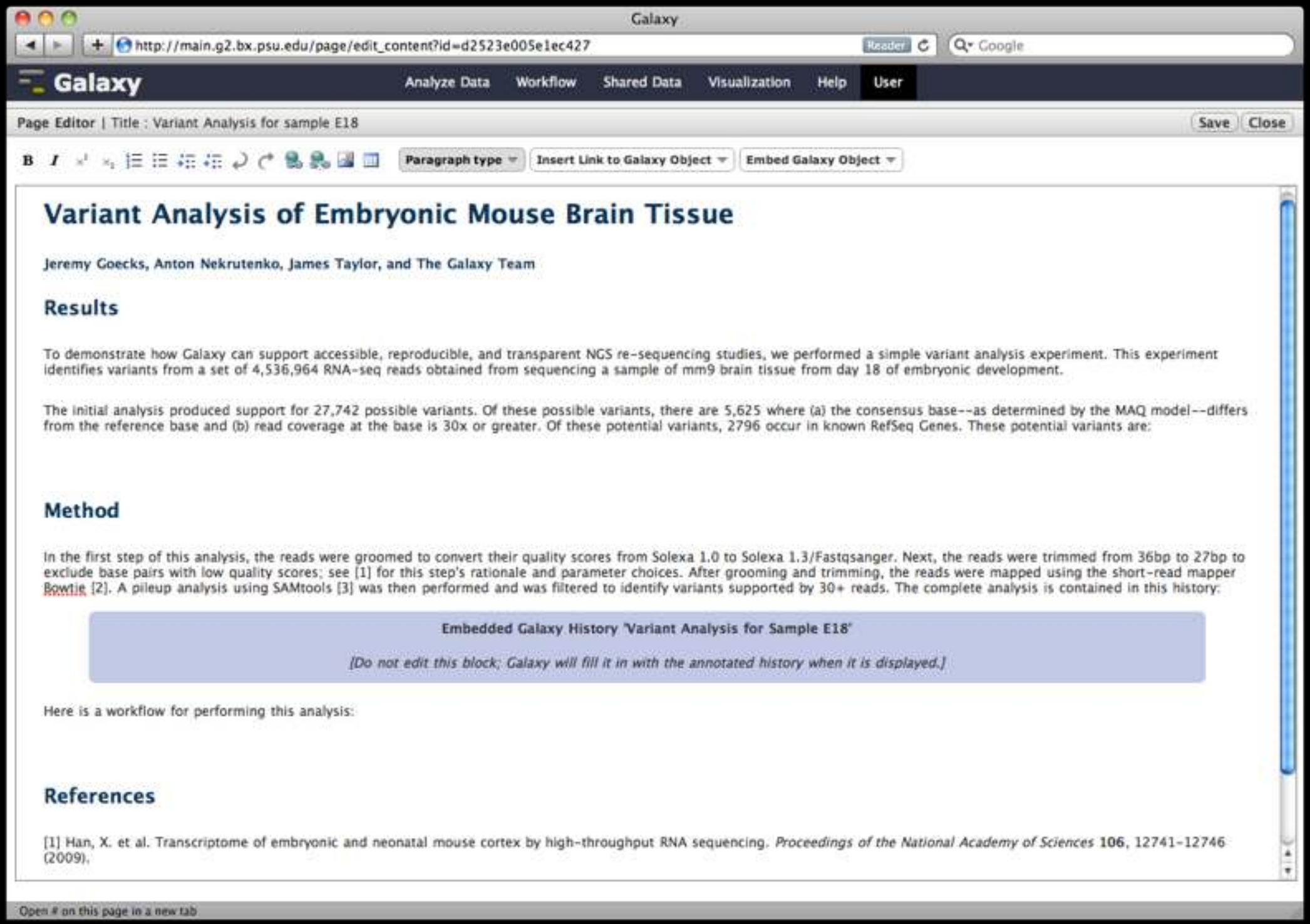
Creating a Page



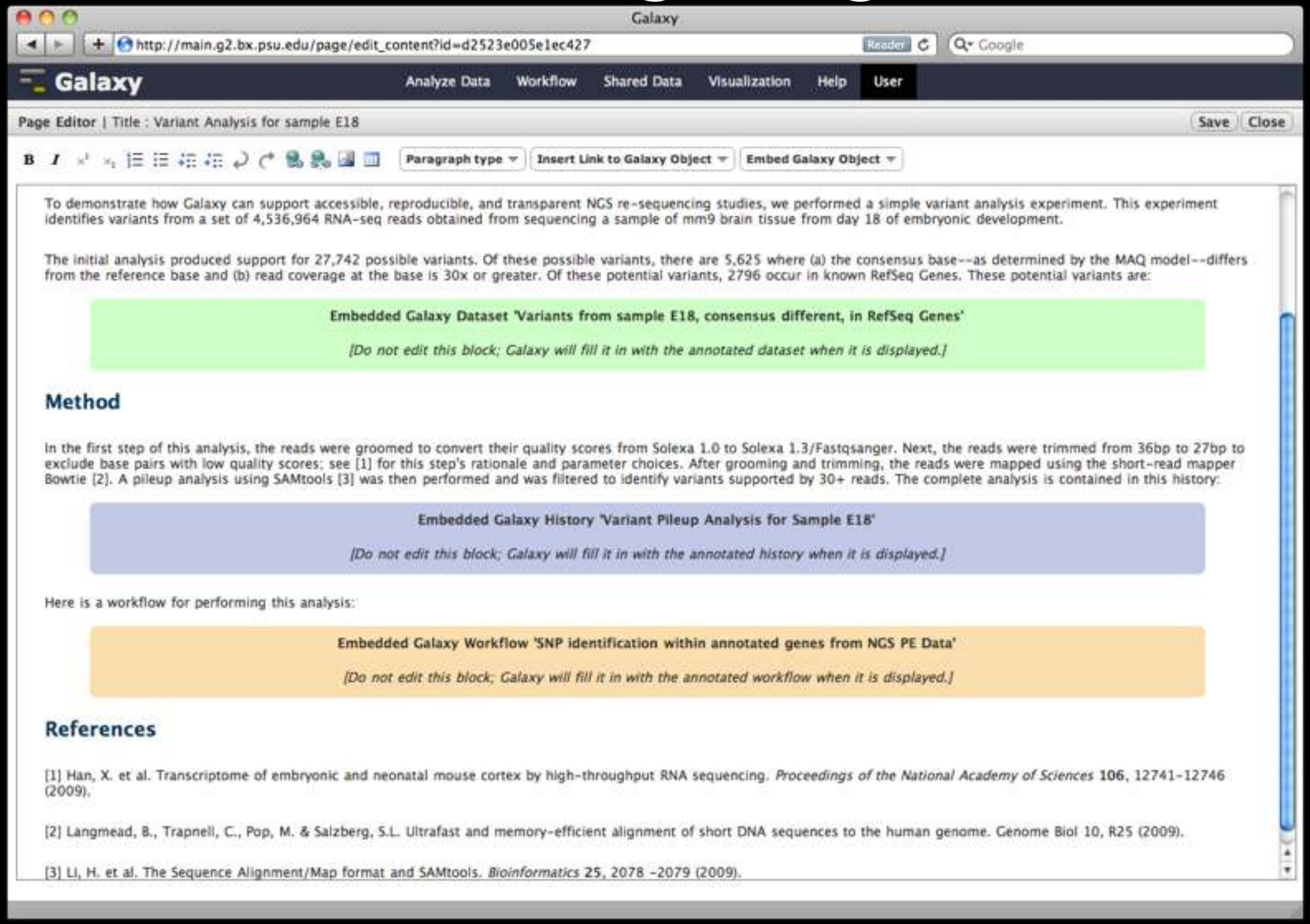
Creating a Page



Creating a Page



Creating a Page



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)

CLUSTER RESOURCES TORQUE

GRID ENGINE

Platform Computing

DRMAA
Distributed Resource Management Application API — www.drmaa.org

Running a Production Server

Use a real database server: PostgreSQL, MySQL

Run on compute cluster resources

External Authentication: LDAP, Kerberos, OpenID

Load balancing; proxy support

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

Lack IT knowledge or resources?

No problem, just use the Cloud

Overview

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

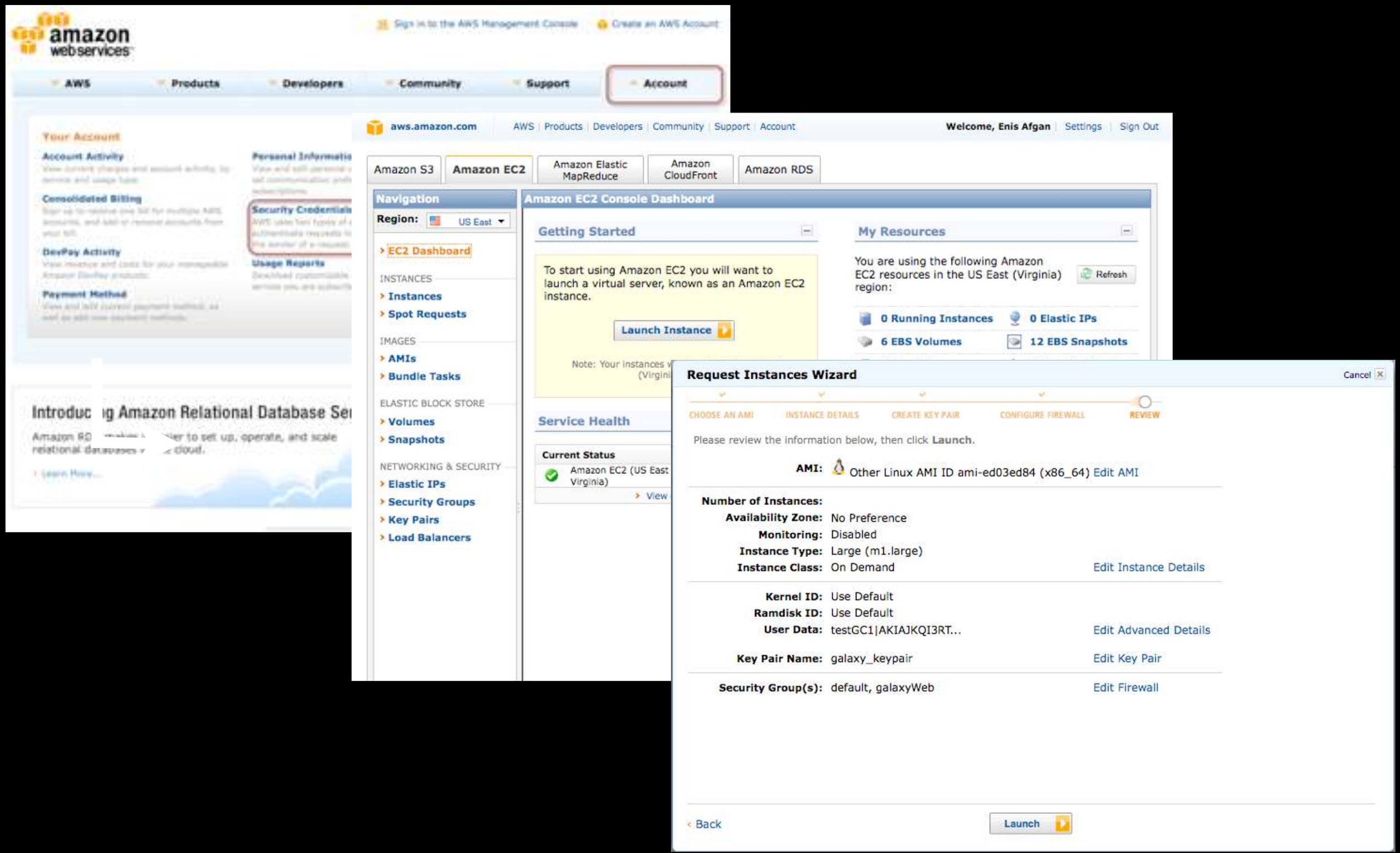
Have oscillating data volume

Deploying Galaxy on the AWS Cloud

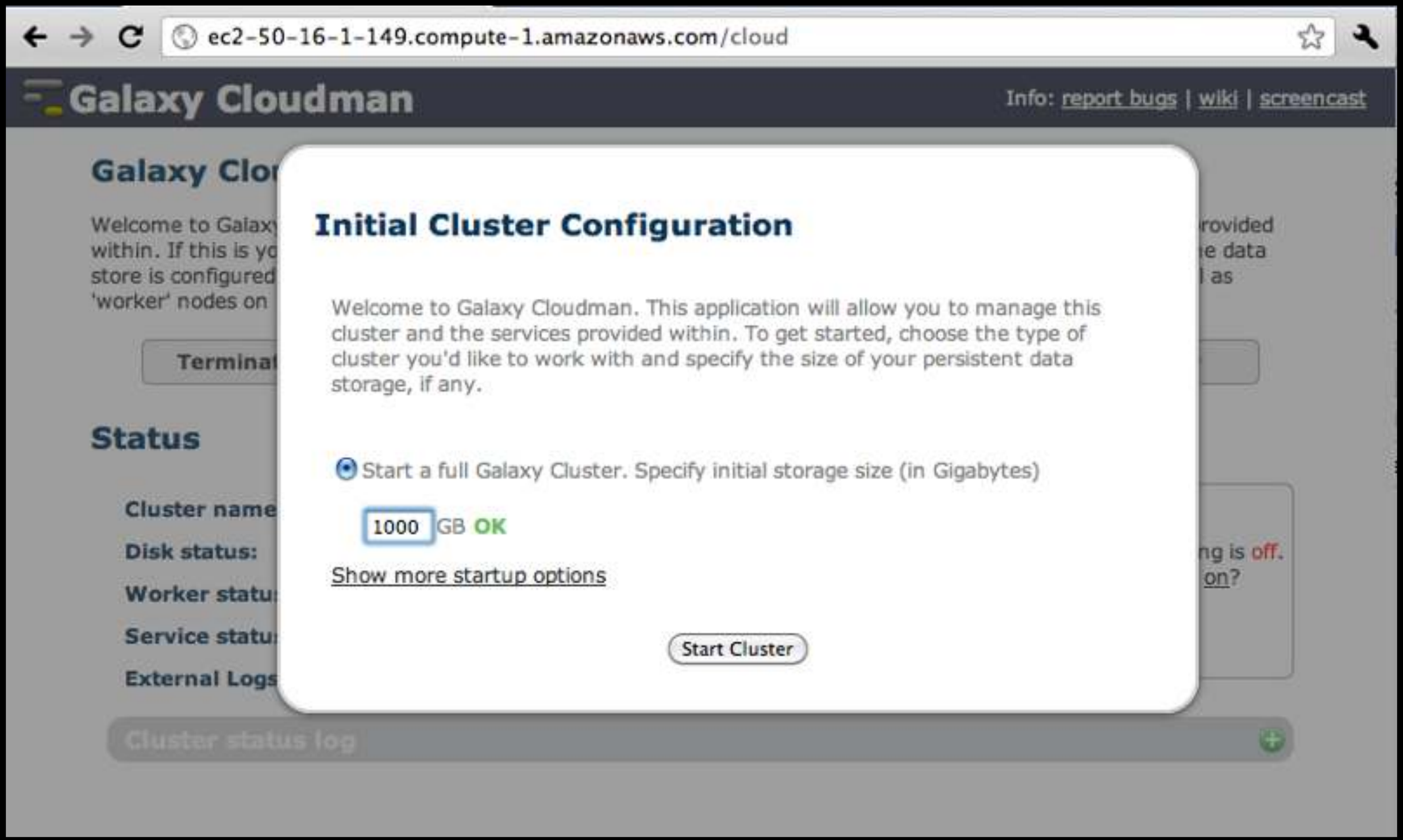
http://usegalaxy.org/cloud

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



3. Configure Your Cluster



Galaxy Cloud

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

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

Cluster status log

Galaxy

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

Google

Galaxy | Analyze Data | Workflow | Data Libraries | Help | User

Tools

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

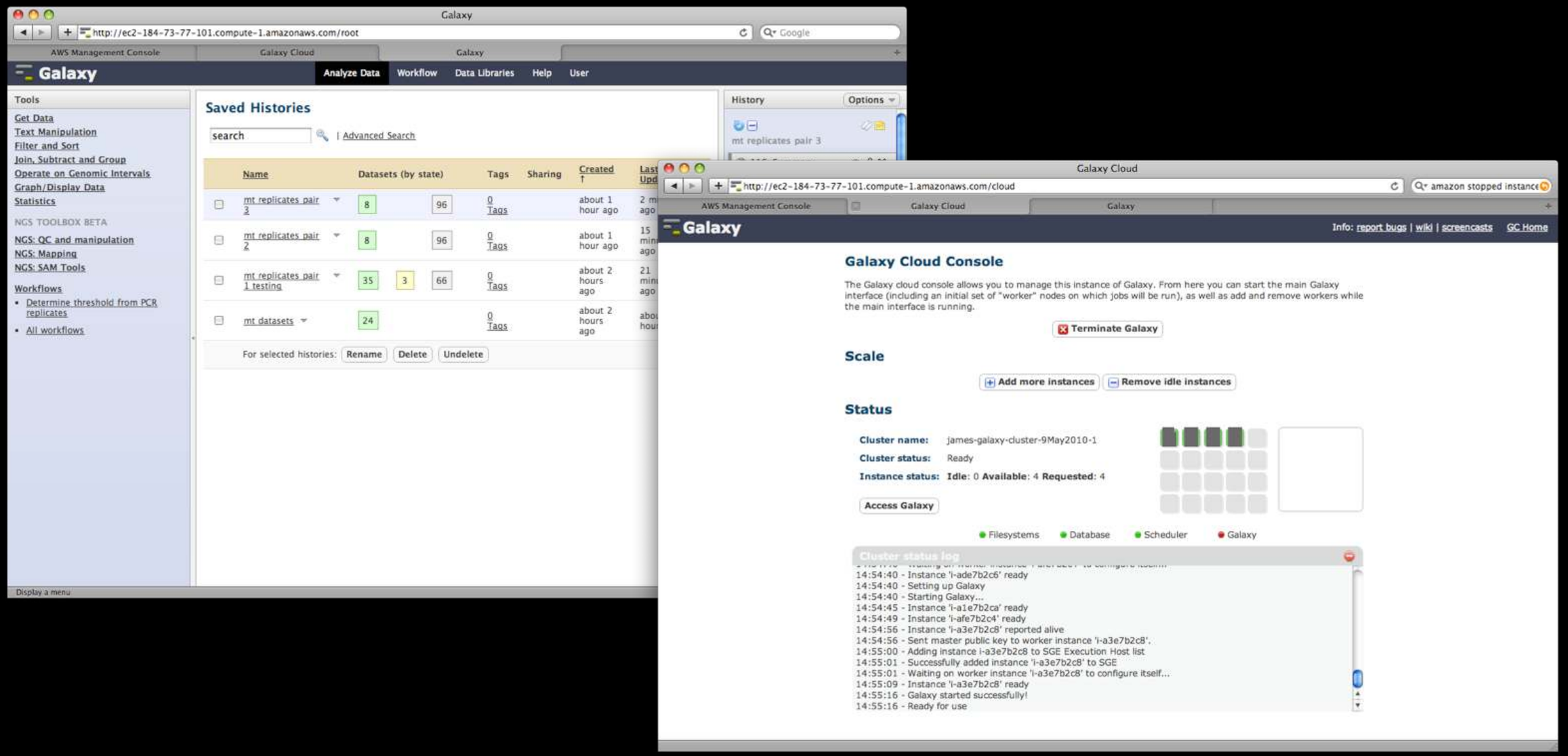
NGS TOOLBOX BETA

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

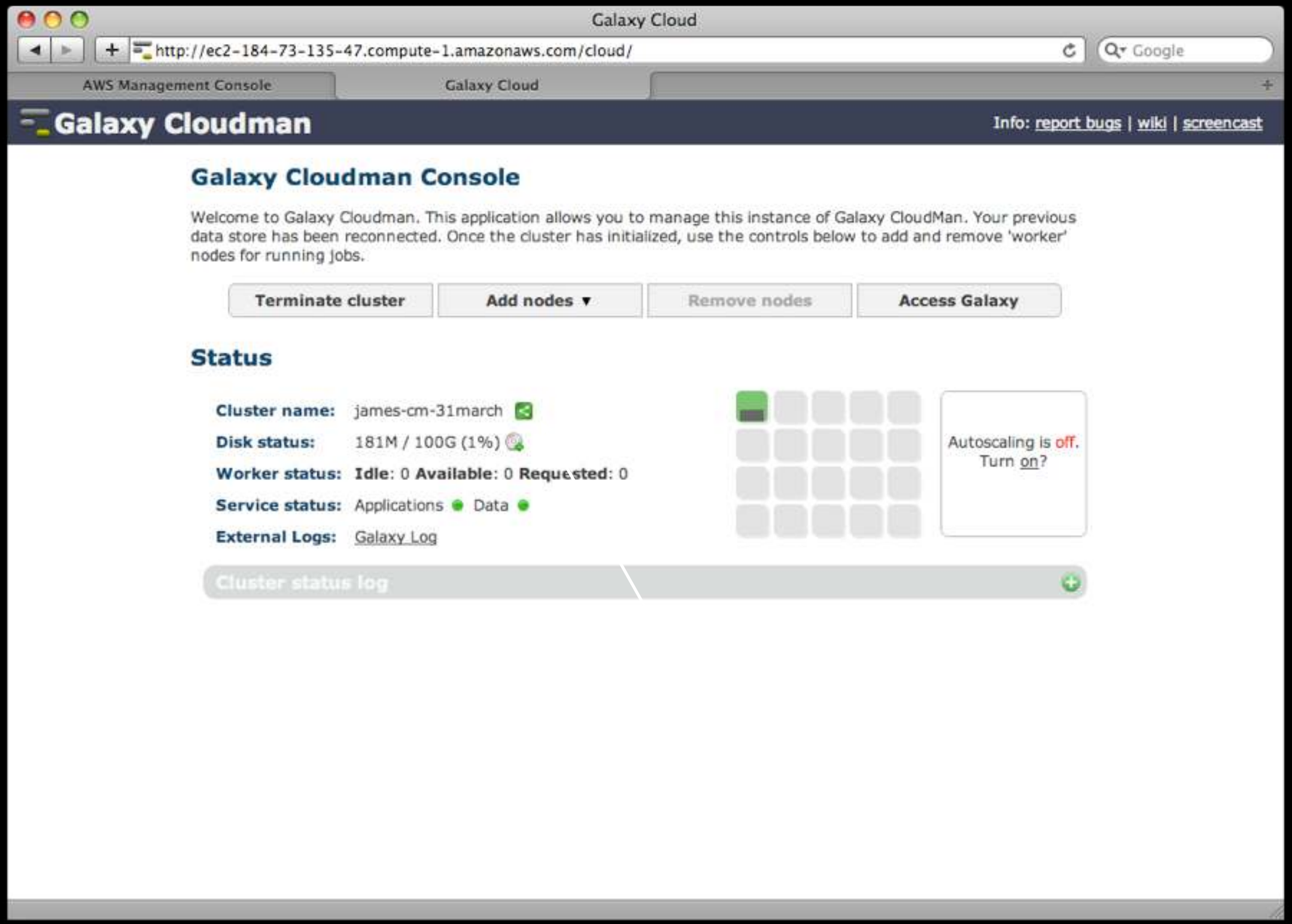
Welcome to Galaxy on the Cloud

History | Options

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



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


Galaxy Cloud
http://ec2-184-73-135-47.compute-1.amazonaws.com/cloud/
Google
AWS Management Console
Galaxy Cloud
Galaxy Cloudman
Info: report bugs | wiki | screencast
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.
Terminate cluster
Add nodes ▾
Remove nodes
Access Galaxy
Status
Cluster name: james-cm-31march
Disk status: 181M / 100G (1%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data
External Logs: Galaxy Log
Autoscaling is off.
Turn on?
Cluster status log

Galaxy Cloud

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

Google

AWS Management Console | Galaxy Cloud

Galaxy Cloudman

Info: report bugs | wiki | screencast

Galaxy Cloudman Console

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

Terminate clust… Access Galaxy

Status

Cluster name: jame…

Disk status: 181…

Worker status: Idle

Service status: Appli…

External Logs: Gala…

Autoscaling is off. Turn on?

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

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

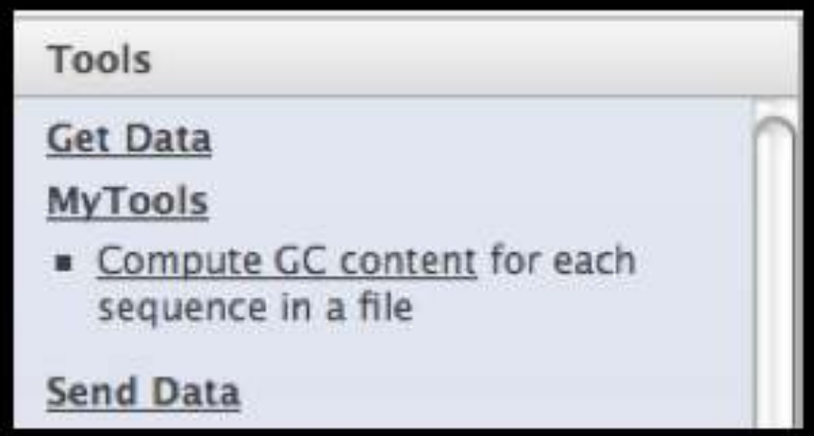
A Basic Tool

```
<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>
```




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

tool_conf.xml

Cluster

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

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

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

---

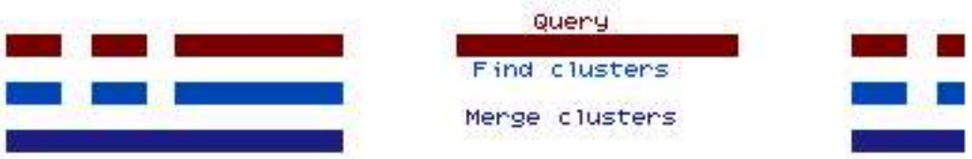
Screencasts!

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

Syntax

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

Example



cluster.xml

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

.. class:: infomark

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

-----

**Screencasts!**

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

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

-----

**Syntax**

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

Line: 87 Column: 8 XML Soft Tabs: 2

Input Parameter types

Basic

- Text
- Integer
- Float
- Select
  - Static
  - Dynamic
- Boolean

- Genome build
- Data column
- Data
- Hidden
- Base URL
- File
- Drill down

- Grouping
  - Conditional
  - Repeat
- Config Files

Datasets and Datatypes

All datasets are associated with a Datatype

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

datatypes_conf.xml and lib/galaxy/datatypes

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

Adding your Own Display Application

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

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

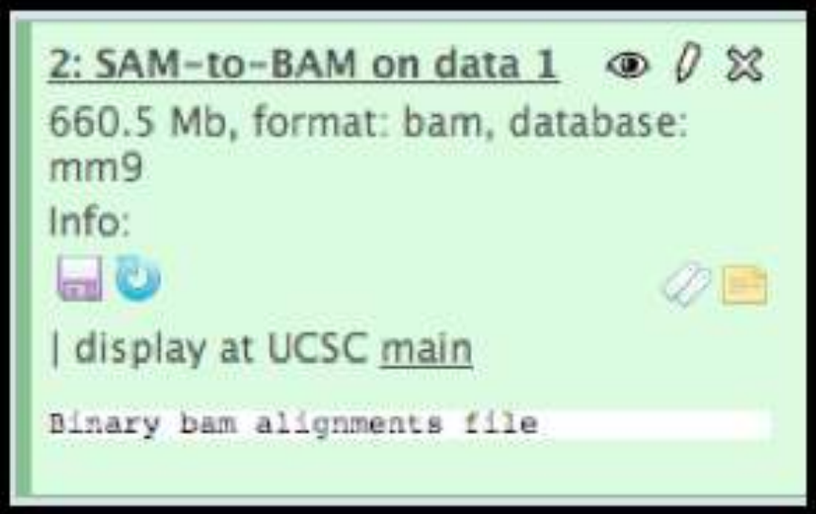
Inform Galaxy about the new display by adding to the appropriate datatype in datatypes_conf.xml

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

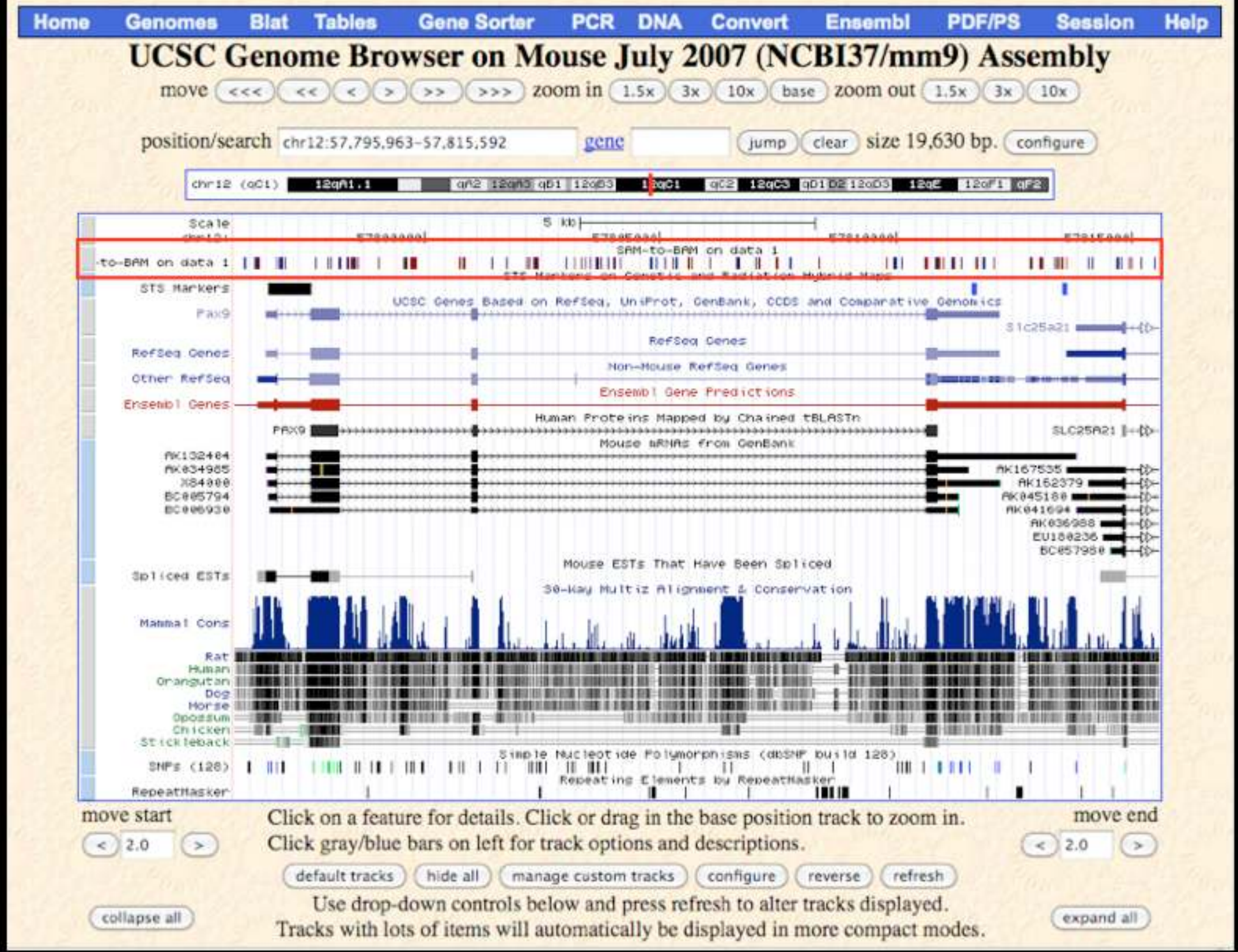
Static External Display Application

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

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



BAM at UCSC


Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl PDF/PS Session Help
UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr12:57,795,963-57,815,592 gene jump clear size 19,630 bp. configure
chr12 (qC1)
Scale
5 kb
SAM-to-BAM on data 1
to-BAM on data 1
STS Markers
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Pax9
RefSeq Genes
Other RefSeq
Non-Mouse RefSeq Genes
Ensembl Genes
Ensembl Gene Predictions
Human Proteins Mapped by Chained tBLASTn
PAX9
SLC25A21
Mouse mRNAs from GenBank
AK132484
AK034985
X84000
BC005794
BC006930
AK167535
AK152379
AK045180
AK041694
AK036988
EU160236
BC057950
Spliced ESTs
Mouse ESTs That Have Been Spliced
30-Way Multiz Alignment & Conservation
Mammal Cons
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback
SNPs (128)
Simple Nucleotide Polymorphisms (dbSNP build 128)
RepeatMasker
Repeating Elements by RepeatMasker
move start
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.
move end
2.0
default tracks hide all manage custom tracks configure reverse refresh
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
collapse all
expand all

Dynamic External Display Application

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

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

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

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

    </dynamic_links>
</display>
```

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

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

Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

Galaxy Tool Shed

http://usegalaxy.org/community

...

private Tool Sheds

1

2

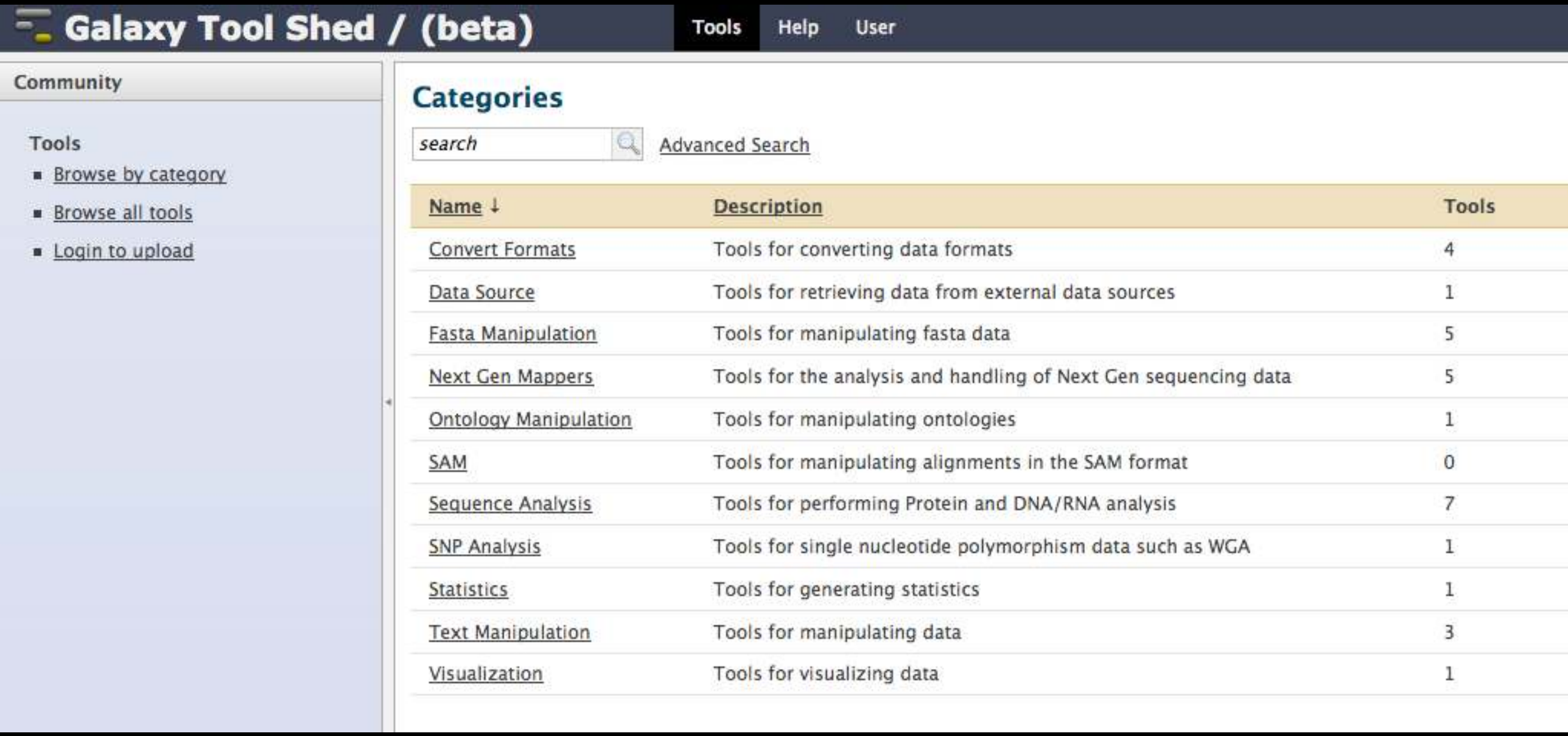
3

...

∞

private Galaxy installations

Get and Contribute Tools



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

Enis Afgan

Dannon Baker

Dave Clements

Jeremy Goecks

Kanwei Li

James Taylor

PENNSTATE.

Dan Blankenberg

Nate Coraor

Jennifer Jackson

Greg von Kuster

Kelly Vincent

Anton Nekrutenko

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

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

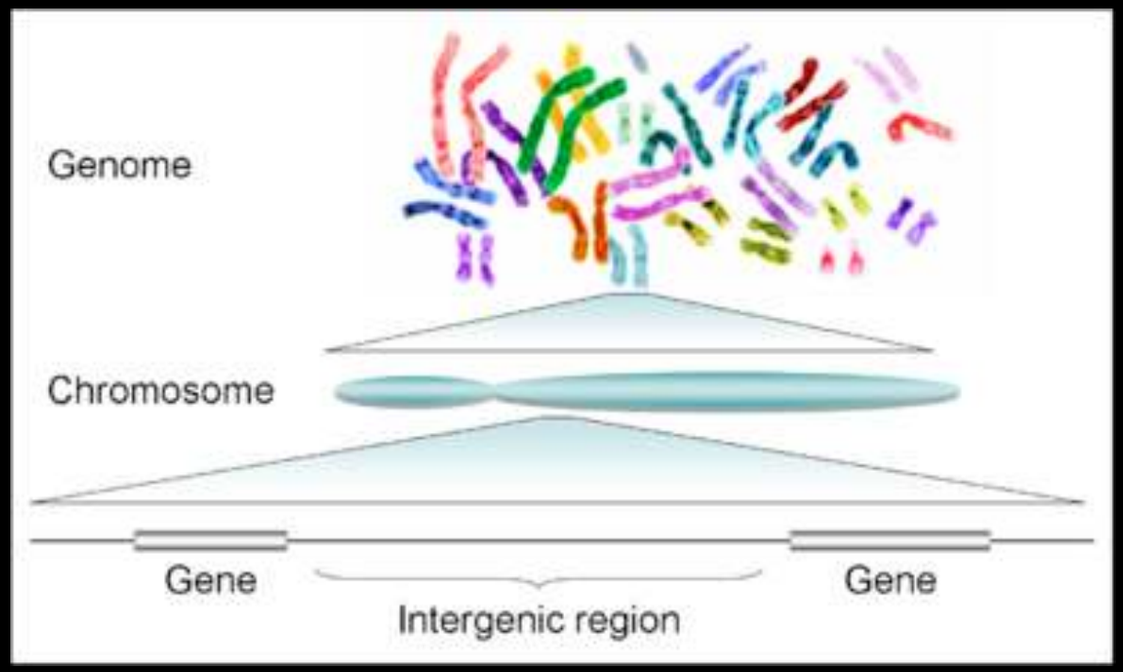
Overview

- 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
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac

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

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

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

Get Genomic data from UCSC Table Browser

Determine each SNP that overlaps with a specific coding exon

Calculate count of overlapping SNPs for each exon

Sort and select exons by greatest SNP counts